Perceptually-Transparent Online Estimation of Two-Channel Room Transfer Function for Sound Calibration

Hai Morgenstern and Boaz Rafaely, *Senior Member, IEEE*

***Abstract*—Sound calibration is employed in many commercial audio systems for improving sound quality. This process includes the estimation of the room transfer function (RTF) between each loudspeaker and a microphone located at the listeners' position. Current methods for RTF estimation employ calibration signals, such as noise or tones, in a dedicated process applied to each loudspeaker separately. Such an estimation disrupts normal playback, is time consuming, and requires user intervention. A perceptually-transparent online RTF estimation method for a two-channel system, which employs calibration signals generated using the original audio signals and complementary filters, is proposed in this article. These calibration signals are uncorrelated across the two channels, which facilitates the online estimation of both channels using a single microphone. Estimation performance is investigated for an experimental system and displays low estimation errors. Finally, a subjective evaluation via a listening test shows that playback of calibration signals is perceptually-transparent under some of the conditions investigated.**

***Index Terms*—Sound calibration, system identification, transfer function estimation, two-channel audio systems, sound perception.**

I. Introduction

SOUND calibration for audio systems is a process in which the parameters of each channel or each loudspeaker are adjusted to improve sound quality. The parameters typically include frequency equalization, loudness level, and time-delays, and are adjusted based on the acoustic environment in the room and the location of the sound system. Multiple methods for sound calibration have been proposed, including perceptually motivated methods that exploit psychoacoustic properties of hearing [1]. Sound calibration usually requires the use of a microphone for estimating the room transfer function (RTF). Current methods for RTF estimation employ excitation, or calibration signals, to estimate the RTF in a dedicated process applied to each loudspeaker separately, and a microphone assumed to be located in the vicinity of the listeners. Excitation signals include maximum-length sequences, inverse-repeated sequences, sweep sines (linear and logarithmic), impulses, periodic impulse excitation, time-stretched pulses, random noise, pseudo-random noise, and periodic random excitation [2]–[10], while the variety of applications include sound calibration, video conferencing/multimedia communication systems, virtual and augmented reality, auralization, and spatialization of sounds. A comparison of methods, including guidelines for choosing a method for specific acoustic conditions, is provided in [7], [11]. Specifically for sound calibration, multiple patents that employ some of the above excitation signals have been awarded [12]–[15]. All the above methods employ a calibration signal, which makes the task of RTF estimation time-consuming, necessitates user intervention, and interrupts the normal use of the audio system.

There are also RTF estimation methods that employ non-dedicated source signals, which allows a more natural measurement of a system during playback [7]. In this case, the source signals should cover all frequencies of interest in the long term, and typically require longer averaging times to achieve good estimates compared to typical excitation signals. Applications typically include echo cancellation and source localization, wherein speech signals are used for estimating the RTFs [16], [17]. Lately, methods for RTF estimation and room parameter estimation that are based on learning algorithms and neural networks have been proposed. Methods for estimating acoustical parameters such as early decay time (EDT) and reverberation time (RT) using neural networks and have been proposed in [18], [19]. In [20], Gaussian processes regression is applied to find regularizers for obtaining RTF estimates using a regularized least square approach. The approach of learning a regularizing function that facilitates better RTF estimates has been extended in [21], where the regularizing function is data-driven and learned via an artificial neural network. Moreover, both parametric and kernel learning methods have been proposed for the identification of linear systems, which show improvement in the reconstruction of system impulse responses compared to standard methods [22], [23]. However, these methods are suited for a system with a single loudspeaker. For multi-channel audio systems, these methods could still be applied, but may require playback using one loudspeaker at a time, which again may interrupt the normal playback of the system. Even more recently, methods that use room images instead of audio signals were

Manuscript received May 15, 2019; revised October 7, 2019; accepted November 8, 2019. Date of publication November 22, 2019; date of current version December 24, 2019. This work was supported in part by the Samsung RUNWAY program #4, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Roland Badeau. *(Corresponding author: Hai Morgenstern.)*

The authors are with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: hai.morgenstern@gmail.com; br@bgu.ac.il).

Digital Object Identifier 10.1109/TASLP.2019.2955286

proposed to infer the geometrical properties of a room [24], [25]. While these methods are inherently perceptually-transparent acoustically, they may facilitate the prediction of room acoustic parameters such as RT and EDT. These methods do not seem to yet facilitate the estimation of a full RTF, which is addressed in this paper and may be required for sound calibration.

An RTF estimation process for multi-channel audio systems that more naturally integrates with the normal operation of the system and that is perceptually-transparent to listeners could be of great interest. Such a process could also be made automatic, without requiring user-intervention, and could therefore be applied when necessary to compensate for time-varying acoustic conditions. In this paper, RTF estimation for two-channel audio systems is developed under the following requirements: (i) a two-loudspeaker audio system with standard audio material played-back (e.g., speech and music), and, (ii) a microphone positioned at a point of interest, such as the vicinity of a listener. The aim is to estimate the RTF with minimal interruption as perceived by a listener. For example, playback with only one loudspeaker at a time to facilitate estimation could be considered as an interruption (see the listening experiment in Section VI), and similarly, playback of standard calibration signals such as sine sweeps instead of the normal audio would also be considered a major interruption.

The approach behind this method is inspired by the use of dichotic stimuli in hearing aids, where spectral splitting with complementary filter banks is applied to produce signals for two channels [26], [27]. In a similar manner to the methods used for these hearing devices, as an initial step, a pre-processing method employs complementary filter banks to produce estimation signals (referred to as calibration signals) for the two loudspeakers. Due to the use of complementary filter banks, the processed calibration signals become uncorrelated across the two channels and can be successfully used for simultaneous estimation of the two-channel RTF using a single microphone. A post-processing method at the microphone first applies standard single-channel methods to estimate multiple functions, which are then combined to construct the two-channel RTF. A very important property of the proposed method is that it is applied online to the system, without interrupting its normal operation and is such a way that is perceptually similar to playback of the original, unprocessed audio signals. This could be explained by the phenomenon of spectral fusion [28].

The paper is organized as follows. In Section II a model for a system with two loudspeakers and a single microphone positioned in a room is presented, followed by a brief description of a common method for RTF estimation, dual-channel fast Fourier transform (FFT) analysis, in Section III. Original work starts from Section IV, where a pre-processing method for producing calibration signals using complementary filter banks is initially presented, followed by a description of a post-processing method for estimating the RTFs using the calibration signals recorded by the microphone. An investigation of RTF estimation performance for measured RTFs is presented in Section V, including an example implementation using 1/2-octave band finite impulse response (FIR) filters that were designed for the problem at hand. Then a subjective evaluation of the proposed method is presented in Section VI, showing how playback of calibration signals compares with playback of the original unprocessed signals with respect to perception.

II. System Model

The system model comprises two loudspeakers (left and right channels) positioned in a room and a single microphone, assumed to be positioned in the vicinity of a listener. The signal measured by the microphone, $y(f)$, is given by:

$$y(f) = H_L(f)\, s_L(f) + H_R(f)\, s_R(f), \tag{1}$$

where f denotes frequency, $s_L(f)$ and $s_R(f)$ are the input signals of the left and right channels, respectively, and $H_L(f)$ and $H_R(f)$ are the RTFs of the left and right loudspeakers, respectively. For mono playback, $s_L(f) = s_R(f) = s(f)$. Finally, note that in practice a noise component that can model acoustic, modelling, and transducer noise may be added to $y(f)$. However, this noise component is not added to the equations throughout the paper for simplicity, and since it is not required for the development of the proposed method.

III. Current Methods for RTF Estimation

A current method for RTF estimation, dual-channel FFT analysis [29], is presented as background. This method has been chosen since it works with various calibration signals, including real-world audio signals, which is the area of interest in this paper. Consider, for simplicity, a single-channel system, which will be extended below to the two-channel case. Following the notation of Eq. (1), the single-channel system equation is given by:

$$y(f) = H(f)\, s(f). \tag{2}$$

As detailed at the end of this section, the estimated RTF is calculated using the auto spectra of the input signal, and the cross spectrum between the input signal and the output signal. The auto spectra of $s(f)$ is defined as:

$$S_{ss}(f) = E[s^*(f)s(f)], \tag{3}$$

where $E[\cdot]$ is the statistical expectation and $(\cdot)^*$ denotes the complex conjugate. In practice, the statistical expectation is approximated by averaging spectra over multiple time blocks, as described in Section V. The cross spectrum is given by:

$$S_{sy}(f) = E[s^*(f)y(f)], \tag{4}$$

and is approximated in practice using time averaging, as in the case of the auto spectra. Given the spectra, there are several ways to estimate the RTF, $H(f)$, with the specific implementation chosen according to where noise is added in the processing chain. One estimate of the RTF, $\hat{H}(f)$, is given by:

$$\hat{H}(f) = \frac{S_{sy}(f)}{S_{ss}(f)}, \tag{5}$$

and is derived based on a noise-reduction criterion and under the assumption that noise is added to the output signal, i.e., the microphone [30].

This section presented RTF estimation when a single loudspeaker is employed. When both loudspeakers are employed, the output $y(f)$ will include contributions from both loudspeakers. For demonstrating this, we employ the notation of the left and right RTFs and input signals as in Eq. (1); in particular, $H_L(f)$ is to be estimated, but when both $s_L(f)$ and $s_R(f)$ are non-zero. Including the contribution of a non-zero $s_R(f)$ in $y(f)$, as in Eq. (1), the estimated RTF from Eq. (5) is now denoted as $\hat{H}_L(f)$, and is given as:

$$\hat{H}_L(f) = H_L(f) + H_R(f)\frac{S_{s_L s_R}(f)}{S_{s_L s_L}(f)}, \tag{6}$$

where $S_{s_L s_R}(f)$ is the cross spectrum between the left and right input signals. This demonstrates that when signals of both channels are simultaneously employed, there is an error in the estimation if the left and right signals are correlated. If $s_L(f)$ and $s_R(f)$ are uncorrelated, $S_{s_L s_R}(f)$ from the last equation is zero, and the RTF could be estimated using the single output signal, $y(f)$. This is the main principle behind the proposed method presented in the following sections, where decorrelation is achieved by separation in frequency.

IV. Proposed Method for RTF Estimation

The proposed estimation methods includes a pre-processing method for producing calibration signals, and a post-processing method that estimates RTFs, as described in this section.

A. Pre-Processing - Calibration Signals

A method for producing two uncorrelated signals given a two-channel input signal is proposed; the method employs a filter bank and a complementary filter bank. At this stage, the filter banks are presented in a generic manner. Various sets of center frequencies, filter bandwidths, number of bands, and filter implementations can be considered, with an example implementation provided in Section V.

To construct the filter banks, I center frequencies, $f_i, i = 0, \ldots I-1$, are initially selected, with their values restricted to the range between zero and half of the sampling frequency, $f_s/2$. For each center frequency, a corresponding frequency band, O_i, is defined. For each center frequency and frequency band, a corresponding filter, $B_i(f)$, is constructed.

Filters for the left and right loudspeakers are constructed next. The filter for the left loudspeaker is constructed as a summation of filters for even center frequency indices:

$$G_L(f) = \sum_{i=0,2,4,\ldots,I-2} B_i(f). \tag{7}$$

Similarly, a filter for the right loudspeaker is comprised of a summation of filters for odd indices:

$$G_R(f) = \sum_{i=1,3,5,\ldots,I-1} B_i(f), \tag{8}$$

where it has been assumed, for simplicity, that I is even. Next, two signals are generated as loudspeaker input signals given the two-channel input signal and $G_L(f)$ and $G_R(f)$:

$$l(f) = G_L(f)\, s_L(f), \text{ and} \tag{9}$$

$$r(f) = G_R(f)\, s_R(f), \tag{10}$$

for left and right loudspeakers, respectively. The cross-spectrum between the two signals is given by:

$$E[l^*(f)r(f)] = G_L^*(f)G_R(f)\, E[s_L^*(f)s_R(f)]. \tag{11}$$

Now, since $s_L(f)$ and $s_R(f)$ cannot be assumed to be uncorrelated, as they could be identical for a mono signal, or highly correlated for a stereo signal, achieving decorrelation between these signals requires:

$$G_L^*(f)G_R(f), \ \ \forall f \in [0, f_s/2]. \tag{12}$$

Substituting Eqs. (7) and (8) in Eq. (12), the condition in Eq. (12) can be reformulated as:

$$B_i^*(f)\, B_m(f) = 0, \ \ \forall i \text{ even and } m \text{ odd}. \tag{13}$$

This is a sufficient condition to ensure zero correlation if the filters are equal to zero outside their bandwidth, or low correlation if they have a small gain outside their bandwidth. The proposed de-correlated calibration signals form the basis for RTF estimation.

B. Post-Processing - RTF Estimation With Proposed Calibration Signals

As an initial step, RTF estimation is described in this section for the case of ideal filters, with the following assumptions:

i) Filters $B_i(f)$ used for generating the calibration signals are ideal, with:

$$B_i(f) = \begin{cases} 1, \ \forall f \in O_i(f) \\ 0, \ \forall f \notin O_i(f). \end{cases} \tag{14}$$

ii) The frequency bands used for defining these filters do not overlap, i.e.:

$$O_i \cap O_m = 0, \ \ \forall i \neq m. \tag{15}$$

iii) The union of all frequency bands covers the entire frequency range, i.e.:

$$\cup_{i=0}^{I-1} O_i = [0, f_s/2]. \tag{16}$$

From Eqs. (14) and (15), i.e. ideal filter with non-overlapping bands, it is clear that Eq. (13) is satisfied, and the correlation of the calibration signals as in Eq. (11) becomes zero. Next, RTF estimation is developed by defining two sets of calibration signals:

$$l_1(f) = G_L(f)\, s_L(f), r_1(f) = G_R(f)\, s_R(f), \tag{17}$$

and

$$l_2(f) = G_R(f)\, s_L(f), r_2(f) = G_L(f)\, s_R(f). \tag{18}$$

The two sets are defined as in Eqs. (9) and (10), but with interchanged filters between left and right channels. The two filters sets are employed at different system operating times. Due to the ideal filters which satisfy Eq. (12), the correlation of the calibration signals as in Eq. (11) is zero, i.e. $E[l_1^*(f)r_1(f)] = E[l_2^*(f)r_2(f)] = 0$.

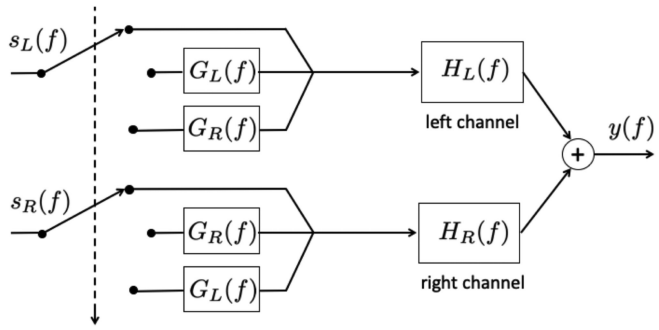

Fig. 1. Block diagram illustrating the system for three different states: (1) normal playback (no calibration, Eq. (1)), (2) calibration signals set 1 (Eq. (17)) with $y(f) = y_1(f)$ from Eq. (19); and, (3) calibration signals set 2 (Eq. (18)) with $y(f) = y_2(f)$ from Eq. (20).

The calibration signals are played back by the loudspeakers, instead of playback of the original audio signals, as in Eq. (1), and are recorded by the microphone. The new system model in this case becomes:

$$y_1(f) = H_L(f)\, l_1(f) + H_R(f)\, r_1(f), \tag{19}$$

$$y_2(f) = H_L(f)\, l_2(f) + H_R(f)\, r_2(f). \tag{20}$$

A system block diagram is shown in Fig. 1. In the diagram, the processing of the left and right input signals before playback is controlled by a single switch with three different states: (1) normal playback (Eq. (1)), (2) calibration signals set 1 (Eq. (19)), and, (3) calibration signals set 2 (Eq. (20)).

RTFs are estimated using the measured signals in Eqs. (19) and (20). Initially, RTFs between these measured signals and the calibration signals from Eqs. (17) and (18) are estimated. For $y_1(f)$, for example, $H_L^1(f)$ is defined similarly to Eq. (5), using:

$$H_L^1(f) = \frac{S_{l_1 y_1}}{S_{l_1 l_1}} = \frac{E\left[l_1^*(f)\, y_1(f)\right]}{E\left[l_1^*(f)\, l_1(f)\right]} \tag{21}$$

$$= \frac{H_L(f)\, E\left[\left|l_1(f)\right|^2\right] + H_R(f)\, E\left[l_1^*(f) r_1(f)\right]}{E\left[\left|l_1(f)\right|^2\right]}$$

$$= H_L(f),\ \forall f \in O_i,\ \text{ for even } i, \tag{22}$$

where the last line was derived using Eq. (11) and the conclusion derived above that the correlation is zero for the case of ideal filters. Note that in order to avoid a zero denominator due to the construction of $r_1(f)$ in Eq. (17), only frequency bands of even indexes are considered. Repeating the same estimation process for $y_2(f)$ and $l_2(f)$, $y_2(f)$ and $r_2(f)$, and $y_1(f)$ and $r_1(f)$, will lead to estimation of the two RTFs for the entire frequency range, due to Eq. (16):

$$\hat{H}_L(f) = \begin{cases} H_L^1(f),\ \forall f \in \{O_i(f)\}_{i=0,2,4,\ldots,I-2} \\ H_L^2(f),\ \forall f \in \{O_i(f)\}_{i=1,3,5,\ldots,I-1}, \end{cases} \tag{23}$$

and

$$\hat{H}_R(f) = \begin{cases} H_R^2(f),\ \forall f \in \{O_i(f)\}_{i=0,2,4,\ldots,I-2} \\ H_R^1(f),\ \forall f \in \{O_i(f)\}_{i=1,3,5,\ldots,I-1}. \end{cases} \tag{24}$$

Note that while this estimation process with ideal filters leads to zero estimation error, filters in practice are never ideal.

RTF estimation is described now for the case of practical filters, with the following assumptions:

i) Filters $B_i(f)$ used for generating the calibration signals include a pass band, $O_i^{pass}(f)$, and a stop band, $O_i^{stop}(f)$, and are defined in a generic manner as:

$$B_i(f) \approx 1,\ \forall f \in O_i^{pass}(f),$$
$$|B_i(f)| < \epsilon,\ \forall f \in O_i^{stop}(f), \tag{25}$$

where ϵ is a predefined threshold. It is important to note that $O_i^{pass}(f)$ and $O_i^{stop}(f)$ do not cover the entire frequency range; there are frequencies in the transition between these two frequency bands.

ii) The pass bands used for defining these filters do not overlap, i.e.:

$$O_i^{pass} \cap O_m^{pass} = 0,\ \forall i \neq m. \tag{26}$$

iii) The union of all pass bands covers the entire frequency range, i.e.:

$$\cup_{i=0}^{I-1} O_i^{pass} = [0, f_s/2]. \tag{27}$$

Under these assumptions, the derivation of the estimation method for ideal filters is repeated with the following differences. $B_i^*(f), B_j(f) = 0$ from Eq. (13) is now replaced by $|B_i^*(f) B_j(f)| < \epsilon$ at $i, j \in S$, where S is a subset of bands which are sufficiently distant from one another such that any one pass band overlaps only with stop bands from the other bands. This leads to Eq. (12) being replaced by $|G_L^*(f) G_R(f)| < \epsilon$, and $E[l^*(f) r(f)] < \epsilon E[s_L(f) s_R(f)]$ replacing Eq. (11) for frequency bands in S. This leads to a small correlation between signals $l(f)$ and $r(f)$. RTF estimation is repeated for more band subsets, until the entire frequency range is covered. This process is demonstrated in the next section for a specific filter implementation, and can be extended to multiple sets of calibration signals.

V. Analysis of RTF Estimation

An analysis of RTF estimation using the proposed method is presented for an experimental system comprising a stereo loudspeaker system and a single microphone positioned in a room, for multiple audio signals, and for example filter banks designed for the proposed estimation method. The analysis includes an investigation of the frequency-dependent error between estimated and measured RTFs, the total error, and errors in room acoustic parameters calculated using the RTFs, the RT and the EDT.

A. Setup

The analysis was conducted for two different environments, in order to test its performance under various acoustic conditions.

TABLE I
SIX ACOUSTIC CONDITIONS OF TWO-CHANNELS RTFS FOR ENVIRONMENT 2

Conditions	RT [ms]	r [m]
1	160	1
2	360	1
3	610	1
4	160	2
5	360	2
6	610	2

Environment 1-Meeting Room at Ben-Gurion University: The first set of RTFs was measured in a meeting room at Ben-Gurion University of the Negev, with dimensions of (7.2, 6.4, 3.1) m, an approximate volume of 143 m^3, and an RT of about 0.5 s. The RTFs were measured using a system comprised of two KRK ROKIT 6″ loudspeakers and a Bruel & Kjaer 1/2-inch diffuse-field microphone, which were connected via an ESI U24XL soundcard to a laptop. The RT was calculated from measured RIRs using the Schroeder backward integration [31]. The left and right loudspeakers were positioned at (0.3, 1.6, 0.7) m and (0.3, 0.8, 0.7) m, with a distance of 0.8 m between the loudspeakers. The system configuration was chosen since it models a two-channel setup such as a stereo TV. The microphone was positioned at (2.8, 1.2, 1.05) m, which is 2.5 m away from the middle point connecting the loudspeakers. The system was placed slightly towards the room corner due to a large meeting table positioned in the room.

Environment 2-Speech & Acoustic Lab at Bar-Ilan University: The method was applied also to multiple RTFs from the database described in Ref. [32]. In particular, six different RTFs from this database are used in this analysis, which correspond to three different RTs of 160, 360, and 610 ms, and can model an office and meeting rooms. The configuration in Ref. [32], involves a linear microphone array and a array of loudspeakers mounted on two half circles with a radius, r, of 1 and 2 m, around the center of the microphone array. Out of this configuration, the setup employed the fourth array microphone, positioned close to the origin of the circle, and of the loudspeakers positioned at angles -60 and -45° for a radius of 1 m, and -30 and 15° for a radius of 2 m. See Ref. [32] for a diagram of this setup. The specific loudspeakers were chosen since the distances between the loudspeakers are similar to configuration 1, and to practical systems such as TV stereo. Overall, the six two-channel RTFs correspond to six acoustic conditions, which are presented in Table I.

For both environments, the analysis is provided for various stereo input signals that include white noise, speech, and music. Three different 20-second duration stereo audio signals were used as input signals:

1) white (uniform) noise,
2) jazz music (taken from [33]),
3) pop music (taken from [34]), and
4) speech (taken from [35]).

B. Methods

The methods presented in this section are organized as follows. First, an implementation of filter banks designed for the task in hand is presented. Then, the measurement procedure of the RTFs and associated pre-processing of these functions is described. Finally, the application of the method to these functions is presented, including the derivation of the errors.

TABLE II
CENTER FREQUENCIES FOR TWO SETS OF 1/2-OCTAVE FILTER BANKS

Octave index	Set 1(f_i) [Hz]	Set 2 ($\tilde{f}_i$) [Hz]
0	75	89
1	106	126
2	150	178
3	211	251
4	300	355
5	422	501
6	596	708
7	841	1001
8	1189	1413
9	1679	1996
10	2371	2820
11	3350	3983
12	4732	5627
13	6683	7948
14	9441	11227
15	13335	–

The example design of filter banks for the proposed estimation method is based on 1/2-octave filters with discrete-Fourier-transform (DFT) based implementation. Two sets of filter banks are employed, following the general framework presented for practical filters in the previous section. $I = 16$ center frequencies are selected for the first set of filter banks, and $I = 15$ center frequencies are selected for the second set of filter banks, with the sampling frequency, f_s, set to 44100 Hz. The center frequencies are presented in Table II for both sets. Note that the center frequencies of the set 2 are shifted by a 1/4 octave compared to those in set 1. The reason for this choice will be apparent shortly. For both sets, and for each center frequency, a corresponding frequency band, O_i, is defined as follows:

$$O_i := [2^{-1/4} f_i,\ 2^{1/4} f_i),\ i = 0, 1, \ldots, I-1, \tag{28}$$

where O_0 is modified to start at 0 Hz and O_{I-1} is modified to end at $f_s/2$. For both sets and for each frequency band, a corresponding filter, $b_i[t]$, $i = 0, \ldots I-1$, where t is the discrete time index, is constructed using a DFT-based implementation as [36]:

$$b_i[t] = \sum_{q:=f_s\, q/N \in O_i} e^{j2\pi qt/N},\ t = 0, \ldots, N-1, \tag{29}$$

where $N = 7055$ is the DFT length used for constructing these filters, which, combined with f_s, yield a time duration of 160-ms, and $q = 0, \ldots, N-1$ is the DFT frequency index. In particular, DFT frequency indices that correspond to negative frequencies are also included in the summation in Eq. (29), so that the filters are real in the time domain. The filters in the frequency domain, $B_i(f)$, are calculated by applying a DFT on $b_i[t]$. Then, $G_L(f)$ and $G_R(f)$ are constructed as in Eqs. (7) and (8), respectively, using filters $B_i(f)$.

As an illustrative example, filter $B_4(f)$ from set 1 is presented in Fig. 2 for a DFT of length 5 N for improved frequency

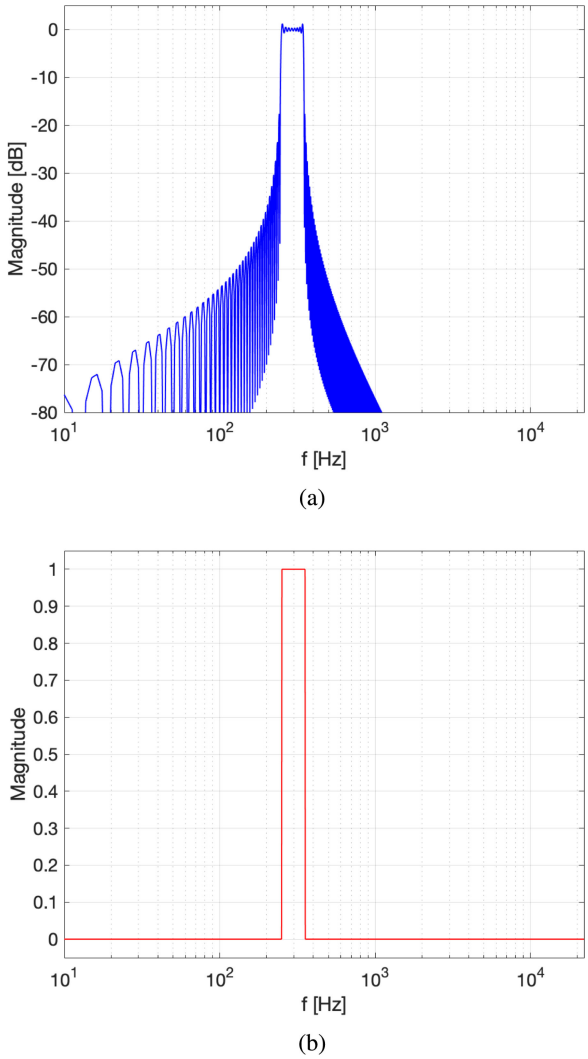

Fig. 2. $B_4(f)$ (set 1) for: (a) 1/2-octave filters with DFT-based implementation (logarithmic y-axis), and (b) ideal filters (linear y-axis).

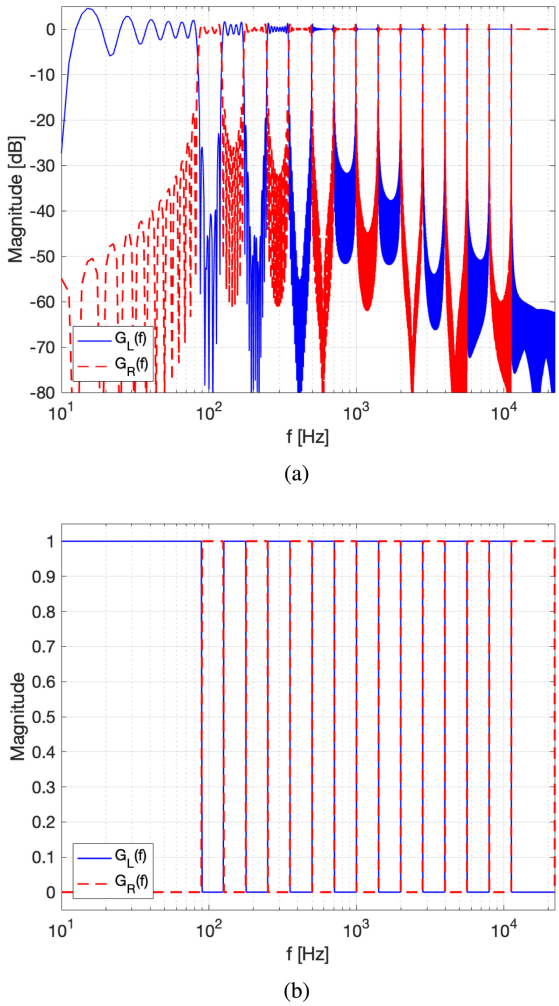

Fig. 3. $G_L(f)$ and $G_R(f)$ (set 1) for: (a) 1/2-octave filters with DFT-based implementation (logarithmic y-axis), and (b) ideal filters (linear y-axis).

resolution. A magnitude close to unity is evident at the corresponding octave band, while a relatively high attenuation is observed away from the pass band. A logarithmic y-axis is used in this figure. An ideal filter corresponding to $B_4(f)$ is also presented, but with a linear y-axis. The remaining filters, $B_i(f)$, show similar behavior. Filters $G_L(f)$ and $G_R(f)$ for set 1 are presented in Fig. 3 for a DFT of length $5\,N$. Due to the construction using filters from Eq. (29), $G_L(f)$ and $G_R(f)$ sum to unity at the original DFT frequencies, and the inverse DFT of the summation of these functions gives an impulse function, which was a consideration for this specific implementation, as discussed in the next section. $G_L(f)$ and $G_R(f)$ constructed using ideal filter are also illustrated using a linear y-axis scale. Filters from set 2 are not presented as they are similar in behavior. In Fig. 4, $|G_L^*(f)G_R(f)|$, which is proportional to the correlation (c.f. Eq. (11)), is presented for both sets. High values of $|G_L^*(f)G_R(f)|$ are evident for both sets at frequencies in the transition between the frequency bands. At these frequencies, correlation is expected to be higher, and this may degrade RTF estimation in Eq. (21). However, at frequencies where $|G_L^*(f)G_R(f)|$ is high for set 1, it is low for set 2 and vice versa. Fig. 5 presents the pointwise minimum of $|G_L^*(f)G_R(f)|$ between both sets, which can be considered as the effective $|G_L^*(f)G_R(f)|$ for both sets. The pointwise minimum of $|G_L^*(f)G_R(f)|$ for both sets is lower than approximately -20 dB at all frequencies, which can facilitate improved estimation using two sets compared to estimation using a single set, as demonstrated in the next section. This is the reason that the filters are divided into two sets in the manner described in Table II.

Next, the mapping of frequencies that have minimum $|G_L^*(f)G_R(f)|$ for each set can be used for constructing estimated RTFs by merging the estimates from each set; at frequencies where $|G_L^*(f)G_R(f)|$ of set 1 is minimal, RTFs can be estimated as described in Section IV for signals corresponding to set 1, and, similarly, at frequencies where $|G_L^*(f)G_R(f)|$ of set 2 is minimal RTFs can be estimated using signals corresponding to set 2. Other methods for merging RTF estimates of multiple

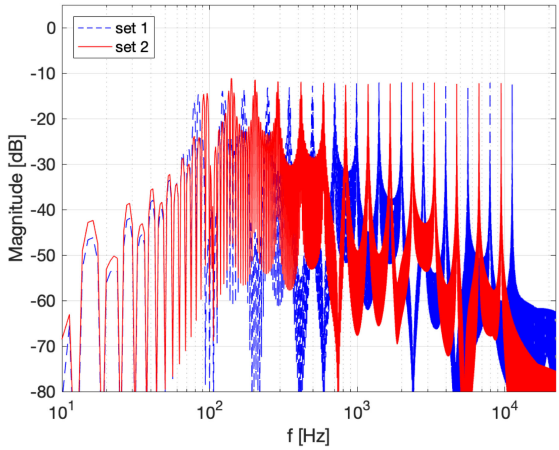

Fig. 4. $|G_L^*(f)G_R(f)|$ for 1/2-octave filters with DFT-based implementation for both sets of filters with center frequencies given in Table II.

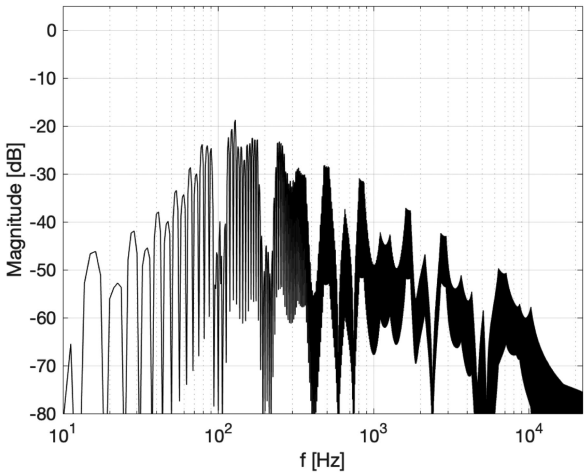

Fig. 5. Pointwise minimum of $|G_L^*(f)G_R(f)|$ for both sets.

sets can be considered, such as a fixed mapping using the center frequencies of each set. Finally, note that the curves that correspond to ideal filters cannot be seen in Figs. 4 and 5, as $G_L^*(f)\,G_R(f) = 0\ \forall f$.

For environment 1, the RTFs of the two channels were measured as follows. Each loudspeaker (separately) played back a 20-second white-noise signal. The signal recorded by the microphone along with the input signal of the loudspeakers were recorded by the computer. RTFs from the two loudspeakers to the microphone position were then computed using these signals and MATLAB's *tfestimate* function, with a 1-second duration window and an overlap factor of 2. For each transfer function, a corresponding room impulse response (RIR) was calculated using an inverse DFT, and these are denoted $h_L[t]$ and $h_R[t]$ for the left and right channels, respectively. The system buffer, or time delay, due to all hardware components including the sound card was compensated for. For environment 2, RIRs from Table I were initially downloaded (see Ref. [32]), and then resampled to meet the sampling rate of the filters. The resampled RIRs were truncated to a 1-second duration, and corresponding RTFs were then calculated.

To investigate estimation using the proposed method, signals $y_1(f)$ and $y_2(f)$ as in Eqs. (19) and (20), respectively, were generated in a computer simulation using the measured RTFs and calibration signals generated using the filters presented in this section. Noise was generated using a zero-mean Gaussian distribution and added to the output signals to model measurement noise. The variance of the noise was set equal over frequency to produce a 40 dB average signal-to-noise ratio (SNR). $H_L^1(f)$, $H_L^2(f)$, $H_R^1(f)$, and $H_R^2(f)$ were calculated using MATLAB's *tfestimate* function. As an initial step, MATLAB's *tfestimate* function computes the auto and cross spectra using estimates of the statistical expectation, which are calculated by averaging spectra over multiple time blocks. In particular, all of the signals are first split up into B segments with an overlap factor of 1/2, which are then used for approximating the expectation. E.g., the auto-spectra of $s_L(f)$ is approximated as:

$$S_{s_L s_L(f)} = \sum_{n=1}^{B} s_{L,n}^*(f) s_{L,n}^*(f), \tag{30}$$

where n is the time-block discrete index, and $s_{L,n}^*(f)$ is the input signal at block n; moreover, due to the overlap factor of 1/2, all neighboring segments of a signal share half of the samples. Furthermore, the block time duration is assumed to be sufficiently long compared to the RTF, $H(f)$, in time, as required by the multiplicative transfer function assumption [37]. Finally, a Hamming window function was applied to each time block in practice for a modified periodogram with desired frequency characteristics [38]. Given the auto and cross spectra, $H_L^1(f)$, $H_L^2(f)$, $H_R^1(f)$, and $H_R^2(f)$ were calculated for both sets as in Eq. (21) Estimated left and right RTFs, $\hat{H}_L(f)$ and $\hat{H}_R(f)$, respectively, were then estimated twice. First, $\hat{H}_L(f)$ and $\hat{H}_R(f)$ were estimated for set 1 only, as in Eqs. (23) and (24). Then, $\hat{H}_L(f)$ and $\hat{H}_R(f)$ were estimated for both sets, mapping frequencies that have a pointwise minimum of $|G_L^*(f)G_R(f)|$ of both sets (c.f. Fig. 5 and the discussion that follows).

A frequency dependent error was defined as $|H_L(f) - \hat{H}_L(f)|$ and $|H_R(f) - \hat{H}_R(f)|$, for the left and right channels, respectively. A normalized error was calculated as $|H_L(f) - \hat{H}_L(f)|/|H_L(f)|$ and $|H_R(f) - \hat{H}_R(f)|/|H_R(f)|$ for the two channels, respectively, and was averaged in 1/3 octaves so as to avoid ill-conditioning at frequencies at which the RTFs magnitudes are low. The total error was calculated as $||H_L(f) - \hat{H}_L(f)||/||H_L(f)||$ and $||H_R(f) - \hat{H}_R(f)||/||H_R(f)||$ for the left and right loudspeakers, respectively, where $||\cdot||$ denotes the 2-norm.

To evaluate errors in RT and EDT, measured and estimated RIRs were initially calculated for the two channels by applying an inverse FFT on the corresponding RTFs. Measured RIRs are denoted as $h_L[t]$ and $h_R[t]$, and estimated RIRs are denoted $\hat{h}_L[t]$ and $\hat{h}_R[t]$ for the left and right channels, respectively. RTs were calculated for the estimtated RIRs using the Schroeder backward integration [31], as the RTs for the measured responses. Errors in RT were calculated as the absolute value of the difference between these estimated RTs and the measured RT. The RT and EDT were calculated using a 20-dB and a 10-dB dynamic range, respectively, see Ref. [31] for more detail.

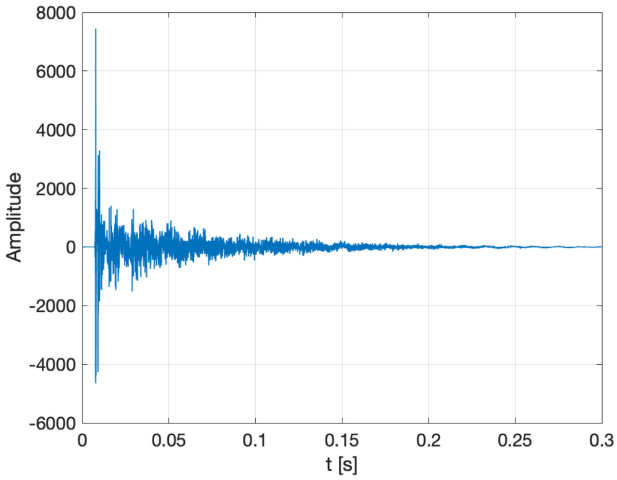

Fig. 6. Measured RIR for left channel ($h_L[t]$).

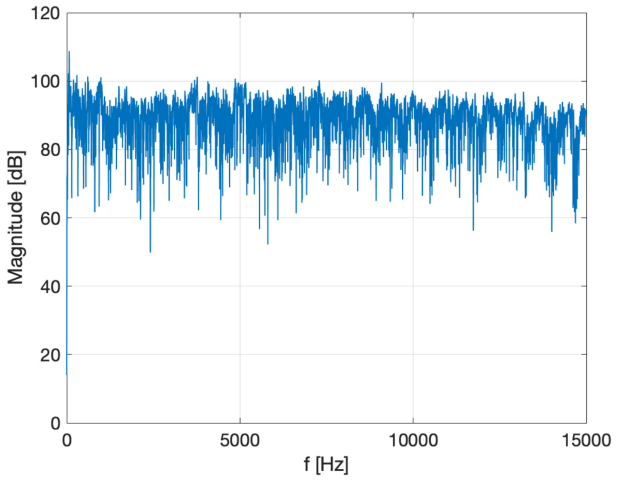

Fig. 7. Measured RTF for left channel ($H_L(f)$).

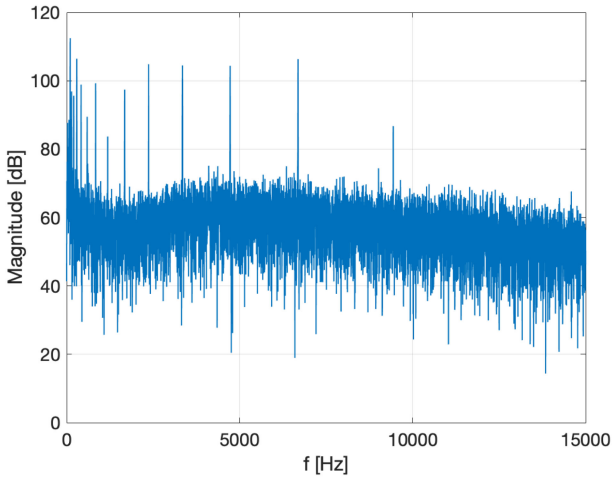

Fig. 8. Difference between measured RTF and RTF estimated using set 1.

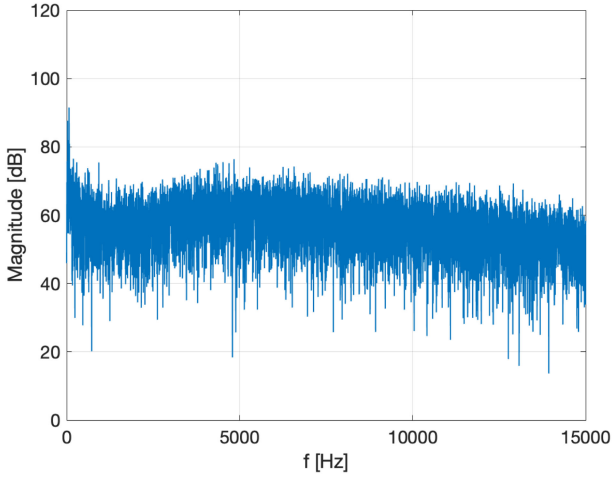

Fig. 9. Difference between measured RTF and RTF estimated using both sets.

C. Results

In this section the frequency-dependent error is initially analyzed for the RTFs measured in environment 1. Then, errors in frequency-independent measures, the total error, the RT, and the EDT are presented for both environments.

$h_L[t]$ and $H_L(f)$ measured in environment 1 are presented in Figs. 6 and 7, respectively. The frequency dependent error for the left channel, $|H_L(f) - \hat{H}_L(f)|$, estimated using set 1 only is presented in Fig. 8. The difference signal shows high errors at frequencies near the transition between the pass and stop bands of the filters used to construct $G_L(f)$ and $G_R(f)$. Similar results are evident for the right channel RTF and for calibration signals generated for the remaining input signals, and are thus not presented. In Fig. 9, the difference between the measured RTF and the RTF estimated using both sets is presented. The difference signal in this case is 20 dB lower than the measured RTF at most frequencies, including the transitions between the frequency bands. The normalized error averaged in 1/3 octaves is presented in Fig. 10 for all the input signals. The figure shows that the maximum error for these examples is less than -15 dB,

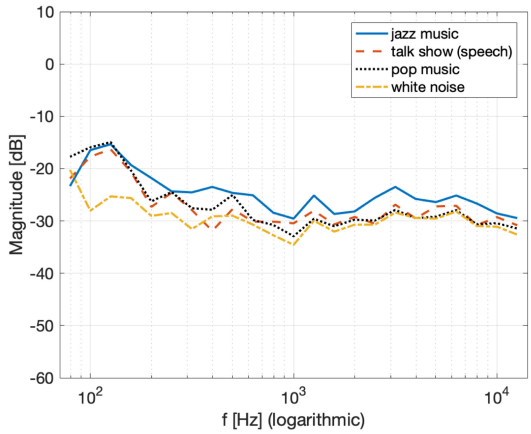

Fig. 10. Normalized errors averaged in 1/3 octaves.

but is typically significantly lower. The total estimation error for the left channel, the error in the RT, and the error in the EDT are presented in Table III for all input signals. Finally, the mean and variance of the total estimation error for the left channel, the error in the RT, and the error in the EDT over the six conditions

TABLE III
TOTAL ESTIMATION ERROR, THE ERROR IN RT, AND THE ERROR IN EDT OVER THE FOUR INPUT SIGNALS FOR ENVIRONMENT 1

Input signal	Total error [dB]	Error in RT [%] (RT = 0.5 s)	Error in EDT [%]
White noise	-25.8	0.5	4.4
Jazz music	-28.0	7.6	6.0
Pop music	-27.5	5.0	4.8
Speech	-29.5	9.7	6.9

TABLE IV
AVERAGE TOTAL ESTIMATION ERROR (ROOT MEAN SQUARE ERROR), AVERAGE ERROR IN RT, AND AVERAGE ERROR IN EDT OVER THE SIX CONDITIONS AND THE FOUR INPUT SIGNALS CHOSEN FROM THE DATABASE OF ENVIRONMENT 2, SEE TABLE I

Input signal	Total error [dB]	Error in RT [%] (RT = 0.5 s)	Error in EDT [%]
White noise	-28.1	3.3	2.5
Jazz music	-31.1	3.6	2.8
Pop music	-31.8	3.7	2.8
Speech	-32.7	4.6	3.3

chosen for environment 2 are presented in Table III for all input signals.

In conclusion, frequency dependent errors, total (frequency-independent) errors, and errors in room acoustic parameters have been presented in this section for multiple acoustic conditions. These errors are reasonably low, demonstrating the effectiveness of the proposed estimation method.

VI. LISTENING TEST

A listening test was performed with the aim of evaluating the perceptual-transparency of the calibration signals when played back using the system loudspeakers with speech and music input signals. The listening tests were conducted in the same meeting room described in the previous section, but with the left and right loudspeakers positioned at (3.2, 6.1, 0.7) m and (4, 6.1, 0.7) m (keeping the same distance of 0.8 m between the loudspeakers), and the listeners were positioned at (3.6, 3.93, 0.7) m, which is 2.17 m on the axis of the loudspeaker middle point.

A. Methodology

The listening tests were designed for evaluating the perceptual-transparency of the developed method, and for comparing its performance with that of the baseline solution of signal playback from a single loudspeaker. For this purpose, three systems are considered:

- **Reference** playback of the original stereo audio signals,
- **Test** playback of the signals after applying the developed method, i.e. playback for the calibration signals, and
- **Anchor** playback from the left loudspeaker only.

An ABX comparison-based test was conducted [39] that employs three audio signals with a 5-sec duration, played back with pauses of 1.5-sec between the signals, as presented in Table V. In the table, signal X is either Reference or Test with equal probability for the sequences in lines 1–2, or Reference or Anchor for the sequences in lines 3–4. Lines 1–2 serve to investigate if the Test system is perceptually-transparent compared to the Reference system, and, similarly, lines 3–4 serve to investigate if the Anchor system is perceptually-transparent compared to the Reference system. The signal types employed were the speech and music input signals defined in Section V. Different segments of the same audio material were used for producing signals A, B, and X, and in each test an ABX sequence was played only once. After playback, listeners were required to answer two questions:

1) Is system X the same as system A, or is it the same as system B?
2) Did you hear a difference between system A and system B, or did you guess?

24 normal hearing subjects (19 male, 4 female) participated in a listening experiment. Prior to the experiment, the subjects were trained using six examples of Anchor and Test systems. Testing began only after the subjects' training results were examined, showing that the subjects understood the instructions. For each listener, the experiment included 24 trials, corresponding to four repetitions of each of the six conditions (Test/Anchor $\times$ Jazz/Pop/Speech). The order of the tests and the order of signals A and B in each test were selected randomly. The results were averaged across subjects.

TABLE V
A, B, AND X SIGNALS FOR INVESTIGATING BOTH TEST AND ANCHOR SIGNALS COMPARED TO REFERENCE

	A	B	X
1	Reference	Test	Reference/Test
2	Test	Reference	Reference/Test
3	Reference	Anchor	Reference/Anchor
4	Anchor	Reference	Reference/Anchor

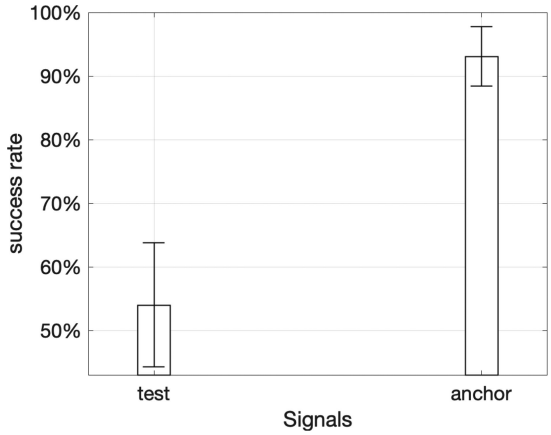

Fig. 11. Success rates for jazz music. Guess report: 45% (Test) and 11% (Anchor).

B. Results

Figs. 11, 12, and 13 show the average success rate of Test and Anchor systems for jazz music, pop music, and speech, respectively. Fig. 14 shows the overall success rate, which is the average for all signal types. In each figure, 95% confidence intervals for discrete variables are plotted (for more details, see Wilson Score Intervals [40]), and a guess report is also included in the caption of the figures corresponding to the second question the listeners were asked. A success rate of 50% is interpreted

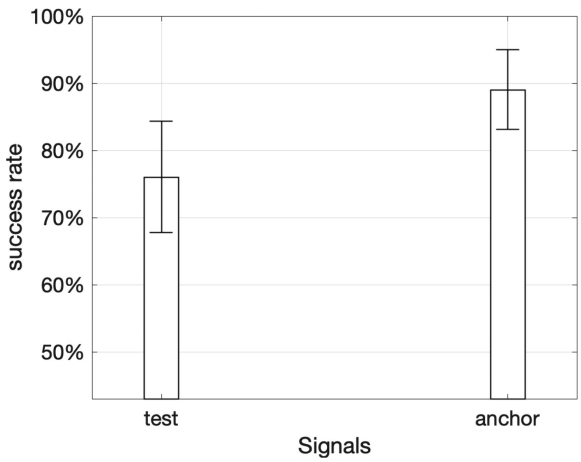

Fig. 12. Success rates for pop music. Guess report: 28% (Test) and 6% (Anchor).

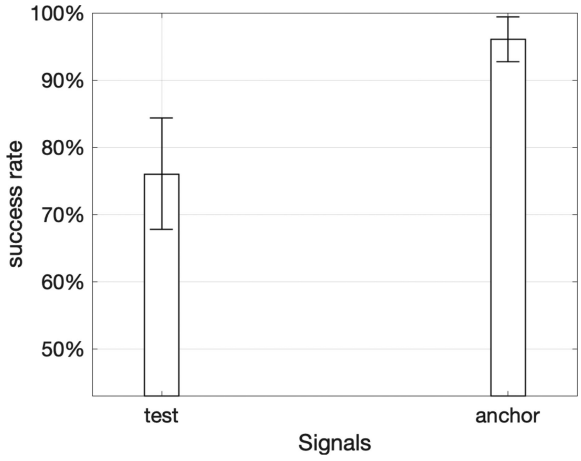

Fig. 13. Success rates for speech. Guess report: 16% (Test) and 6% (Anchor).

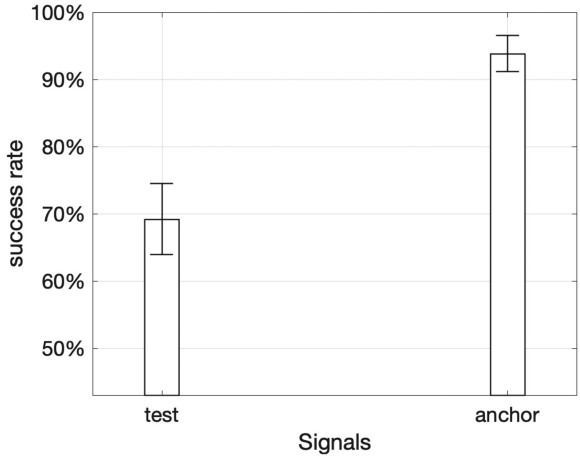

Fig. 14. Average success rates (for jazz music, pop music, and speech). Guess report: 30% (Test) and 8% (Anchor).

as perceptually-transparent since it corresponds to the average success rate of an equiprobable guess. A success rate of 100% is interpreted as completely non-transparent, and a 75% rate is interpreted as partially-transparent. The results show that the method developed is perceptually-transparent for the jazz music audio recording, and is partially perceptually-transparent for the pop music and speech. On the other hand, the anchor system, or baseline solution is very close to being completely non-perceptually-transparent for all input signals.

VII. Conclusion

An estimation method for a two-channel RTF using a single microphone was presented. The method was validated on measured RTFs in a simulation study, showing low estimation errors. The method was also investigated in a listening test, and was shown to be perceptually-transparent for an example of a music audio recording, and partially-perceptually transparent for other examples. The proposed approach therefore enables the estimation of a two-channel RTF in a perceptually-transparent manner, using the original audio signals.

One direction for future work could be the application of the method sequentially over limited frequency regions. While this may extend the estimation time, it may improve perceptual transparency. Another direction for future work is the extension of the method to more loudspeaker which should be relatively straightforward.

The proposed method has the following limitations. First, due to the use of practical filters for generating the calibration signals with non-zero stop band magnitude, there is signal leakage, or crosstalk between the filters used for the two channels. This leakage introduces an interference in the RTF estimation function, limiting the estimation accuracy. A second limitation is related to the application of the method, e.g. to RTF equalization. Due to the use of a single microphone, minor changes in the measurement geometry may lead to high variation of the measured RTF in particular at high frequencies. The measured transfer function at the microphone position may therefore be different than the one at a listener's position, especially at high frequencies. Finally, the perception of the calibration signals demonstrates additional limitations. A further study is required to understand these perceptual limitations, and their dependence on signal type, filter design, etc. Future work could also include the implementation and evaluation of the method for sound calibration in a room.

References

[1] S. Cecchi, A. Carini, and S. Spors, "Room response equalization—A review," *Appl. Sci.*, vol. 8, no. 1, pp. 10–15, 2018.
[2] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *J. Audio Eng. Soc.*, vol. 37, no. 6, pp. 419–444, 1989.
[3] C. Dunn and M. J. Hawksford, "Distortion immunity of MLS-derived impulse response measurements," *J. Audio Eng. Soc.*, vol. 41, no. 5, pp. 314–335, 1993.
[4] R. C. Heyser, "Acoustical measurements by time delay spectrometry," *J. Audio Eng. Soc.*, vol. 15, no. 4, pp. 370–382, 1967.
[5] A. Berkhout, M. M. Boone, and C. Kesselman, "Acoustic impulse response measurement: A new technique," *J. Audio Eng. Soc.*, vol. 32, no. 10, pp. 740–746, 1984.
[6] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. Conv. 108*, Audio Engineering Society, 2000, pp. 1–24.
[7] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 443–471, 2001.

[8] T. Gustafsson, J. Vance, H. Pota, B. Rao, and M. Trivedi, "Estimation of acoustical room transfer functions," in *Proc. IEEE Decis. Control, 39th Conf.*, 2000, vol. 5, pp. 5184–5189.
[9] N. Aoshima, "Computer-generated pulse signal applied for sound measurement," *J. Acoust. Soc. America*, vol. 69, no. 5, pp. 1484–1488, 1981.
[10] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1995.
[11] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 249–262, 2002.
[12] B. S. Lavoie and W. R. Michalson, "Auto-calibrating surround system," U.S. Patent 7,158,643, Jan. 2, 2007.
[13] C. Hak, R. Wenmaekers, J. Hak, L. Van Luxemburg, and A. Gade, "Sound strength calibration methods," in *Proc. Int. Congr. Acoust.*, 2010, pp. 1–6.
[14] N. Zacharov and P. Suokuisma, "Method for loudness calibration of a multichannel sound systems and a multichannel sound system," U.S. Patent 6 639 989, Oct. 28, 2003.
[15] J. Johnston and S. Smirnov, "Room acoustics correction device," U.S. Patent App. 11/289 328, May 31, 2007.
[16] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and application to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.
[17] Y. Lin and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 839–847, Mar. 2006.
[18] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 219–230, 2001.
[19] P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers, and F. F. Li, "Room acoustic parameter extraction from music signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Proc.*, 2006, vol. 5, pp. V - 801–V - 804.
[20] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and Gaussian processes–revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.
[21] C. Andersson, N. Wahlström, and T. B. Schön, "Data-driven impulse response regularization via deep learning," in *Proc. 18th IFAC Symp. Syst. Identification*, Stockholm, Sweden, 2018, pp. 1–6.
[22] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.
[23] G. Pillonetto, A. Chiuso, and G. De Nicolao, "Prediction error identification of linear systems: A nonparametric gaussian regression approach," *Automatica*, vol. 47, no. 2, pp. 291–305, 2011.
[24] H. Kim *et al.*, "Acoustic room modelling using a spherical camera for reverberant spatial audio objects," in *Proc. Audio Eng. Soc. Conv. 142*, Audio Engineering Society, 2017, pp. 1–10.
[25] H. Kim, L. Remaggi, P. Jackson, and A. Hilton, "Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360 images," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2019, pp. 120–126.
[26] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, "Binaural dichotic presentation to reduce the effects of spectral masking in moderate bilateral sensorineural hearing loss," *Int. J. Audiol.*, vol. 51, no. 4, pp. 334–344, 2012.
[27] A. Amano-Kusumoto, J. M. Aronoff, M. Itoh, and S. D. Soli, "The effect of dichotic processing on the perception of binaural cues," in *Proc. Thirteenth Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1476–1479.
[28] P. C. Loizou, A. Mani, and M. F. Dorman, "Dichotic speech recognition in noise using reduced spectral cues," *J. Acoust. Soc. Am.*, vol. 114, no. 1, pp. 475–483, 2003.
[29] H. Vold, J. Crowley, and G. T. Rocklin, "New ways of estimating frequency response functions," *Sound Vib.*, vol. 18, no. 11, pp. 34–38, 1984.
[30] H. Herlufsen, "Technical review - Dual channel FFT analysis (Part II)," *Brüel Kjær Tech. Rev.*, no. 1984-2, 1984.
[31] *Acoustics—Measurement of Room Acoustic Parameters–Part 2: Reverberation Time in Ordinary Rooms*, ISO 3382-2, *International Organization for Standardization, Genève*, 2008.
[32] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.
[33] Avishai Cohen Trio & Ensemble, "Remembering," At Home, Razdaz Recordz, 2005.
[34] Amy Winehouse, "Stronger than me," Frank, Island Records Ltd., 2003.
[35] Trevor Noah is Lupita Nyong'o's Son (2018), added by The Tonight Show Starring Jimmy Fallon. Accessed: Oct. 31, 2018. [Online]. Available: https://www.youtube.com/watch?v=ctTKPs66rCU
[36] M. Smith and T. Barnwell, "A new filter bank theory for time-frequency representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 3, pp. 314–327, Mar. 1987.
[37] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[38] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, Jun. 1967.
[39] W. Munson and M. B. Gardner, "Standardizing auditory tests," *J. Acoust. Soc. Am.*, vol. 22, no. 5, pp. 675–675, 1950.
[40] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statist. Sci.*, vol. 16, pp. 101–117, 2001.

Hai Morgenstern was born in Rishon LeZion, Israel, in 1985. He received the B.Sc. degree (summa cum laude) and the M.Sc. degree (cum laude) in electrical and computer engineering from Ben-Gurion University (BGU) of the Negev, Israel, in 2012 and 2013, respectively. He received the Ph.D. degree in electrical and computer engineering from BGU in 2017. He spent six months as an Exchange Research Student at the Institute of Electronic Music (IEM), Graz, Austria, in 2011. In 2014, he was awarded the Chateaubriand Fellowship and spent a year at the Institute de Recherche et Coordination Acoustique/Musique (IRCAM) in Paris, France, investigating acoustic multiple-input multiple-output systems for room acoustics analysis. He has been awarded the short-term Postdoctoral Negev grant in 2017, and spent two years as a Researcher at the Acoustics Laboratory at BGU investigating source localization and separation and processing methods for 3D analysis of room acoustics. In 2019, he joined BeyondMinds, Tel Aviv-Yafo, Israel, as a Research Scientist, and is currently investigating deep neural networks with applications in speech and natural language processing.

Boaz Rafaely (SM'01) received the B.Sc. degree (cum laude) in electrical engineering from Ben-Gurion University, Beer-Sheva, Israel, in 1986; the M.Sc. degree in biomedical engineering from Tel-Aviv University, Israel, in 1994; and the Ph.D. degree from the Institute of Sound and Vibration Research (ISVR), Southampton University, U.K., in 1997. At the ISVR, he was appointed Lecturer in 1997 and Senior Lecturer in 2001, working on active control of sound and acoustic signal processing. In 2002, he spent six months as a Visiting Scientist at the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, investigating speech enhancement for hearing aids. He then joined the Department of Electrical and Computer Engineering at Ben-Gurion University as a Senior Lecturer in 2003, and was appointed as Associate Professor in 2010, and Professor in 2013. He is currently heading the acoustics laboratory, investigating methods for audio signal processing and spatial audio. During 2010–2014, Dr. Rafaely has served as an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and during 2013–2018 as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He also served as an Associate Editor for IEEE SIGNAL PROCESSING LETTERS during 2015–2019, for *IET Signal Processing* during 2016–2019, and currently for *Acta Acustica united with Acustica*. During 2013–2016 he has served as the chair of the Israeli Acoustical Association, and is currently chairing the Technical Committee on Audio Signal Processing in the European Acoustical Association. He was awarded the British Councils Clore Foundation Scholarship.