EQUALIZING LOUDSPEAKERS IN REVERBERANT ENVIRONMENTS USING DEEP CONVOLUTIVE DEREVERBERATION

Silvio Osimi

Università di Parma, Parma, IT
silvio.osimi@unipr.it

Leonardo Gabrielli

Università Politecnica delle Marche, Ancona, Italy
l.gabrielli@univpm.it

Samuele Cornell

Carnegie Mellon University, Pittsburgh, USA
samuele.cornell@ieee.org

Stefano Squartini

Università Politecnica delle Marche, Ancona, Italy
s.squartini@univpm.it

ABSTRACT

Loudspeaker equalization is an established topic in the literature, and currently many techniques are available to address most practical use cases. However, most of these rely on accurate measurements of the loudspeaker in an anechoic environment, which in some occurrences is not feasible. This is the case, e.g. of custom digital organs, which have a set of loudspeakers that are built into a large and geometrically-complex piece of furniture, which may be too heavy and large to be transported to a measurement room, or may require a big one, making traditional impulse response measurements impractical for most users. In this work we propose a method to find the inverse of the sound emission system in a reverberant environment, based on a Deep Learning dereverberation algorithm. The method is agnostic of the room characteristics and can be, thus, conducted in an automated fashion in any environment. A real use case is discussed and results are provided, showing the effectiveness of the approach in designing filters that match closely the magnitude response of the ideal inverting filters.

1. INTRODUCTION

The accurate reproduction of audio signals through loudspeaker systems has been a longstanding pursuit in audio engineering. Variations in transducer characteristics, enclosure designs and materials introduce deviations from the ideal response, leading to coloration, distortion, and a loss of fidelity in the reproduced sound. To address these challenges, the field of loudspeaker equalization has seen significant advancements over the years and is now very mature.

Early methods for loudspeaker equalization relied on electrical networks to compensate for inherent transducer deficiencies. However, with the advent of digital signal processing, a wide range of sophisticated equalization techniques has emerged, offering precise control over the frequency and/or phase response, transducer distortion, directivity, etc.

Most equalization techniques rely on an accurate measurement of the device, which must be done in acoustically anechoic or semi-anechoic environments. This is generally not a problem, since loudspeakers are generally designed by companies which have the facilities for conducting measurements on prototypes and then produce them serially.

However, there are different use cases where accurate impulse response measurements inside an anechoic room are not feasible. This may be the case, e.g. of amateur loudspeaker projects, or workshops building site-specific loudspeakers directly in site. Another use case, that will be taken as use case in this paper, is the manufacturing of large custom digital church organs. These include a sound emission system based on a set of loudspeakers that must be carefully tuned, but the size and weight of such instruments require large measurement rooms and complex transportation procedures. Furthermore, the wooden cabinet may be custom-made and assembled by artisans at the final destination (e.g. theater or church), making the measurement impossible. Unfortunately, performing IR measurement of the loudspeakers in a reverberant environment makes it difficult to separate the response of the loudspeakers and the cabinet (which reflects and filters sound waves propagating from the loudspeakers) from that of the room.

1.1. Prior Art and Scope of the Work

A plethora of equalization methods can be found in the literature that require anechoic measurement of the system to be inverted [1]. Measurements can be conducted by assuming the system to be either linear or non-linear. In the first case the IR (or its frequency-domain transform) is sufficient to provide an exact mathematical description of the system. The validity of this assumption must be questioned case by case, but for the sake of simplicity, in many works this is assumed true. Indeed, in this work, we will consider the systems under test linear.

The problem of non-anechoic measurement of loudspeakers has been addressed less often. Theoretical foundations of acoustic measurements in reverberant environments were laid by Richard Heyser [2], later leading to time-selective techniques [3]. These techniques are based on merging a near-field measurement at low frequency and a time-windowed far-field measurement at high frequency. Although being scientifically solid, there are several reasons why such a method is not applicable to our scenario. First of all, it is not automatic, but requires at least two separate measurements (or more, for ported speakers), with two microphone positions. It also requires a fair amount of knowledge in acoustics to figure out the right position for the microphones to be posi-

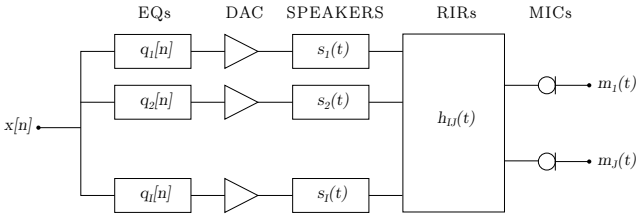


Figure 1: Diagram of the problem setting. Digital equalizing filters $q_i[n]$ must be designed in order to invert the loudspeakers IR $s_i(t)$, while leaving the room impulse responses (RIR) $\underline{h}_{ij}(t)$ unaffected. The effect of the room must be removed to estimate viable filters.

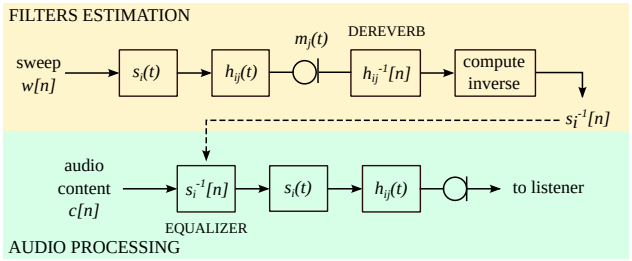


Figure 2: Diagram of the proposed method.

tioned, partly subject to room size and properties. Merging the two measurements is not automatic as well, thus the frequency overlap region where both measurements are still valid and a match must be found manually. Finally, this test technique has been devised for on-axis measurement of simple loudspeakers. Whenever complex sound sources have to be measured at some specific point in space, as will be in our use case, things may not work as expected. Later approaches have been proposed based on the same premises [4, 5, 6]. Another relevant work to the field of room and loudspeaker equalization can be found in [7], where a very insightful introduction to the topic is given. However, the method also inverts the room together with the loudspeaker, which we are not interested in.

The assumptions behind this work follow: (1) no prior knowledge of the room type and size, therefore the method must be general; (2) we leave all the decisions to an automatic system, therefore no human intervention or subjectivity must be involved; (3) we deal with off-axis loudspeaker measurement; (4) we require the system to work on swept sines to perform the measurement once and estimate the optimal equalization filters with a well-known reference method. Following these assumptions, we propose an automatic system to equalize a loudspeaker system in a non-anechoic environment based on a blind dereverberation algorithm. The algorithm requires no prior knowledge of the room size and characteristics and can, thus, be performed automatically during the final deployment of the sound system, being entirely automated except from the positioning of the measuring microphones in the space. To the best of our knowledge no prior work proposed a method to work in such a generic setting.

The remainder of this paper is organized as follows: Section 2 describes the proposed method for equalization in reverberant environments. Section 3 describes in more detail the use case which provides motivation for this work. Section 4 provides a description of the experimental setup for the specific use case, while Section 5 reports the results and discusses them. Finally, Section 6 provides concluding remarks and outlines directions for future research.

2. PROPOSED METHOD

Let us first introduce the setting of the problem, depicted in Figure 1. A loudspeaker or a system of multiple loudspeakers is fitted in a room. The IR of the loudspeakers $s_i(t)$ is unknown and its measurement is affected by the room impulse responses (RIR) $h_{ij}(t)$ between the sources and the measurement microphones, which capture the signals $m_j(t)$. Digital filters $q_i[n]$ of length L can be applied before the digital to analog conversion (DAC) to equalize the $s_i(t)$. Please note that the discrete-time index will be indicated as $[n]$, while the continuous time as (t). An IR measurement is possible for each loudspeaker-microphone pair, e.g., by using a sine sweep as input $x[n]$ according to the method from [8]. However, the resulting IR will be the convolution of the loudspeaker and the RIR.

Without *a priori* knowledge of the room, it is not possible to obtain a direct estimate of the loudspeaker response. However, assuming that the microphones are close to the loudspeakers, a dereverberation algorithm should be able to separate the effects of the room, that are expected to arrive much later than the direct signal from the loudspeaker or the first reflections (which may be partly due to the loudspeaker cabinet itself, as will be in the described use case).

Please note that the method is not particularly effective when the early reflections take longer than L samples to get into the microphone, since in that case the loudspeaker IR can be separated from room IR by windowing the first L samples and computing the inverse. Therefore in the target use case some of the room walls or obstacles are quite close to the loudspeaker.

2.1. Overview

The proposed method is depicted in Figure 2 and is divided in two stages: a filter design stage, and the actual real-time audio processing stage. A swept sine measurement is first conducted through the I loudspeakers in the room and a matrix of IRs $h_{IJ}(t)$ is obtained from the J microphones. These are processed through a dereverberation algorithm to cancel the effect of the room and obtain an estimate of the loudspeakers IRs that can be fed to a method to compute inverting filters. The method employed in the rest of the paper is the one from Kirkeby [9], which finds the optimal inverse filters in the Least Square Error (LSE) sense. Although the method is not recent, is regarded as a standard method for audio equalization in the scientific literature and is well known, making it suitable to allow us investigating the potential of the proposed method.

Once equalizing filters have been designed using Kirkeby's method, they can be employed in the application to equalize the loudspeakers. At this stage any digital audio content $c[n]$ is fed to the equalizers and then to the loudspeakers. The room impulse response will alter the audio content but the effect of the loudspeakers should be compensated by the digital filters.

2.2. Deep Dereverberation Algorithm

Several methods exist in the scientific literature for speech dereverberation [10, 11] and (less often) music dereverberation. Among the numerous methods one stands out for the wide adoption as a reference and is based on regular Digital Signal Processing (DSP)

techniques, which is commonly referred to as Weighted Prediction Error (WPE) [12, 13].

WPE computes a filter to estimate late reverberation from past observations and subtracts this part from the speech affected by reverberation, to obtain the target speech. The filter design is based on a variance-normalized linear prediction, suitable to identify correlations in the signal that are manifestation of the room reflections. The algorithm relies on a prediction delay which, unfortunately, poses limitations to the shortest delay that it can estimate, and thus to the separation between early and late reflections (to be removed). WPE will be considered as a baseline method, and compared, in the evaluation phase, to a more advanced Deep Neural Network (DNN)-based method.

Methods based on DNN are more flexible in tasks such as dereverberation [14, 15, 16] (i.e. target speech prediction), speech enhancement or speaker diarization. However, these suffer the common issue of being black-boxes. DNN methods leveraging WPE have been also envisioned [17, 18], with [18] being one of the latest incarnation and, according to its authors, it addresses several shortcomings.

The method selected for this work, from now on referred to as DNN-FCP, is based on estimating the direct-path clean signal using a DNN and then approximate the RIR with a forward filter. This procedure is called Forward Convolutive Prediction (FCP) and is introduced in [19]. It is worth to discuss the difference between WPE and DNN-FCP.

WPE computes a K-tap inverse linear filter to estimate the late reverberation at the current frame from the past observations. The estimated late reverberation is then subtracted from the mixture for dereverberation, i.e. for a single-speaker scenario,

$$\hat{S}_{\text{WPE}}(t,f) = Y(t,f) - \hat{\mathbf{g}}(f)^{\mathsf{H}}\widetilde{\mathbf{Y}}(t-\Delta,f), \tag{1}$$

where $\hat{\mathbf{g}}(f) \in \mathbb{C}^K$ is a K-dimensional filter, $\mathbf{Y}(t,f)$ is the last frame of the recorded mixture in the time-frequency domain, $\Delta\,(\geq 1)$ a prediction delay, and $\widetilde{\mathbf{Y}}(t,f) = [Y(t,f), Y(t{-}1,f), \ldots, Y(t{-}K+1,f)]^{\mathsf{T}}$. Under certain assumptions, WPE computes the maximum likelihood estimation filter through the minimization problem

$$\underset{\mathbf{g}(f),\lambda(\cdot,f)}{\operatorname{argmin}} \sum_t \frac{|Y(t,f) - \mathbf{g}(f)^{\mathsf{H}}\,\widetilde{\mathbf{Y}}(t-\Delta,f)|^2}{\lambda(t,f)} + \log\lambda(t,f), \tag{2}$$

where λ is the time-varying zero-mean PSD of the signal. Unfortunately, this objective does not have a closed-form solution and a solution must be iteratively found by minimizing alternatively either of the two objectives.

DNN-based works aim at estimating the PSD with a DNN model, therefore simplifying the objective function and allowing for a closed-form solution [17]:

$$\underset{\mathbf{g}(f)}{\operatorname{argmin}} \sum_t \frac{|Y(t,f) - \mathbf{g}(f)^{\mathsf{H}}\,\widetilde{\mathbf{Y}}(t-\Delta,f)|^2}{\hat{\lambda}(t,f)}, \tag{3}$$

where $\hat{\lambda}(t,f)$ is provided by a dedicated DNN. The dereverberated result $\hat{S}_{\text{DNN}}(t,f)$ is obtained similarly to Eq. 1. A step further can be obtained by also estimating the target speech with a DNN, thus removing the delayed $\widetilde{\mathbf{Y}}(t-\Delta,f)$. Now the problem requires estimating the linear filter $\mathbf{g}(f)$ that minimizes the following [19]:

$$\underset{\mathbf{g}(f)}{\operatorname{argmin}} \sum_t \frac{|Y(t,f) - \mathbf{g}(f)\,\widetilde{\hat{\mathbf{S}}}_{\text{DNN}}(t,f)|^2}{\hat{\lambda}(t,f)}, \tag{4}$$

The dereverberation result is obtained as

$$\hat{S}_{\text{FCP}}(t,f) = Y(t,f) - \Big(\hat{\mathbf{g}}(f)^{\mathsf{H}}\widetilde{\hat{\mathbf{S}}}_{\text{DNN}}(t,f) - \hat{S}_{\text{DNN}}(t,f)\Big), \tag{5}$$

where $\widetilde{\hat{\mathbf{S}}}_{\text{DNN}_b}(t,f) = [\hat{S}_{\text{DNN}_b}(t,f), \hat{S}_{\text{DNN}_b}(t{-}1,f), \ldots, \hat{S}_{\text{DNN}_b}(t{-}K{+}1,f)]^{\mathsf{T}}$. Please note that $\hat{\mathbf{g}}(f)^{\mathsf{H}}\widetilde{\hat{\mathbf{S}}}_{\text{DNN}}(t,f)$ is an estimate of the reverberant target speech, and $\hat{\mathbf{g}}(f)^{\mathsf{H}}\widetilde{\hat{\mathbf{S}}}_{\text{DNN}}(t,f) - \hat{S}_{\text{DNN}}(t,f)$ is the estimated reverberation of the target speaker. Therefore, FCP computes a forward filter (hence the name), as opposed to WPE that computes an inverting filter. The advantage over WPE is multiple: it is more effective in removing early reflections since the estimated target speech is not delayed by Δ; the estimated filter $\mathbf{g}(f)$ is not an inverse filter, therefore the estimation is more accurate in presence of interfering signals [19].

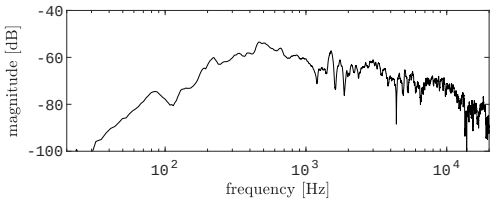
3. USE CASE

In the manufacturing of large digital church organs anechoic measurements cannot always be done. Often, the organ is large and heavy, and only large semianechoic chambers can host such instruments. Furthermore, their transportation to a measurement room is complex and the risk of scratching or damaging the wood should be minimized. Finally, these organs can be custom tailored, therefore each product is different from another. Even if multiple products share the same sound emission system (amplifiers, loudspeaker types and disposition), the frequency response may differ due different shape, size and materials of the wooden cabinet. As an example: the shape of some wooden parts may reflect the sound coming from the loudspeakers producing dips and notches in the frequency response, the different distance from the bench may alter the time of arrival of different sound waves, the wood may damp some frequency component, etc. Finally, in large custom installation the final assembly is done at the venue, therefore an anechoic measurement is impossible.

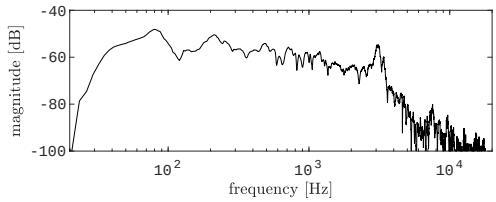
When anechoic IR measurement is not viable, the proposed method may improve the overall sound reproduction quality. In the rest of the work we shall consider the combined effect of the loudspeakers and the cabinet as one, and try to distinguish this from the effect of the room. Please note that the goal of our method is not to invert the RIR, but only to invert the sound emission system and make its magnitude response flat. Therefore, once inverting filters are designed, these are implemented in the organ tone generation DSP to pre-equalize.

Digital organs have a number of loudspeakers S with different specifications and positioning. In this work we have employed a digital organ that is small enough to fit inside a small semianechoic room, for the purpose of measuring the IRs and have material for a real-world evaluation of our technique. The organ has six loudspeakers, in particular two woofers and 2 mid-range speakers.

The target for our equalization algorithm is the organ player, since the audience of an organ concert will sit far away from the console, where the effect of the room, or the effect of other loudspeakers positioned in the space becomes predominant. Anyway, it would be too hard to equalize for an entire audience or parts of it, since many measurements should be taken. Therefore we shall



(a)



(b)

Figure 3: Magnitude Frequency Response of two of loudspeakers measured at the left microphone: (a) left midrange, (b) left woofer.

Figure 4: Picture of the organ under test. The loudspeakers IRs are measured to assess the validity of the method.

consider the position of the organ player ears and thus consider $J = 2$.

Figure 3 shows the frequency response of two loudspeakers measured at the left ear of the organ player in the anechoic environment of Figure 4. The objective of the work is to equalize these responses without recurring to their anechoic measurement.

4. EXPERIMENTS

In this work we performed experiments on a synthetic dataset composed of simulated loudspeakers IRs playing through simulated shoebox-type rooms. Then we move onto the use case provided by semianechoic measurements of real digital organ loudspeakers, to verify the effectiveness of the approach.

4.1. Dataset and Training

Differently from other works dealing with acoustics and loudspeaker equalization, here we need a training dataset for the DNN-FCP method to be employed to effectively suppress reverberation from an audio system. In speech dereverberation scenarios, the DNN-FCP is trained on a large corpus of speech. However, in our work, the dereverberation algorithm must remove the reverb from a measurement of the system obtained with microphones. In principle, the DNN-FCP can work with any time-domain signal, therefore the IR measured at the microphone could be used to feed the DNN-FCP. However, from our experiments a DNN-FCP trained on a speech corpus failed to cancel any reverb, probably due to the unseen nature of the IRs and to their extremely short duration.

For this reason, we decided to train the DNN-FCP on a dataset containing log-sweep signals affected by a number of loudspeakers and rooms. From early experiments this choice proved more effective in removing reverb from unseen log-sweeps and it motivated to construct a synthetic dataset of moderate size. The dataset is made of log-sweeps signals filtered by synthetic IRs of loudspeakers generated randomly and by synthetic IRs of shoebox rooms with different parameters, in order to provide enough generalization for the network to perform properly on real-world signals. Specifically, the test signal where dereverberation is performed during inference is a reverberated version of the loudspeaker signal coming from the organ described in Section 3, while the training dataset is generated according to the parameter ranges in Table 1. More specifically, the loudspeakers have been simulated by assuming their behavior linear. Several lumped elements models exists, such as the Thiele-Small model [20, 21, 22], however they fail to entirely characterize the complex frequency spectrum of a loudspeaker, which is given by the contribution of a large number of parasitic components, natural modes and complex physical phenomena. For this reason, we opted for a more generic approach where woofers, mid-range and full-range speakers are characterized by a low-pass and a high-pass rolloff, and an operating range where several complex conjugate poles affect the frequency response by cutting or boosting the energy by several dB. With these constraint we designed IIR filters in Matlab that attempt to mimic the loudspeakers character, whose IR have been extracted and convolved with the sweep signals.

The rooms, instead, have been generated using a widely adopted Python package, PyRoomAcoustics [1]. In this case the rooms have all been designed in a shoebox shape with three different sizes and RT60, in order to simulate three different kinds of venues, from a medium room to an auditorium, with different damping. In all cases the source signal is placed at the coordinates $(2 + L/3, 2 + W/3, 1.5)m$, where L and W are the room length and width, while the microphones are placed in a circle centered at coordinates $(3 + L/3, 2 + W/3, 1.5)$, with radius 0.1 m. Although, in principle, many more degrees of freedom can be devised to increase the dataset size, this dataset was large enough to allow the DNN-FCP to be trained and to perform dereverberation on sweep inputs affected by room reverb. The total number of loudspeakers is 30, while the rooms and microphones are 3 and 8, respectively.

4.2. Training and Test

For the DNN-FCP, we used TF-GridNet [23, 24] as the DNN used to obtain $\hat{S}_{\text{DNN}}(t, f)$ in Eq. 4. TF-GridNet is a state-of-the-art model originally developed for speech separation and enhancement. Here we relied on the open-source implementation made available by the ESPNet-SE++ toolkit [25][2]. In our experiments we used the TF-GridNet configuration that offers the best trade-off between computational requirements and performance according

[1] pyroomacoustics.readthedocs.io/

[2] github.com/espnet/espnet/blob/master/espnet2/enh/separator/tfgridnetv2_separator.py

PARAMETER	MIN	MAX
LOUDSPEAKERS		
low-frequency rolloff range [Hz]	100	200
high-frequency rolloff range [Hz]	10k	20k
nr. of pole pairs	1	30
gain range [dB]	-6	+6
ROOMS and MICROPHONES		
room size [m]	7x6x2.5	21x18x7.5
RT60 [s]	0.3	0.9

Table 1: Parameters for synthetic dataset generation.

to [24], i.e. the row 7 model in Table XIII of [24], with the exception that here we used an STFT window size of 32 ms and 8 ms hop size. For the FCP algorithm we use 3 taps ($K = 3$ in [19]). TF-GridNet was trained with the Adam optimizer, a learning rate (η) of 10^{-3}, L2 norm gradient clipping of 1 and batch size 8. The η is halved if no improvement is observed for 5 epochs and early stopping is triggered if no improvement is encountered for 10 epochs. We used an NVIDIA A100 40GB GPU for the training, which took $\sim$ 6 h. Inference was instead performed on a laptop with a i7-8750H CPU.

The filters were designed using Kirkeby's method as implemented by the Aurora plugins [3]. The FIR filters have length $L = 2048$ at 48 kHz sampling rate, therefore the FIR group delay is 2.1 ms, which is low enough for real-time applications such as sound synthesis in a musical instrument. The regularization term is $\beta = 0.01$ for the range $50 - 20kHz$. The delay Δ for WPE is 3 frames, which is the most common choice in the literature.

4.3. Metrics

To evaluate the dereverberation algorithms we will employ several metrics, such as the PEMOQ and the SDR, along with traditional acoustic descriptors such as the T_{20}, i.e. the time for the signal to decrease by 20 dB.

PEMOQ is an index proposed in [26], where PEMO stands for PErceptual MOdel and Q stands for Quality assessment. The index was conceived for the objective assessment and prediction of perceived audio quality using an auditory model. At the time of its introduction it was shown to better predict human judgements than the more widely spread PEAQ index [27]. We selected PEMOQ over PEAQ for its ability to predict wider audio degradation ranges, i.e. from more severe to smaller impairments and for its recommendation for music signals in addition to speech signals. In this paper the PEMOQ scores are provided as real values in the range 0-1. To calculate the PemoQ metrics we adopted the freely available toolkit PEASS [4].

The SDR, or Signal-to-Distortion-Ratio, is a well-known metric used in various fields to objectively evaluate the validity of source separation algorithms[28]. Since blind dereverberation can be seen as a source separation task, where the target signal and the reverberation are meant to be identified and separated the use of SDR is licit and widely adopted. Finally, to evaluate the dereverberation capabilities of WPE and DNN-FCP we also employed the T_{20} in octave bands.

[3] pcfarina.eng.unipr.it/Aurora_XP/

[4] gitlab.inria.fr/bass-db/peass

5. RESULTS AND DISCUSSION

In the following we first evaluate the validity of the dereverberation of swept sines to estimate the loudspeaker IR, and then we provide results of the equalization of the dereverberated swept sines, showing that the inverse filters are close to the ones that would be computed on anechoic measurements.

5.1. Dereverberation Performance

To assess the validity of the dereverberation algorithm applied to swept sines we compare a set of samples from the dataset, not seen in the training set of the DNN-FCP. These are obtained from 3 synthetic loudspeakers, with 3 different rooms and 6 randomly picked microphones, for a total of 12 samples for each case. We first extracted the T_{20} decay time for each octave band, showing how the time gets reduced by WPE and DNN-FCP, as shown in Table 2. The data shows that the DNN-FCP is far superior than WPE in reducing the decay, especially at mid to high frequency. It must be observed that the WPE is bounded by its prediction delay, therefore the reverb tail cannot be shortened below its threshold. However, the prediction delay can be reduced if some assumptions on the geometry of the space and the microphone position can be made.

To evaluate the effectiveness of the dereverberation from an audio quality perspective we also use the PEMOQ, and SDR metrics [28]. The data is reported in Table 3 and shows the superior performance of the DNN-FCP. The PEMOQ score is obtained by comparing the audio to the anechoic source. A score of 1 means that two audio sources are identical. The SDR score is expressed in dB (the higher, the better).

Overall, these results show that the DNN-FCP method is far superior to WPE in removing the reverb tail from the output and - more importantly - that training a DNN-FCP network with a fairly small synthetic dataset of swept sines can make it work with swept sines coming from reverberated measurements. In the next section we will compare the two dereverberation methods for what concerns spectral coloration and the ability to be used in the proposed method to design equalizing filters.

5.2. Equalization and Filter Design

By converting the dereverberated log sweep signals to frequency responses we can obtain equalization filters based on the method from [9]. We first ran a test with the synthetic loudspeakers, as a proof-of-concept of the proposed method. The inverting FIR filters designed with Kirkeby's method are shown in Figure 5 for three different loudspeakers. As can be seen, the equalizing filters designed on the dereverberated IRs are similar to those designed on the anechoic measurement of the loudspeaker. However, depending on the example and frequency range, the WPE is sometimes closer to the reference curve than the DNN-FCP. On the other hand, the DNN-FCP presents a smoother magnitude frequency response.

Figure 6 shows third-octave band plots of the equalized signals, using the filters in Figure 5(a). As can be seen, in this case the spectral flatness of DNN-FCP is superior to that of WPE. All the spectra are obtained by convolution between the anechoic loudspeaker IR and the equalizing filter. Table 4 provides the MSE computed for all three loudspeakers from the test set. These results are computed by convolution between the equalizing filters

	bands (Hz)	**31.5**	**63**	**125**	**250**	**500**	**1000**	**2000**	**4000**	**8000**	**16000**	**average [s]**
REVERB	Room 1	0.76	0.58	0.31	0.39	0.36	0.33	0.30	0.30	0.33	0.32	**0.40**
	Room 2	1.02	1.02	0.73	0.60	0.64	0.65	0.61	0.62	0.62	0.63	**0.71**
	Room 3	1.37	1.26	0.59	0.92	0.96	0.93	0.80	0.87	0.91	0.88	**0.95**
WPE	Room 1	1.97	0.35	0.24	0.34	0.28	0.29	0.24	0.26	0.26	0.25	**0.45**
	Room 2	1.80	0.61	0.40	0.43	0.50	0.39	0.38	0.39	0.39	0.38	**0.57**
	Room 3	1.69	0.47	0.58	0.47	0.54	0.49	0.45	0.45	0.44	0.45	**0.60**
DNN-FCP	Room 1	0.53	0.23	0.95	0.07	0.05	0.03	0.01	0.01	0.01	0.00	**0.19**
	Room 2	0.44	0.23	0.12	0.06	0.03	0.02	0.01	0.01	0.01	0.00	**0.09**
	Room 3	0.44	0.23	0.12	0.06	0.03	0.02	0.01	0.01	0.01	0.00	**0.09**

Table 2: Average T20 decay times for octave-bands of the reverberated audio signals and the WPE- and DNN-FCP-dereverberated audio signals. The signals are divided by room (with Room 1 being the one with shortest decay, and Room 3 being the one with the longest decay). For each room 4 random combinations of loudspeaker and microphone from the dataset are averaged.

	REVERB	**WPE**	**DNN-FCP**
PEMOQ	0.81 (0.08)	0.81 (0.09)	0.97 (0.02)
SDR [dB]	7.97 (3.54)	8.69 (4.16)	17.29 (3.06)

Table 3: Average value of the PEMOQ and SDR (the standard deviation is reported in brackets).

	Anechoic	**NO-EQ**	**WPE**	**DNN-FCP**
Figure 5(a)	1.0	3.5	9.3	2.8
Figure 5(b)	1.0	25.9	23.0	17.5
Figure 5(c)	0.4	7.9	5.8	5.6

Table 4: MSE referred to third-octave spectra computed after equalization of the synthetic loudspeaker from Figures 5(a), (b), (c).

and the anechoic response of the loudspeaker, to simulate the flatness of the spectrum in absence of the room. As can be seen, the performance of the exact inverse computed using the method from Kirkeby cannot be matched, however the filters estimated after dereverberation provide an improvement of several dB over the case without equalization (except for one case with WPE).

5.3. Real-world use case

The proposed method is finally evaluated on the use case of the organ loudspeakers. The dereverberation has been carried out with either WPE or DNN-FCP, and the latter has shown to be significantly superior in terms of dereverberation (average $T_{20} = 0.08\,s$ for DNN-FCP vs. $0.37\,s$ for WPE). The inverting filters are shown in Figure 7. As can be seen, both WPE and DNN-FCP filters are close to the reference ones, computed on the anechoic measurement. This seems to confirm the validity of the approach. The MSE obtained for all approaches are: 15.1 dB, 16.5 dB, 19.9 dB for the anechoic, WPE and DNN-FCP cases, respectively. In this case, the DNN-FCP is penalized by a larger error in the low-frequency range.

6. CONCLUSIONS

This work proposed a method for equalizing loudspeakers by means of FIR filters in cases when the anechoic IR cannot be computed. A reverberated swept-sine signal is dereverberated using two alternative state-of-the-art methods, WPE and DNN-FCP, before computing the inverse filters. We show the idea to be viable using both methods (with some differences in the results) and the obtained filters show similar magnitude response to the ones computed from inversion of the anechoic loudspeakers IR. Our work also shows that training the Deep Neural Network used by DNN-FCP with a fairly small dataset of swept sines is sufficient to allow the method to be adapted to the dereverberation of swept-sine measurements.

This proof of concept is open to future improvements. In principle end-to-end methods for dereverberation and filter design can be envisioned. At the moment methods based on Deep Learning for the design of equalizing filters have been proposed, however they are not directly applicable to the problem addressed in this work [29]. Other methods could be envisioned that adapt to the environment and invert it by means of Differentiable Digital Signal Processing (DDSP) techniques [30]. In this case the techniques can be employed similarly to a system inversion problem using adaptive filters [31] removing the need for swept-sine measurements, but requiring a microphone always available during adaptation and at the position to be equalized.

7. ACKNOWLEDGMENTS

We would like to thank the staff at Viscount International for the data and the useful discussions.

8. REFERENCES

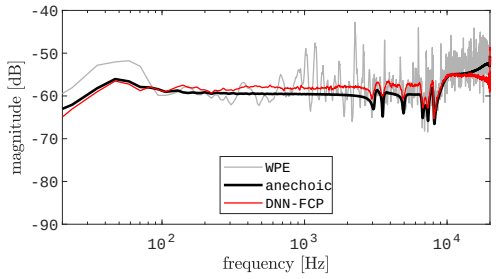
[1] Vesa Välimäki and Joshua D. Reiss, "All about audio equalization: Solutions and frontiers," *Applied Sciences*, vol. 6, no. 5, 2016.

[2] Richard Heyser, "Acoustical measurements by time delay spectrometry," *Journal of the Audio Engineering Society*, vol. 15, pp. 370–382, october 1967.

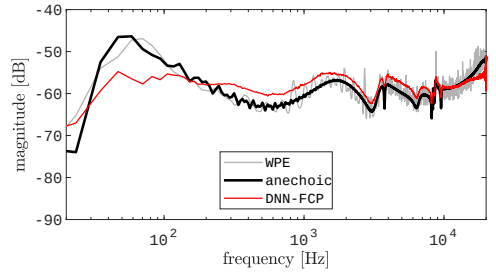
[3] Christopher J Struck and Steve F Temme, "Simulated free field measurements," *Journal of the Audio Engineering Society*, vol. 42, no. 6, pp. 467–482, 1994.

[4] Eric Benjamin, "Extending quasi-anechoic measurements to low frequencies," in *Audio Engineering Society Convention 117*, Oct 2004.

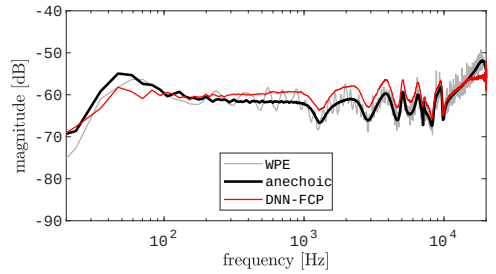
[5] Richard Stroud, "Quasi-anechoic loudspeaker measurement using notch equalization for impulse shortening," in *Audio Engineering Society Convention 129*, Nov 2010.



(a)

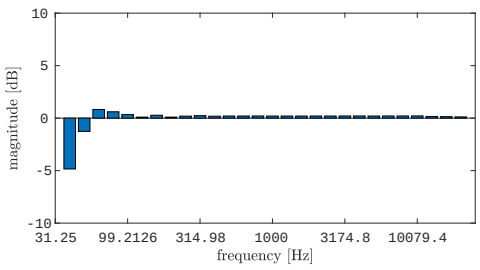


(b)

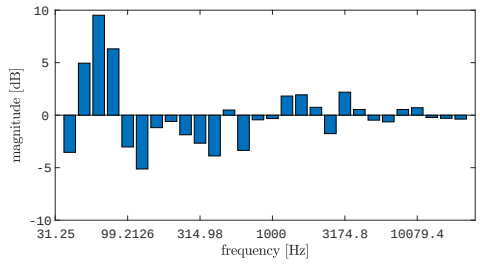


(c)

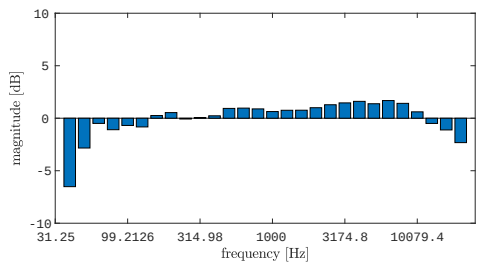
Figure 5: Magnitude frequency response of the FIR filters designed with the method from Kirkeby et al. for three of the synthetic loudspeakers in the database designed on: the anechoic IR (solid black line), the WPE output (gray), the DNN-FCP output (red).



(a)



(b)



(c)

Figure 6: Third-octave band spectra resulting after equalization using the filters shown in Figure 5(a), i.e. the filter based on the anechoic IR (a), the WPE output (b), the DNN-FCP output (c).

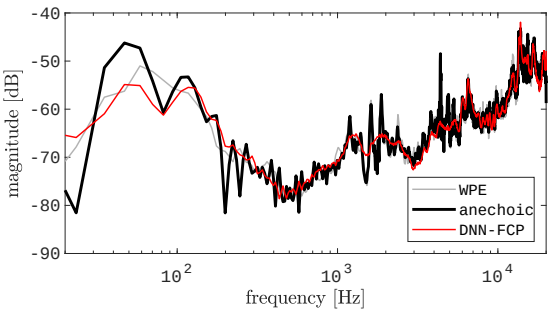


Figure 7: Equalized frequency response of one of the loudspeakers of the organ under test.

[6] Florian Denk, Birger Kollmeier, and Stephan D Ewert, "Removing reflections in semianechoic impulse responses by frequency-dependent truncation," *Journal of the Audio Engineering Society*, vol. 66, no. 3, pp. 146–153, 2018.

[7] Balázs Bank, "Combined quasi-anechoic and in-room equalization of loudspeaker responses," in *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[8] Angelo Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio engineering society convention 108*. Audio Engineering Society, 2000.

[9] Ole Kirkeby, Philip A. Nelson, Hareo Hamada, and Felipe Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 189–194, March 1998.

[10] Yuan Li and Lunhui Deng, "An overview of speech dereverberation," in *Proceedings of the 8th Conference on*

Sound and Music Technology: Selected Papers from CSMT. Springer, 2021, pp. 134–146.

[11] Emanuël A.P. Habets and Patrick A. Naylor, *Dereverberation*, chapter 15, pp. 317–343, John Wiley Sons, Ltd, 2018.

[12] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 85–88.

[13] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[14] Zhong-Qiu Wang and DeLiang Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 941–950, 2020.

[15] Yan Zhao, DeLiang Wang, Buye Xu, and Tao Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1598–1607, 2020.

[16] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] Keisuke Kinoshita, Marc Delcroix, Haeyong Kwon, Takuma Mori, and Tomohiro Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation.," in *Interspeech*, 2017, pp. 384–388.

[18] Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3476–3490, 2021.

[19] Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux, "Convolutive prediction for reverberant speech separation," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 56–60.

[20] Neville Thiele, "Loudspeakers in vented boxes: Part 1," *Journal of the Audio Engineering Society*, vol. 19, no. 5, pp. 382–392, 1971.

[21] Neville Thiele, "Loudspeakers in vented boxes: Part 2," *Journal of the Audio Engineering Society*, vol. 19, no. 6, pp. 471–483, 1971.

[22] Richard H Small, "Vented-box loudspeaker systems–part 1: Small-signal analysis," *Journal of the Audio Engineering Society*, vol. 21, no. 5, pp. 363–372, 1973.

[23] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[25] Yen-Ju Lu, Xuankai Chang, Chenda Li, Wangyou Zhang, Samuele Cornell, Zhaoheng Ni, Yoshiki Masuyama, Brian Yan, Robin Scheibler, Zhong-Qiu Wang, et al., "Espnet-se++: Speech enhancement for robust speech recognition, translation, and understanding," 2022.

[26] Rainer Huber and Birger Kollmeier, "Pemo-q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[27] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[28] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR – half-baked or well done?," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[29] Giovanni Pepe, Leonardo Gabrielli, Stefano Squartini, Carlo Tripodi, and Nicolò Strozzi, "Deep optimization of parametric iir filters for audio equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1136–1149, 2022.

[30] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts, "Ddsp: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020.

[31] Simon Haykin, "Adaptive filter theory," *Prentice Hall google schola*, vol. 2, pp. 67–94, 2002.

[32] Angelo Farina, "Advancements in impulse response measurements by sine sweeps," *Journal of the Audio Engineering Society*, May 2007.

[33] Vesa Välimäki and Jussi Rämö, "Neurally controlled graphic equalizer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2019.

[34] John Vanderkooy and Stanley Lipshitz, "Can one perform quasi-anechoic measurements in normal rooms?," in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[35] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[36] M Martinez Ramirez and Joshua Reiss, "End-to-end equalization with convolutional neural networks," in *21st International Conference on Digital Audio Effects - DAFx-2018, Aveiro, Portugal*, 2018.