2018-04-01

Development of a Real-Time Auralization System for Assessment of Vocal Effort in Virtual-Acoustic Environments

Jennifer Kay Whiting
Brigham Young University

Development of a Real-Time Auralization System for Assessment of Vocal Effort in Virtual-Acoustic Environments

Jennifer Kay Whiting

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Timothy W. Leishman, Chair
Christopher D. Dromey
David G. Long
Tracianne B. Neilsen

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Development of a Real-Time Auralization System
for Assessment of Vocal Effort in
Virtual-Acoustic Environments

Jennifer Kay Whiting
Department of Physics and Astronomy, BYU
Master of Science

This thesis describes the development of the real-time convolution system (RTCS) for a little-studied talker/listener in virtual acoustic environments. We include descriptions of the high-resolution directivity measurements of human speech, the RTCS system components, the measurement and characterization of oral-binaural room impulse responses (OBRIRs) for a variety of acoustic environments, and the compensation filter necessary for its validity. In addition to incorporating the high-resolution directivity measurements, this RTCS improved on that developed by Cabrera et al. [1] through the derivation and inclusion of the compensation filter. Objective measures in the time- and frequency-domains, as well as subjective measures, were developed to asses the validity of the RTCS. The utility of the RTCS is demonstrated in the study on vocal effort, and the results of an initial investigation into the vocal effort data are presented.

Keywords: auralization, real-time convolution, oral-binaural room impulse responses

ACKNOWLEDGEMENTS

I wish to acknowledge the multiple sources of help and support without which I would not have been able to complete this document.

First, a special thank you to Dr. Eric Hunter for collaborating with Dr. Timothy Leishman and providing the research grant from which this research was funded. Funding for this research came from the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Number R01DC012315. They would like to remind the reader that the content of this publication is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

In addition, I wish to thank the faculty of the Department of Physics and Astronomy at BYU who taught principles of acoustics and the gospel in their classes, namely, Dr. Gee, Dr. Neilsen, Dr. Sommerfeldt, and Dr. Leishman.

Thank you to the many students and research assistants who performed many hours of research in the anechoic chamber and elsewhere to perfect this project: Kieren Smith, Travis Hoyt, Brooks Butler, Claire Pincock, Michael Rollins, Joy Clay, Rachel Averett, Adam Chrsitensen, Sam Graham, Samuel Bellows, and Mark Berardi.

I wish to thank my family, especially my husband Eric and daughter Maddy who have sacrificed much to allow me to complete my degree. Thank you to my parents and brothers for always supporting me in my goals.

Thank you to my friends who spent many hours of child care during the late stages of writing this document: Vanessa Wiest, Brooke Hanks, Andrea Murray, Stephanie Ricks, Joy Kent, and Liza Kofford. I would not have been able to do this without the help of everyone mentioned here.

Table of Contents

List of Figures

List of Tables

Glossary of Symbols

This glossary contains variables that are used repeatedly in the thesis.

$\hat{a}(f)$ Complex, frequency-dependent signal at a hypothetical point in space near the mouth of a talker

$\hat{a}'(f)$ Signal at a hypothetical point in space near the mouth of a talker in the presence of the RTCS microphone and headphones

$a'_K(f)$ Radiated acoustic pressure signal at a point near the KEMAR mouth simulator in the presence of the RTCS microphone and earphones

$A_K(f)$ Composite transfer function from EASERA used on laptop computer, through PreSonus FireFace, Crown D-45 power amplifier, and KEMAR mouth simulator.

$\hat{a}_K(f)$ Radiated acoustic pressure signal at a point near the KEMAR mouth simulator

$\hat{a}_s(f)$ Digital waveform used to drive KEMAR mouth simulator

$\hat{b}_{K,L}^{\mathrm{ANCH}}(f)$ and $\hat{b}_{K,R}^{\mathrm{ANCH}}(f)$ Signal at the entrances to the KEMAR ear microphones while in the presence of the RTCS microphone and headphones, in an anechoic environment.

$\hat{b}_{s,K,L}^{\mathrm{ANCH}}(f)$ and $\hat{b}_{s,K,R}^{\mathrm{ANCH}}(f)$ Signal recorded by KEMAR ear microphones while in the presence of the RTCS microphone and headphones, in an anechoic environment.

$\hat{b}_{L}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{R}^{\mathrm{RTCS}}(f)$ Signals at the entrances to the blocked left and right ear canals of a talker using the RTCS

$\hat{b}_{K,L}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{K,R}^{\mathrm{RTCS}}(f)$ Signal at the entrances to the KEMAR ear microphones while in the presence of the RTCS microphone and headphones.

$\hat{b}_{s,K,L}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{s,K,R}^{\mathrm{RTCS}}(f)$ Signal recorded by KEMAR ear microphones while using the RTCS with a room OBRIR

$\hat{b}_{K,L}^{\mathrm{RTCS}\,\delta}(f)$ and $\hat{b}_{K,R}^{\mathrm{RTCS}\,\delta}(f)$ Signal at the entrances to the KEMAR ear microphones while in the presence of the RTCS with a delta function.

$\hat{b}_{s,K,L}^{\mathrm{RTCS}\,\delta}(f)$ and $\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)$ Signal recorded by KEMAR ear microphones while using the RTCS with a delta function

$\hat{b}_{L}^{\mathrm{room}}(f)$ and $\hat{b}_{R}^{\mathrm{room}}(f)$ Complex, frequency-dependent signals at the entrances to the blocked left and right ear canals of a talker in a room

$\hat{b}^{\text{room}}_{K,L}(f)$ and $\hat{b}^{\text{room}}_{K,R}$	Signals at the entrances to the KEMAR ear microphones while KEMAR is in a room
$B_{K,L}(f)$ and $B_{K,R}(f)$	Composite transfer functions from the KEMAR ear canal openings through the left and right KEMAR ear microphones, the corresponding phantom-to-ICP power converters, FireFace preamplifier and A/D converter, and laptop computer running EASERA
$\hat{b}^{\text{room}}_{s,K,L}(f)$ and $\hat{b}^{\text{room}}_{s,K,R}(f)$	Signal recorded by KEMAR ear microphones while KEMAR is in a room
$C_1(f)$	Transfer function for the incoming hardware components of the RTCS: RME QuadMic II preamplifier, RME ADI-8 Q A/D converter, RME PCI Express Sound Card
$C_{2,L}(f)$ and $C_{2,R}(f)$	Transfer function for the outgoing hardware components of the RTCS: RME PCI Express Sound Card, RME ADI-8 Q D/A converter, Crown D-75 amplifier
$D_L(f)$ and $D_R(f)$	Transfer functions representing the propagation of the signal $\hat{a}(f)$ around the head of a talker and to the left and right ears, respectively
$D'_L(f)$ and $D'_R(f)$	Transfer functions representing the propagation of the signal $\hat{a}'(f)$ around the head of a talker to the left and right ears, modified by the presence of the RTCS microphone and headphones
$D_{K,L}(f)$ and $D_{K,R}(f)$	Transfer functions representing the propagation of the signal $\hat{a}_K(f)$ around the head of a KEMAR mannequin and to the left and right ears, respectively
$D'_{K,L}(f)$ and $D'_{K,R}(f)$	Transfer functions representing the propagation of the signal $\hat{a}'_K(f)$ around the head of a KEMAR mannequin and to the left and right ears, modified by the presence of the RTCS microphone and headphones
$F_L(f)$ and $F_R(f)$	Filter designed to flatten or equalize the response of the RTCS
$\text{HRTF}_L(f)$ and $\text{HRTF}_R(f)$	Head-related transfer function corresponding to the Fourier transform of the head-related impulse response of a talker.
$\text{HRTF}_{K,L}(f)$ and $\text{HRTF}_{K,R}(f)$	Head-related transfer function of a KEMAR mannequin
$H^{\text{RTCS}}_L(f)$ and $H^{\text{RTCS}}_R(f)$	Frequency response function (FRF) for a talker using the RTCS, relating the signal at the entrance of the blocked ear canals to the signal at a hypothetical point in space near the mouth of a talker
$H^{\text{RTCS}}_{K,L}(f)$ and $H^{\text{RTCS}}_{K,R}(f)$	Frequency response function (FRF) for KEMAR using the RTCS with a room OBRIR, relating the signal recorded by the ear microphones to the signal driving the mouth simulator

$H_{K,L}^{\mathrm{RTCS}\,\delta}(f)$ and $H_{K,R}^{\mathrm{RTCS}\,\delta}(f)$	Frequency response function (FRF) for KEMAR using the RTCS with a delta function, relating the signal recorded by the ear microphones to the signal driving the mouth simulator
$H_{L}^{\mathrm{room}}(f)$ and $H_{R}^{\mathrm{room}}(f)$	Frequency response functions for a talker in a room, relating the signals at the entrances of the blocked ear canals to the signal at a hypothetical point in space near the mouth of a talker
$H_{K,L}^{\mathrm{room}}(f)$ and $H_{K,R}^{\mathrm{room}}(f)$	Frequency response function (FRF) for KEMAR in a room, relating the signal recorded by the ear microphones to the signal driving the mouth simulator
$M(f)$	Transfer function of the propagation path from the hypothetical point in space near the mouth of a talker and the response of the head-worn microphone
$\widehat{m}_s(f)$	Complex, frequency dependent signal recorded by the head-worn microphone
$R(f)$	Fourier transform of an arbitrary room's impulse response from a hypothetical point in space near the mouth of a talker to the unobstructed central head position
$T_L(f)$ and $T_R(f)$	Transfer functions of the left and right AKG K1000 headphone transducers, including propagation paths from the headphones to the entrances of the blocked ear canals of a talker.
$T_{K,L}(f)$ and $T_{K,R}(f)$	Transfer functions of the left and right AKG K1000 headphone transducers, including propagation paths from the headphones to the entrances of the KEMAR ear simulators.

Glossary of Abbreviations

AFMG	Ahnert Feistel Media Group
ANCH	Anechoic
ASM	Assumptions
BRIR	Binaural Impulse Response
EASE	Enhanced Acoustic Simulator for Engineers
EASERA	Electronic and Acoustic System Evaluation and Response Analysis
ERB	Equivalent Rectangular Bandwidth
FRF	Frequency Response Function
HATS	Head and Torso Simulator
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
IR	Impulse Response
KEMAR	Knowles Electronic Mannequin for Acoustic Research
OBRIR	Oral-Binaural Room Impulse Response
RIR	Room Impulse Response
rms	Root mean square
RTCS	Real-Time Convolution System
SIR2	Real-time convolution VST plugin
TF	Transfer Function
VST	Virtual Studio Technology

Chapter 1

Introduction

When listening to your own voice recorded, as on a telephone answering machine, you may notice that you sound different than when you hear yourself talking. The paths the sound takes from your vocal production mechanisms (vocal folds, mouth, and nose) to your auditory receivers (ears) influence the sound of your own voice. When you hear yourself speaking, you have additional sound propagation paths, such as bone conduction and near-field head diffraction that are not present when you listen to the recording [2].

In addition to the paths mentioned above, the acoustic environment in which you are speaking also influences the sound of your own voice. Reflected sound from the room also arrives at your ears, adding another layer. The acoustic simulation of one's own voice requires consideration of all these sound paths.

This thesis describes the development and use of a real-time convolution system (RTCS) to produce real-time auralizations of the human voice in virtual acoustic environments. The utility of such a system is demonstrated in a study on vocal effort. In this study, the response of talkers to simulated acoustic environments was evaluated. As part of the development of the RTCS, high-resolution directivity measurements of human speech were carried out and implemented in architectural acoustic simulation software to create virtual acoustic environments.

1.1 Background

Most sound sources, including the human voice, do not radiate equally in all directions. The relative amount of sound radiated in a given direction at a specific frequency is called directivity. The directivity of a sound source may be measured by comparing acoustic pressure data at various surrounding locations to that measured by a reference microphone. Acoustical researchers and practitioners have long relied on estimated, lower-resolution, single-capture, and polar (single plane) directivity data for speech because complete high-resolution spherical data were unavailable [3-6]. The lower-resolution directivity was necessitated largely by experimental difficulties and costs [7]. Another complicating factor is that talkers, unlike transducers, fail to produce repeatable signals needed for common directivity measurements. The present work demonstrates that complete, reliable, high-resolution speech directivities can be measured with feasible measurement tools and proper attention to experimental and signal-processing details.

These high-resolution speech directivities are useful in improving models of the human voice in architectural acoustic simulations. The resulting auralizations simulate what the human voice may [8-10] sound like in an acoustic environment; such auralizations are commonly produced before construction of a new building. Simply put, an auralization is the creation of an audible sound file from a numerical simulation of sound propagating from a source to a receiver in an acoustical environment [11].

An auralization consists of several steps. The first step is to model the sound production via the frequency-dependent amplitude and directivity of an acoustic source. The next step is a convolution between the direct-sound signal and a room impulse response (RIR) function for the acoustic environment. This convolution is an integral operation that simulates how an acoustic signal would sound in that particular acoustic space. Lastly, the response of the receiver is considered. If the receiver is binaural, as in a human listener with two ears, the response of the receiver is a head-related transfer function (HRTF), and a stereo signal is the auralization result, after convolution with the HRTF.

Each step of the auralization process requires attention to detail and accuracy. The directivity of the human voice as a sound source has already been discussed. For greatest accuracy, the directivity needs to be high-resolution and high quality. The room impulse response (RIR) must also be treated appropriately. Most rooms are modeled geometrically with absorption and scattering coefficients applied to each surface. The source and receiver must be placed and oriented within the virtual room. Once the model is set up, ray tracing, image sources, and diffuse-reflection algorithms are commonly used for generating an RIR. For reduced computational time in this step, many software packages [12-14] use a hybrid method utilizing ray tracing, image sources, and diffuse-reflection algorithms to calculate the RIR. The RIR is calculated by tracing sound from the source (including its directivity pattern) to the room surfaces for multiple reflections until the sound arrives at the receiver location. A final filter, the HRTF, is convolved with both the direct sound and the room reflections to create a binaural sound file. An HRTF is "the ratio of the Fourier transform of the sound pressure level at the ear canal to that which would have been obtained at the head center without the listener present" [11]. It describes the effect of the head on the approaching sound rays just prior to their reception at the ear, as opposed to having no head in place, just a single receiver. The HRTF is most commonly associated with humans' ability to locate sound, since the two ears can separate and identify the direction from which sound is approaching. Individual people have unique HRTFs based on their unique head geometry and pinnae. The HRTF can be measured by using binaural microphones just outside a person's blocked ear canal [15]. Head-tracking systems are sometimes used to create more realistic experiences as listeners reorient themselves within a virtual environment [16].

To summarize, an auralization is produced by convolving a dry, anechoic signal with the RIR containing the source directivity and the receiver HRTF to simulate how it would sound in the space the impulse response was measured in. Kleiner describes auralization as being analogous to visualization [17]. As a computer rendering of a visual scene allows one's eyes to see a virtual space, an auralization allows one's ears to hear a virtual space. Auralizations may be produced well ahead of time and treated like recordings in an acoustic space.

Real-time auralizations occur when the input signal is sampled and passed through a digital computer to be convolved with an IR and the output is played back with minimal latency. The computational load of (and latency introduced by) convolving even small buffer sizes of signal with a long impulse response limits the realism of the real-time output. Thus, convolution is often performed via frequency domain multiplication, which speeds up the computation and reduces the system latency.

Computational advances in recent years have made real-time auralizations possible with ever decreasing latency and increasing realism. One of the principal authors in the field of virtual acoustics is Michael Vorlander. Vorlander treats auralizations as "the technique of creating audible sound files from numerical data" [2, 3]. However, he admits that the latency introduced by computational system components such as audio hardware, filters, and head trackers leave little time for the computation of the acoustic simulation. To be considered real-time, the latency of the total system must "be sufficiently small so that the listener is not disturbed" [18]. Ideally, the total system latency must be small enough that differences from the authentic listening experience are less than the just-noticeable differences (JNDs) for various psychoacoustic parameters, such as tone roughness, loudness, and sharpness [19]. Associated with these JNDs are temporal factors, including the minimum update time for the acoustic scene rendering to occur. Vorlander notes, “To achieve real-time performance, specific run-time conditions must be taken into account to stay within acceptable limits for latency and update rates. Update rates of 60 Hz and total delays of 50 ms are considered acceptable for acoustic virtual sound” [11].

Cabrera et al. focused efforts on a system to simulate a talker’s voice in virtual acoustic environments. They did this by measuring or simulating oral-binaural room impulse responses (OBRIRs), where the source and the receiver were located close together, much like a mouth and ears [20]. They then used real-time convolution to convolve live subjects’ speech or singing with the RIR and presented the result, minus the direct sound, via off-ear headphones [1]. This thesis discusses the development and improvement of a similar RTCS and its use as a tool to assess vocal effort.

1.2 Objectives

This thesis advances the realism of real-time auralizations of a talker’s own voice. The first step in improving auralization is high-resolution directivity. A primary goal for the research effort was to take high-resolution directivity measurements of live talkers and perform the necessary data transformations to prepare the measurements for use in architectural acoustic simulation software packages. Speech scientists and architectural acousticians alike can benefit from these results. Former work by prior students laid the groundwork for the recording and processing of the speech directivity data [7,21]. An anechoic chamber rated down to 80 Hz [22], a large semi-circular array of microphones, and a central turntable, were used to perform repeated-capture directivity measurements.

The second step is to create realistic auralizations of self-generated speech sounds, using the directivity measurement results, measured and simulated OBRIRs, and a RTCS. To be used effectively for this purpose, the shortcomings in the RTCS developed by Yadav and Cabrera were corrected and improved. The anechoic chamber was also used to house the RTCS, so no dereverberation convolution algorithms were needed to reduce reflections from the playback environment at the users' ears. Rather, simulated acoustic reflections were presented with the RTCS headphones to create virtual-acoustic environments. The chamber was also large enough that two or more people could fit comfortably, making an interview scenario possible for the vocal effort study. To investigate the usefulness of the resulting real-time auralization system a vocal effort study was performed in which gender differences in response to virtual acoustic environments are investigated [23-36]. The study was patterned after a similar work in BYU's large reverberation chamber [37].

This thesis reports on each of the research milestones. It discusses the methods used to take directivity measurements, their substantial results, and their use in architectural acoustic simulation software. It provides a literature overview for live speech and singing directivity measurements, and a brief introduction to real-time convolution and auralization. The development of the RTCS is laid out, including major obstacles in signal processing and the approaches to overcome them. The plan for the vocal effort experiment is included, as well as statistically significant results from the study.

1.3 Plan of Development

This chapter has explained the general background and motivations for the research, and the general objectives and scope of the thesis. Chapter 2 provides the details of the human speech directivity study. Chapter 3 explains the measurement of oral-binaural room impulse responses (OBRIRs) to be employed in the RTCS. Chapter 4 describes the background, and development of the RTCS, while Chapter 5 explains the steps taken to validate the system objectively and subjectively. Chapter 6 introduces the use of the RTCS as a tool to study vocal effort. All of these chapters are structured in a format similar to the thesis as a whole. They are thus self-contained reports of specific focuses, with their own background, motivations, objectives, details of measured or simulated data, analysis of the results, and conclusions. Finally, Chapter 7 restates the significant conclusions from the work in Chapters 2 through 6 and explains the impact of the work as a whole. It also provides recommendations for future work.

Chapter 2

Speech Directivity

2.1 Introduction

Most sources of sound, including the human voice, do not radiate equally in all directions. The directivity of a sound source describes the directional variation in the amplitude of radiated sound as a function of frequency. Directivity is measured by comparing acoustic signals at various measurement locations to a reference signal at the source. Acousticians and other professionals have long relied on estimated, low-resolution, single-capture, and polar (single plane) directivity data for speech because high-resolution spherical data have been unavailable. This deficiency has been caused largely by experimental difficulties and costs. Talkers, unlike transducers, fail to produce repeatable signals needed for common directivity measurements, making repeated-capture (or sequential) measurements challenging [7]. This chapter demonstrates that complete, reliable, high-resolution speech directivities can be measured using feasible procedures and proper attention to experimental and signal-processing details.

Directivities are typically measured at far-field distances as a function of frequency. In the far field, the normalized directivity pattern—normalized by the peak value at a given distance—remain consistent as distance increases. Comparisons between normalized directivities of different sources are thus straightforward because exact distances are inconsequential.

Because complete source characterizations also involve spectral variations, speech directivity should be reported over frequency. The data acquired in this study involved 1 Hz narrowband resolution, but could be also summed over other bands as needed. The results reported herein are presented mainly over one-third octave bands.

In the past, several researchers have explored voice directivity using various methods. Some have investigated near-field radiation with applications to telecommunications devices, but efforts related to far-field measurements are more pertinent to the present work. Dunn and Farnsworth were among the first to study live speech directivity [38]. In their experiments, a single human talker repeated 15 seconds of speech while being measured sequentially at 76 surrounding positions in an absorptive but non-anechoic environment. Over 5,000 repetitions were performed as part of the effort and the results were evaluated over octave bands. The authors determined three factors that affect the radiation patterns most significantly: (1) the shadow effect produced by diffraction of the head and body at positions behind the talker, (2) the size of the mouth opening for frequencies above 5,600 Hz, and (3) radiation from locations other than the mouth, such as the throat and chest.

The next studies of talker directivity gave similar results. In 1985, Studebaker [39,40] measured directivity using a 90 s speech passage and signals acquired at 1 m and 45° increments in the horizontal (transverse) plane. The results compared favorably to those of Dunn and Farnsworth. He also compared his measured directivities to those of various loudspeakers. Chu and Warnock [41] evaluated live speech in both the horizontal and vertical (median) planes. Their measurements were taken at 1 m and 15° increments around talkers who spoke for 40 s. The authors investigated differences between male and female directivities, and those between French and English passages. The results of the live speech directivities they measured were similar to those produced by a Brüel and Kjær head-and-torso simulator (HATS), and to the Dunn and Farnsworth measurements [38,41].

Others have similarly contributed to the understanding of both speech and singing directivity. In 1985, Marshall and Meyer [42] measured directivities of professional singers over two octaves, utilizing three vowels and two singing styles. Measurements were made in the horizontal and vertical planes at 20° increments down to 40° degrees below the singer's mouth. Katz et al. presented measurements taken in the horizontal plane in 15° increments for specific sustainable phonemes, while also exploring differences in sung intensities of those phonemes [43,44]. In 2012, Monson et al. examined speech and singing directivities in the horizontal plane at 15° increments, using both long-term averages and distinct phonemes [45]. Kob measured singers performing glissandi over one octave and compared directivity results to those of an artificial mouth radiating white noise. His measurements were made on a partial sphere, extending from -40° to +90° elevation with a single moveable field microphone [46].

Some authors have specifically investigated directivities of HATS and artificial voice simulators for comparison to live speech directivities [47-53]. Halkosaari [50] investigated the directivity of a Bruel and Kjaer HATS for cell phone microphone testing, and compared the HATS directivity measurements to the same measurement locations for 15 test subjects. He found that the HATS was too directional compared to the live speech directivities at higher frequencies for the measurement locations he chose in the near field. Bozzoli et al. [47] examined artificial and live speech directivities for the intent of better assessing speech transmission index in close situations, such as in a car. He used five microphones on a moveable stand at 1 meter and a repeated-capture method in 15° increments for 10 male subjects. In contrast, he tested a Bruel and Kjaer HATS by positioning it on a turntable and keeping the microphone stationary. He concluded that the lack of a norm about artificial mouth's balloon of directivity means that different sources have different behavior, which in turn yields differing results for STI computation in a car depending on the HATS model used.

None of the aforementioned efforts have assessed voice directivities over a complete sphere or with the uninterpolated 5° resolution that has been standardized for loudspeakers [54] which would be so useful for analysis of radiated speech, architectural acoustics simulations, audio recordings, sound reinforcement, etc. (One should note that recorded or reinforced speech is necessarily affected by microphone placement, which is in turn affected by directivity and (near-field) distance.) Instead, researchers have employed lower-resolution directivity, often on individual planes and at inconsistent angular increments. Furthermore, their speech and singing materials were not standardized and spectral resolutions were often limited to one-third or full octave bands. Results were also presented in varied formats, including tables, plots over angle, and plots over frequency, etc.

The present study was part of a larger investigation into the effects of room acoustics on speech communication (see Chapter 6). Its aim was to assess vocal efforts of talkers in virtual acoustics environments using a RTCS that required speech directivities in its models. Greater availability and knowledge of high-resolution speech directivities can also inform the efforts of speech scientists, architectural acousticians, audio engineers, hearing-aid engineers, telecommunications engineers, automotive engineers, and other specialists. Their work can be improved if more detailed and accurate directivities are made available in clear and readily usable formats and implemented wisely. Thus, high-resolution directivity results should lead to enhanced insights regarding human speech, including aiding in the development of better scientific models for speech simulation.

This chapter presents a feasible approach to measure and process high-resolution live-speech directivities while accounting for inherent diffraction and absorption of seated talkers. Results are presented as three-dimensional directivity balloons, associated coherence balloons, and polar plots in the transverse, median, and frontal planes. Data are presented for a composite average of four female talkers, a composite average of four male talkers, and one KEMAR mannequin. The resulting directivities are compared across source type and to lower-resolution results of past researchers for further validation. Section 2.2 provides details of the measurement and data processing methods. Section 2.3 presents selected directivity results. Section 2.4 provides discussion of those results, and a comparison to the directivity of the KEMAR mannequin. Finally, Sec. 2.5 presents conclusions from the work and suggestions for future efforts.

2.2 Methods

An apparatus for the directivity measurement of several sound sources, including live speech, is shown in Fig. 2.1 [7]. The measurements took place in a chamber that is anechoic down to about 80 Hz, which is below the fundamental frequency of most human speech. The apparatus included a semicircular arc with 37 microphones spaced angularly with 5° increments. The radius from the circular center of the arc to the microphones was adjustable and set to 1.2 meters. Each speech subject sat on a chair attached to a turntable that rotated with 5° increments under computer control for each repetition of a speech passage. The subject was positioned such that his or her mouth was at the circular center of the array. The head was held in place with an

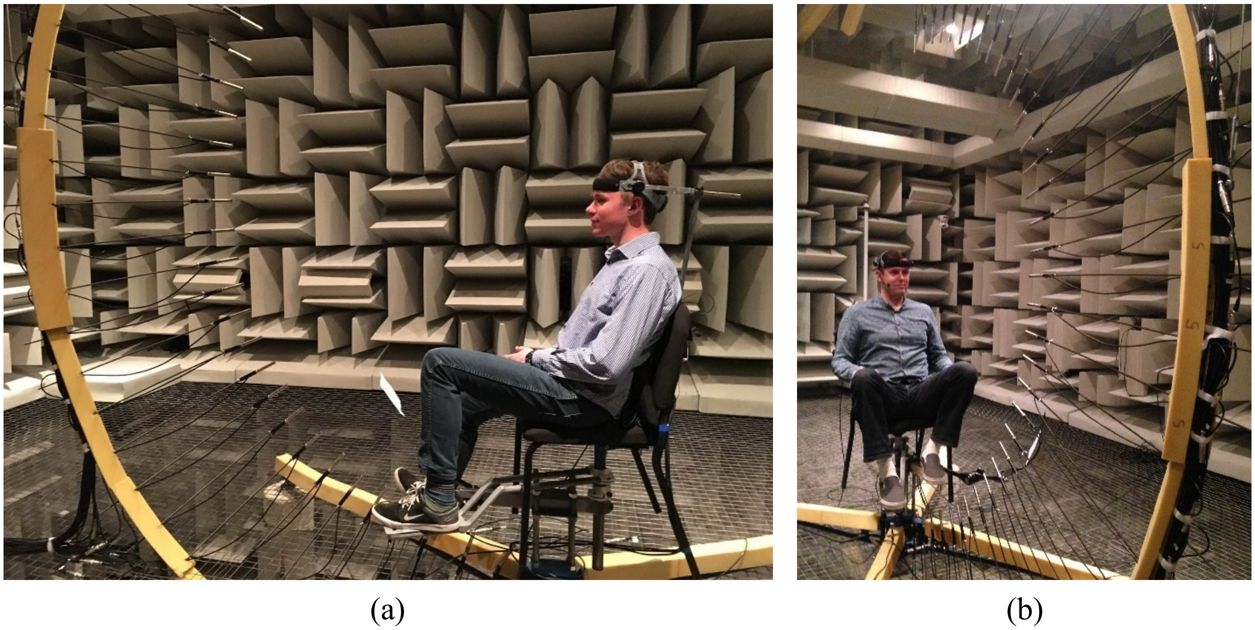

(a) (b)

Figure 2.1. Two views of a male subject in the directivity measurement apparatus. The complete array of microphones at a 1.2 m radius is seen in subplot (b).

adjustable restraint to ensure that the subject remained stationary within the rotating reference frame for the duration of the measurement sequence.

The subject repeated a brief phonetically balanced passage for each rotation [55]. The passage consisted of six sentences containing most of the commonly used phonemes in the English language. The passage had four statements and two questions, shown in Table 2.1. The passage took about ten to fifteen seconds to repeat, making it ideal for the full duration of the measurement sequence, so as not to fatigue the subjects. A full sequence took about 2 hours to complete and contained 2,522 unique measurement points on a sphere around the subject. This directivity measurement configuration met the high-resolution standard of AES56-2008 Type A, normally applied to loudspeakers [54].

Table 2.1. Phonetically balanced passage used for speech directivity measurements.

1. Measure three young kids for height.	4. How do we go there from here?
2. Which boat tour should they join now?	5. Black soot and parks annoy her.
3. Some vagabonds share an apartment.	6. You'll be my love for always.

The subject spoke along to a prompt track, heard through small in-ear headphones. Figure 2.2 shows an example of a female subject sitting in the chair and using the head restraint. The cheek worn microphone is also shown, although the earbuds used for playback of the prompt track are not pictured. The paper containing the six-sentence passage is in the lower right hand corner.

In addition to the 37 microphones on the array, three microphones were positioned near the subject within the rotating reference frame. One reference microphone was selected to produce a reference signal, $a(t)$, for the calculation of frequency-response functions (FRFs),

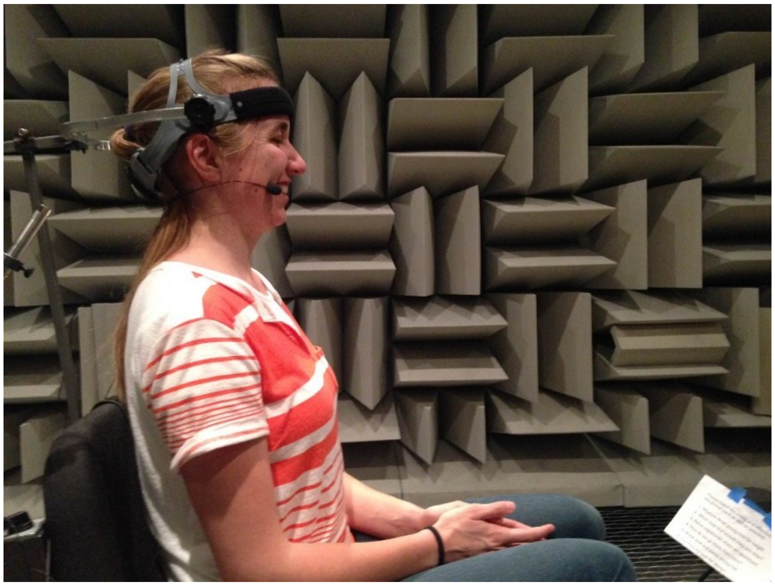


Figure 2.2. A female subject in the directivity measurement chair. The cheek-worn reference microphone, the head positioning apparatus, and the prompt paper are pictured. The earbuds used for playback are not pictured.

$H_{u,v}(f)$ between $a(t)$ and the signals $b_{u,v}(t)$ from the arc-array microphones, where $u = 0, 1, 2, \dots, U - 1$, $v = 0, 1, 2, \dots, V - 1$, where $U = 37$ and $V = 72$ are the number of measurements in the theta and phi directions respectively. The magnitude FRFs in relation to each other over the measurement sphere constitute the directivity balloon for the talker. The method for computing directivity was similar to that described by Leishman et al [56]. Figure 2.3 gives a diagram of the coordinate system used in these directivity measurements.

The FRFs were calculated using the autospectrum $G_{aa}(f)$ from the reference signal and cross-spectra $G_{ab_{u,v}}(f)$ from various array signals as

$$H_{u,v}(f) = H(\theta_u, \phi_v, f) = \frac{G_{ab_{u,v}}(f)}{G_{aa}(f)}, \tag{2.1}$$

where $\theta_u = u\Delta\theta$, $\phi_v = v\Delta\phi$, and $\Delta\theta = \Delta\phi = 5°$. The magnitude of the FRF was then used to compute the decibel directivity as

$$L_{u,v}(f) = 10 \log \left[\frac{\left|H_{u,v}(f)\right|^2}{\left|H_{u,v}(f)\right|^2_{\text{max}}} \right], \tag{2.2}$$

which normalized the FRF magnitudes by their maxima $\left|H_{u,v}(f)\right|_{\text{max}}$.

In addition, coherence, $\gamma^2_{u,v}$ was computed for each measurement point as a measure of the validity of the FRF and directivity at that point. The coherence function estimates the extent to which the output signal, $b_{u,v}(t)$ can be linearly predicted from the input signal $a(t)$. It describes whether the measurement at a specific point (u,v) was contaminated by noise to the degree that the FRF at that measurement point is no longer trustworthy. Coherence is always a value between 0 and 1, due to the Cauchy Schwarz inequality in the least-squares computation:

$$\gamma^2_{u,v}(f) = \frac{\left|G_{ab_{u,v}}(f)\right|^2}{G_{aa}(f)G_{bb_{u,v}}(f)}. \tag{2.3}$$

These signal recordings, $a(t)$ and $b_{u,v}(t)$ were waveforms with normalized units, -1 to 1. However, in the FRF computation, only relative amplitudes were relevant, so with the proper calibration, no discrepancies were introduced. To reduce file sizes and increase the signal-to-noise ratios, the recordings were cut to remove the silence between sentences and then concatenated. The concatenated recording was then split into 2 s blocks with 75% overlap. The Fourier transformation and computation of auto and cross spectra was performed for each block and then averaged across blocks for each measurement location to yield time-averaged spectra. This computation, performed over the entire measurement sphere, gave $L_{u,v}(f)$ and $\gamma_{u,v}^2(f)$ for all u and v. These are the values presented in the balloon plots below.

The FRFs and coherence values for each measurement positions are presented in three-dimensional balloon plots, from which polar plots may be extracted. The data analysis was performed in narrow bands (1 Hz resolution) then summed into third-octave bands for presentation. The directivity data for each group of male and female subjects are averaged

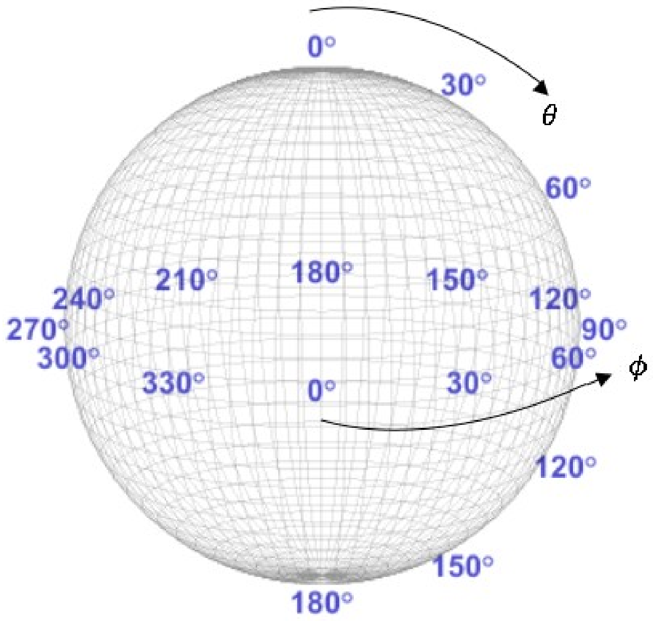


Figure 2.3. Coordinate system for directivity measurements.

energetically across the subjects. The directivity matrix saved for each subject after the first round of processing was linear FRF data. In order to average the directivities from different subjects, the magnitudes of the complex FRFs were squared. The squared functions for each group of subjects of the same gender were then averaged and converted to levels in decibels. The coherence data was simply averaged arithmetically across the subjects. The third-octave data was converted into a format compatible with EASE software for architectural acoustics modeling. More on directivity and EASE models is presented in Section 3.4.

2.3 Results and Analysis

Averaged results for each gender and the KEMAR mannequin are presented below in three-dimensional balloon plots. Directivity and the associated coherence functions are shown for a few frequencies that illustrate the variation across the main frequencies of speech. Animations of the full spectrum of directivity measurements are in Appendix A.

2.3.1 G.R.A.S. KEMAR mannequin type BC

The directivity of a G.R.A.S KEMAR mannequin type BC was measured according to the methods outlined earlier, with a radius of 1.2 meters (see Fig. 2.4). A one-second sine-sweep was used as the input signal, and an average over five sweeps was taken for each measurement

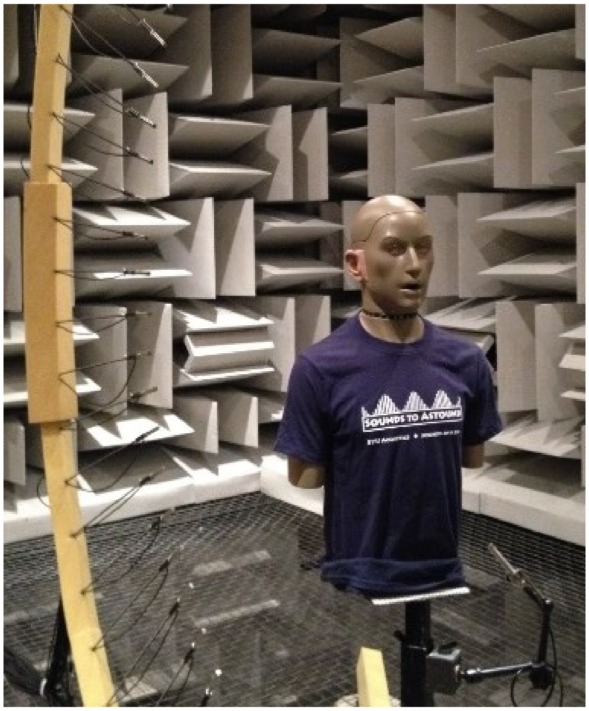


Figure 2.4. KEMAR mannequin centered in arc array for directivity measurements.

orientation. Coherence and directivity results for a few frequencies are shown in Fig. 2.5, with a coherence balloon on the left and a normalized directivity balloon computed from the FRF on the right. The magnitude of each is shown via both the color and radius of the balloon.

In each case, the coherence is close to 1 at each position on the measurement sphere. This is an indication of high signal-to-noise ratio and good FRF quality at each measurement angle. At low frequencies, the directivity is very nearly omnidirectional, as seen in Fig. 2.5 subplots (a) and (b). However, at 500 Hz [subplot (c)], one sees a slight tendency for stronger radiation in the lower hemisphere. Subplot (d) shows that the 1 kHz radiation is strong in the lower hemisphere, but it is also strong in a small region of the upper hemisphere. At 2 kHz [subplot (e)], the radiation is stronger forward and upward, with what appears to be a dipole-like behavior along an angled axis. At 4 kHz [subplot (f)], the directivity pattern is more complex, with multiple directions favored for strong radiation. However, a significant null persists along the angled axis

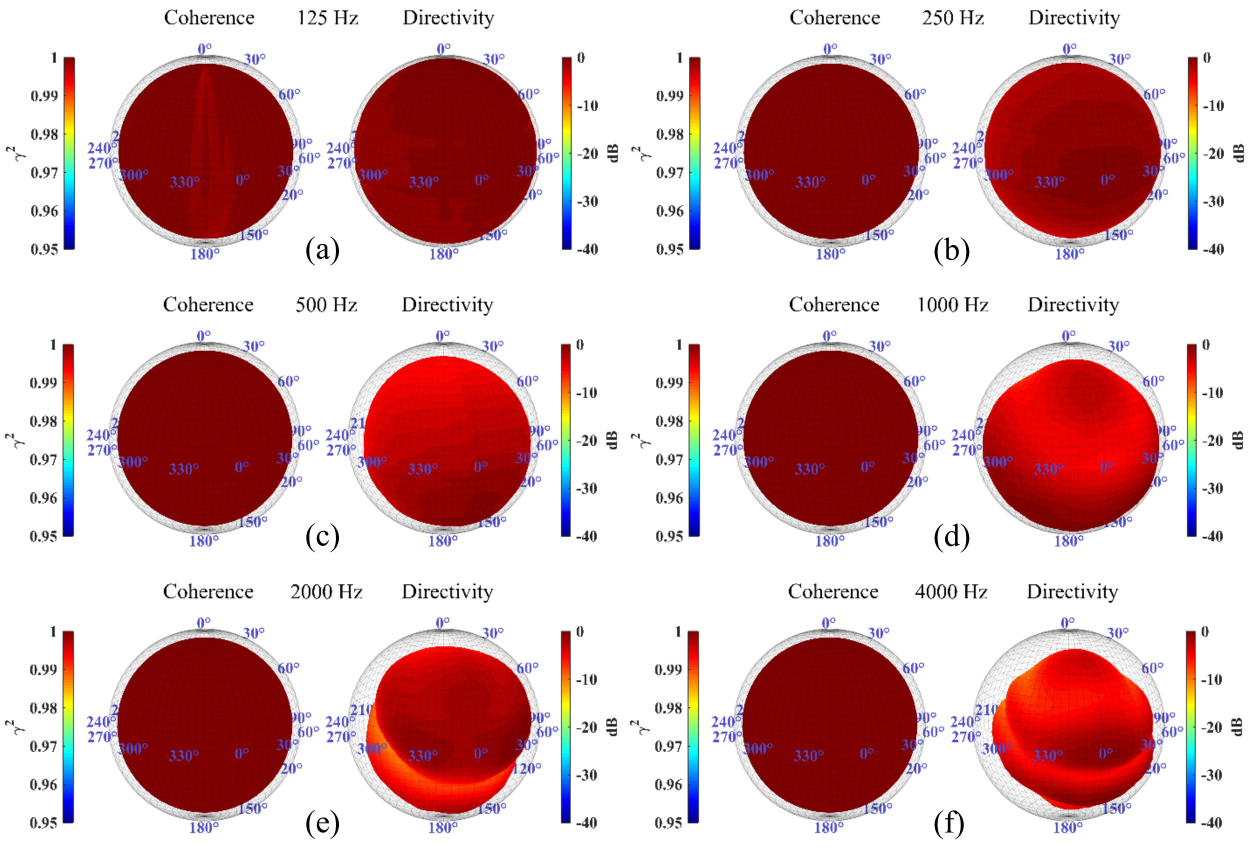


Figure 2.5. Coherence and directivity of the KEMAR measurements for the (a) 125, (b) 250, (c) 500, (d) 1000, (e) 2000, and (f) 4000 Hz third-octave bands. Within each subplot, the left plot depicts the coherence at each measurement point, and the right plot depicts the directivity pattern as calculated from the normalized FRFs at each measurement point.

as seen in subplot (e). Since KEMAR is often used in acoustic testing to mimic how a live talker would behave (Section 3.3.1), it is expected that the directivity results from KEMAR are similar to those of male and female speech.

2.3.2 Female

Three female native-English speakers agreed to participate in the study. They repeated a phonetically balanced passage of six sentences at each of the 72 measurement angles [55]. Their directivities were measured at a radius of 1.2 meters. An example of one of the subjects centered in the arc array is shown in Fig. 2.2. The directivities for the three subjects were averaged together as described in Section 2.2, to create the directivity balloons for female speech [57]. Coherence and directivity results for a few frequencies are shown in Fig. 2.6.

The directivity and coherence at the different frequencies very diverse results. At 125 Hz [Fig. 2.6 subplot (a)], the coherence balloon shows very poor coherence (less than 95%, corresponding to a signal-to-noise ratio of 13 dB). This is most likely due to a poor signal-to-noise ratio at this frequency, because the average female fundamental frequency is closer to 200 Hz. Due to the lack of speech energy at 125 Hz (as shown by the low coherence), there is a large degree of uncertainty associated with the directivity pattern at 125 Hz. However, at 250 Hz [Fig. 2.6 subplot (b)] and 500 Hz [Fig. 2.6 subplot (c)], one sees a tendency for the directivity to have stronger radiation in the lower hemisphere as frequency increases, similar to that shown for the KEMAR mannequin. At 500 Hz, additional side lobes in the upper hemisphere also emerge. In subplot (d) at 1 kHz, coherence starts to break down again. The null near the front of the directivity balloon may partially explain the poor coherence at the same location, because the signal-to-noise ratio is smaller at that location. The poor coherence could also be an indication of a poor measurement, which then resulted in an error in the directivity at that point. At 1 kHz, the tendency is for the sound to radiate more strongly upward and forward.

The differences between the low- and high-frequency results have noticeable effects for mic placement near a female talker. A warmer sound results from placing the mic below the horizontal plane, since the lower speech frqeuencies are more strongly radiated in that direction. In contrast, a cleaner, crisper sound results from placing the mic above the horizontal plane, since the higher speech frequencies are more strongly radiated upwards. This effect can be heard when comparing the sound of radio broadcasts in North America, where the mic is intentionally

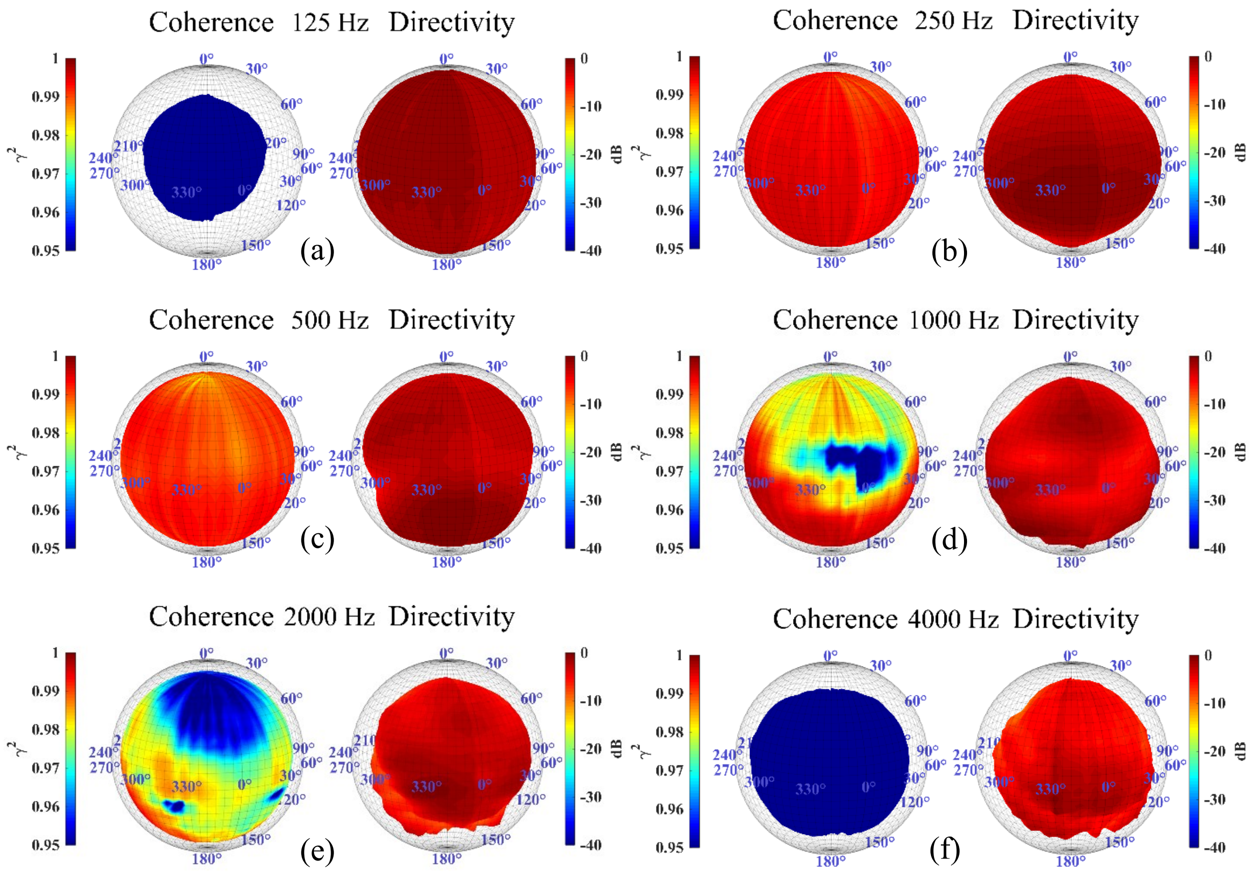


Figure 2.6. Average female coherence and directivity spherical plots for the (a) 125, (b) 250, (c) 500, (d) 1000, (e) 2000, and (f) 4000 Hz third-octave bands. Within each subplot, the left plot depicts the coherence at each measurement point, and the right plot depicts the directivity pattern as calculated from the normalized FRFs at each measurement point.

placed lower to achieve a warmer, more intimate sound, and England, where the mic is placed near the speaker's forehead for a crisper or brighter sound.

At 2 kHz [subplot (e)] and at 4 kHz [subplot (f)], the directivity pattern increases in complexity while the coherence decreases. The increase in complexity is similar to the trends seen in the KEMAR directivity data. Coherence could be lower simply because there is less speech energy at those frequencies, resulting in a lesser signal-to-noise ratio.

An additional factor that differed between the KEMAR and the live female and male directivity measurements is that the live talkers were seated in a chair, whereas KEMAR was positioned directly on a stand on the turntable. The radiation from the seated talkers involved diffraction from their bodies and the chair that was not included in the radiation from KEMAR. This may explain some of the roughness seen behind and below the talkers, although it is difficult to resolve the artifacts from the observation angle of the plots. A more clear view is shown in the animation in Appendix A.

2.3.3 Male

Four male native English speakers agreed to participate in this study. They repeated the same phonetically balanced passage of six sentences at each of the 72 azimuthal measurement angles. Coherence and directivity results for a few frequencies are shown in Fig. 2.7.

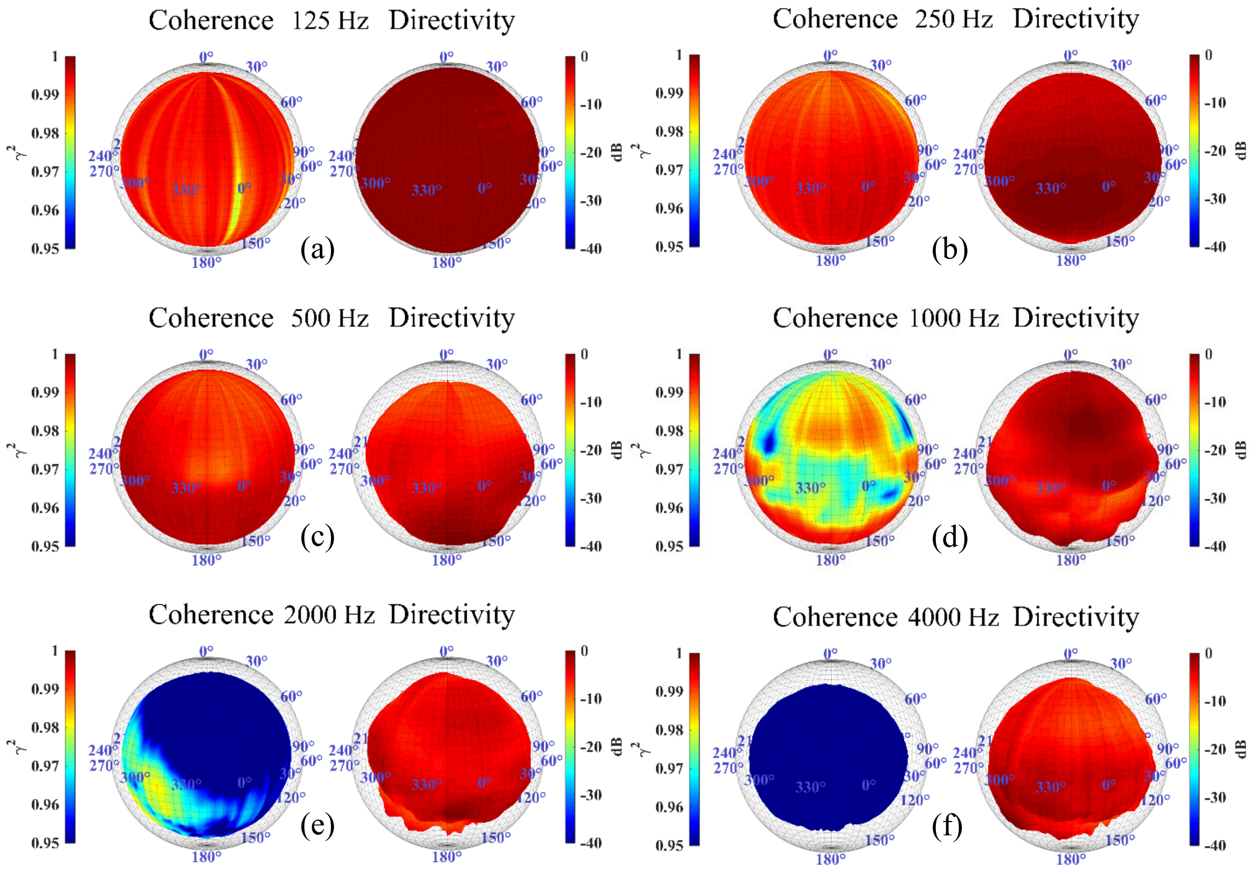


Figure 2.7. Composite male coherence and directivity spherical plots for the (a) 125, (b) 250, (c) 500, (d) 1000, (e) 2000, and (f) 4000 Hz third-octave bands. Within each subplot, the left plot depicts the coherence at each measurement point, and the right plot depicts the directivity pattern as calculated from the normalized FRFs at each measurement point.

The fundamental frequency of the male speaker is lower than that of the female speaker, so the coherence at the lowest frequencies is improved. At low frequencies, the radiation is nearly omnidirectional, similar to KEMAR and the female talkers, with some preference to the frontal direction at 250 Hz. At 500 Hz, the radiation is more strongly directed forward and downward. At higher frequencies, the signal-to-noise ratio lessens, coherence is poorer, and the directivity patterns are more complex. At 1 kHz, one can see a null along a tilted axis, similar to that seen in the KEMAR directivity at 2 kHz.

2.3.4 Comparison to Prior Work

Polar-plot information may be extracted from these three-dimensional balloon plots. An example of this for the KEMAR directivity data at 1 kHz is shown in Fig. 2.8. The upper left-

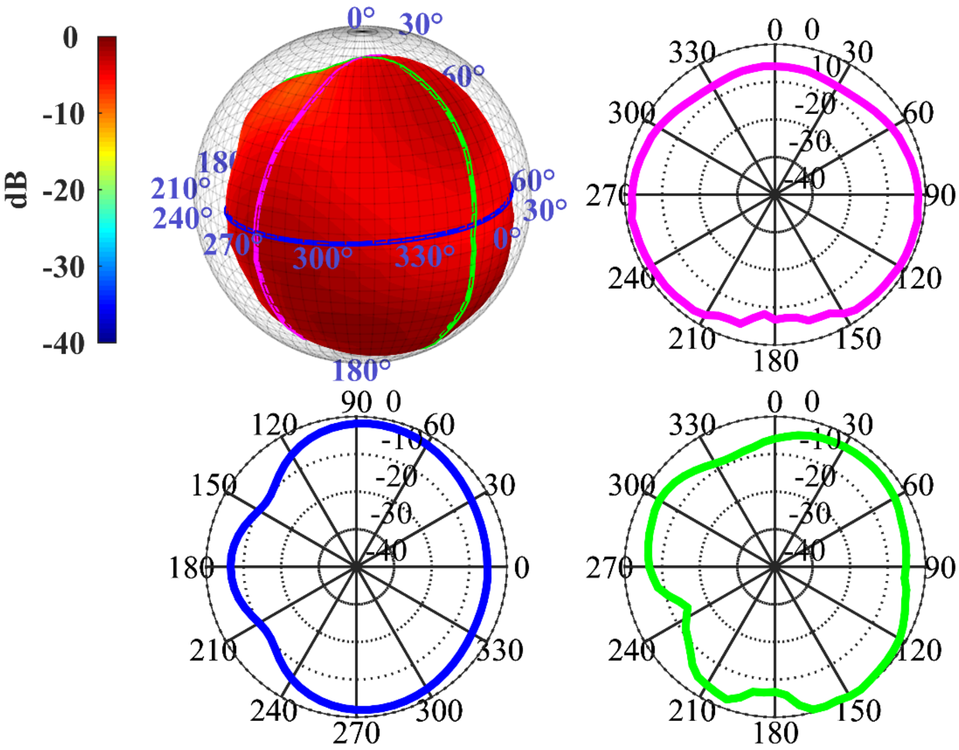


Figure 2.8. Three-dimensional balloon plot of KEMAR directivity at 1000 Hz (upper left), cross section at the transverse plane (lower left), cross section at the median plane (lower right), and cross section at the coronal plane (upper right).

hand plot depicts the balloon plot with superposed bands representing the mapping of the three polar plots. The blue curve divides the upper and lower hemispheres and is referred to as the horizontal, or transverse, plane. The green curve divides the left and right hemispheres and is referred to as the vertical, or median plane. The magenta curve divides the front and back hemispheres and is referred to as the frontal or coronal plane. These polar results are plotted against the results of Dunn and Farnsworth [38], and Chu and Warnock [41] at similar radii for several frequencies. These plots are shown in Figs. 2.9-2.12.

These figures allow for easy comparison between the different measurements. At lower frequencies (Figs. 2.9-10), near omnidirectional behavior is observed in all measurements. At higher frequencies (Figs. 2.11-12), the measured male and female data remain similar to each other, but the KEMAR directivity pattern differs. The Dunn and Farnsworth [38] and Chu and Warnock [41] data provide a reasonable match to these measurements for the main frequencies

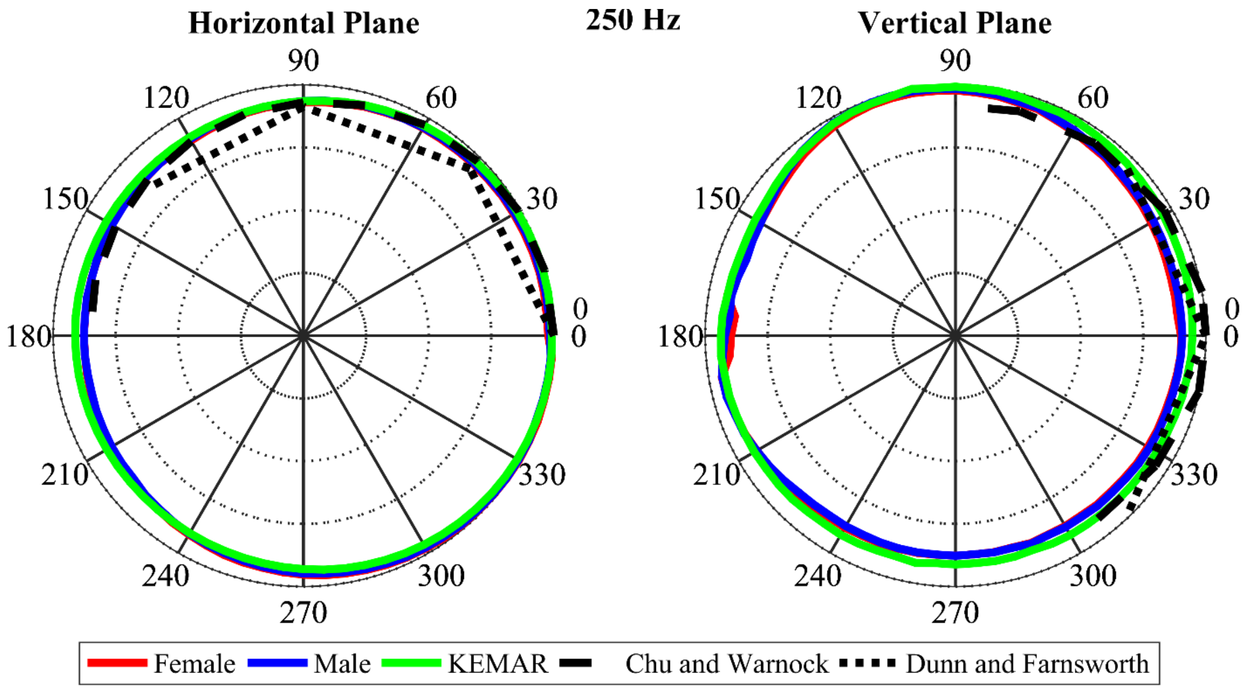


Figure 2.9. Comparison of directivity measurements to those of Dunn and Farsnworth, and Chu and Warnock at 250 Hz in the horizontal and vertical planes.

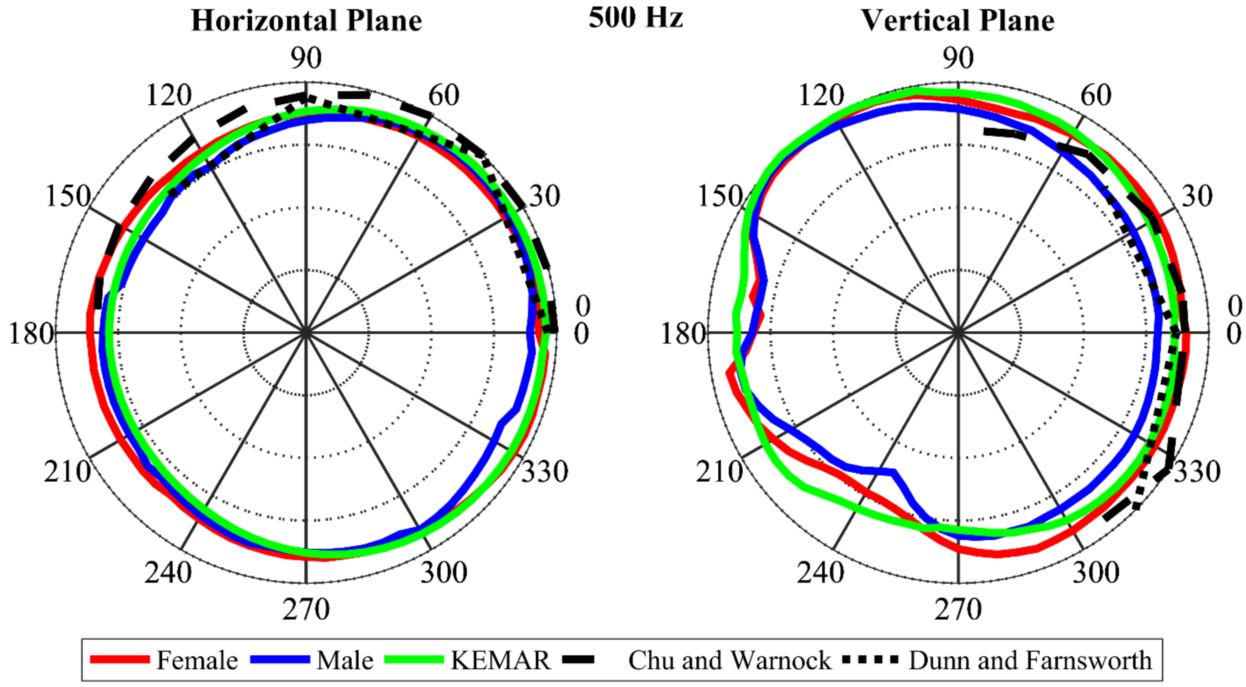


Figure 2.10. Comparison of directivity measurements to those of Dunn and Farsnworth, and Chu and Warnock at 500 Hz in the horizontal and vertical planes.

of human speech. However, since the data measured by Chu and Warnock and Dunn and Farnsworth were sparser, their results are angularly interpolated for a more direct comparison.

2.4 Discussion

The directivity data presented here were taken in the far field. As such, the directivity does not generally change with increased distance from the talkers, making it ideal for acoustic simulation software such as EASE. However, a smaller array of microphones placed at close range could have provided more information about the near field of speech directivity, especially regarding how sound diffracted around the head and torso. Near-field data could be used for studies involving spherical near-field acoustical holography and could be propagated as needed numerically to the far field [58]. Even from the 1.2m radius, the data could be used in spherical-harmonic expansions and exterior problems to better describe radiation. Studies such as these

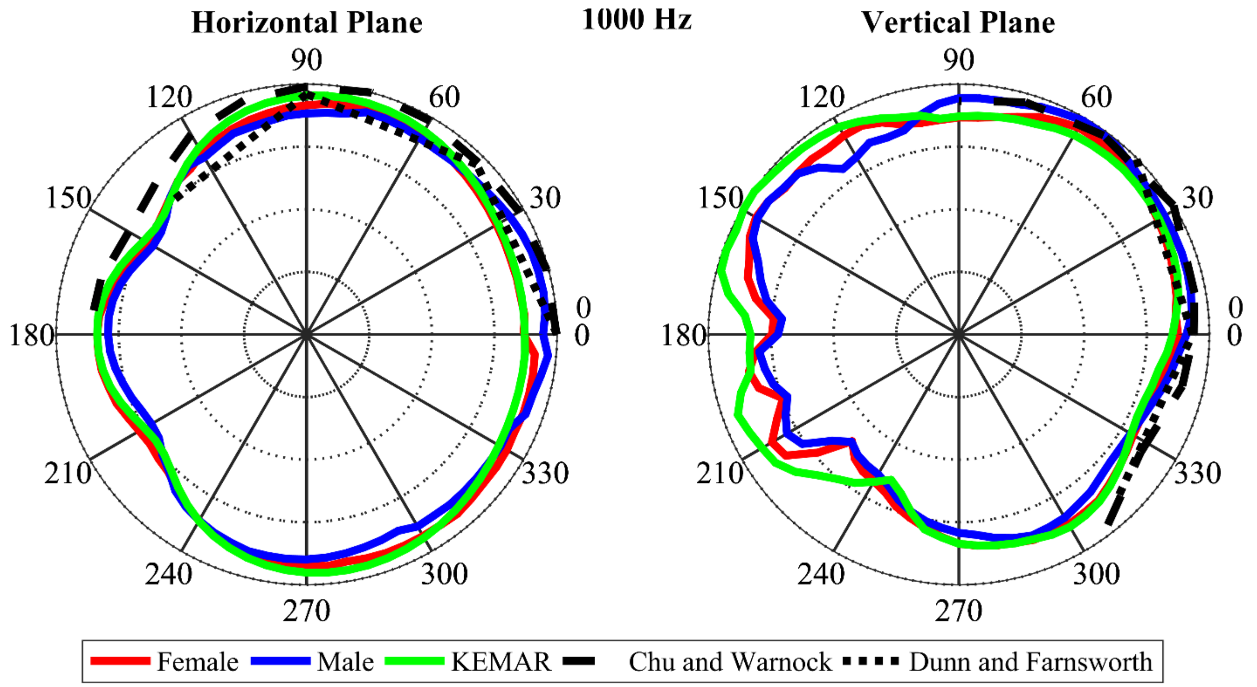


Figure 2.11. Comparison of directivity measurements to those of Dunn and Farsnworth, and Chu and Warnock at 1 kHz in the horizontal and vertical planes.

could establish a boundary between near-field and far-field for various frequencies of the human voice.

In order to keep the total time for the directivity measurements manageable for each subject and to prevent vocal fatigue, the study involved a short speech passage. A longer phonetically-balanced passage could provide more data for long-time average spectra, from which the directivity data would be computed.

While the subjects were somewhat restrained in their movements, some wiggling may have occurred over the course of the 2-hour measurements. An alternative approach could use a head tracking system to ensure the subject remains stationary during each rotation. However, the addition of head tracking could also affect the diffraction effects around the head.

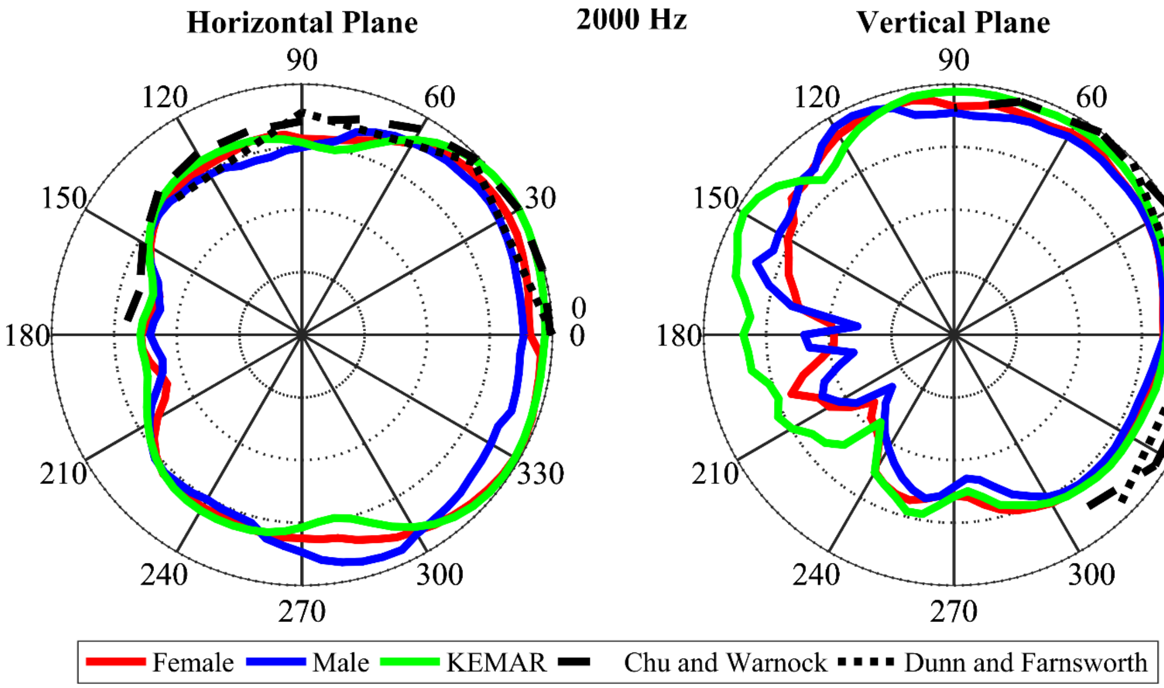


Figure 2.12. Comparison of directivity measurements to those of Dunn and Farsnworth, and Chu and Warnock at 2 kHz in the horizontal and vertical planes.

The directivity patterns include diffraction effects of the chair and rotation apparatus. A more detailed presentation that reveals these and other effects may be seen by viewing animations of the 3D balloon plots varying with frequency, available in Appendix A. An additional study would reveal the differences in the directivity patterns for standing subjects.

Lastly, the directivity patterns here have been "patched" because the bottom-most microphone in the array was obstructed by the rotation apparatus. A larger-radius array would have permitted the use of that microphone position.

2.5 Conclusions

The results presented in this chapter gave a detailed look at live speech directivity and how it changes with frequency. The measurements were presented as balloon plots over a full sphere, allowing one to see more detail in the radiation patterns than had been previously published. In addition, the angular resolution of these measurements was higher than has been previously published. The animations in Appendix A especially highlight this resolution.

Although uncertainty is introduced due to variability in live-talker for the repeated-capture system, general trends can be deduced about live speech. The reliability of the results is increased because the use of FRFs mitigates the variability uncertainty while coherence balloons indicate FRF quality and measurement points with inadequate data. Regions with high coherence lead to valid measurements from which one may draw conclusions about the general trends in live speech. At low frequencies, the speech radiation is nearly omnidirectional. At frequencies near 500 Hz, the radiation is dominant below, while also strong toward the sides and front. At frequencies near 1 kHz, the radiation dominance shifts above the horizontal plane and more strongly radiated forward than toward the sides. Further study on live speech directivity will allow researchers to improve the measurement method and analysis techniques to more fully understand the results.

Chapter 3

Oral-Binaural Room Impulse Responses

3.1 Definitions and Background

A room impulse response (RIR) characterizes the effect of the acoustic environment on an impulse as it travels between an acoustic source and receiver at discrete points in space. As depicted in Fig. 3.1, sound emitted from a source reflects off surfaces of the acoustic space and eventually arrives at the receiver. These RIRs are particularly useful in that several architectural acoustic parameters, including reverberation time (RT), clarity, and others, can be derived from them. These parameters help characterize the acoustic space.

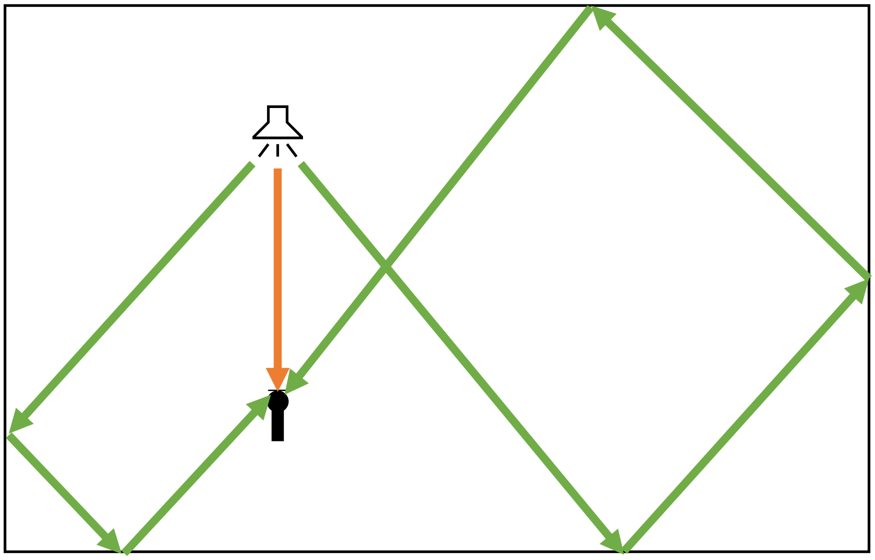


Figure 3.1. Example of ray tracing reflections in a room. An acoustic source emits sound. The orange ray indicates direct sound arriving at the receiver first. The green rays depict two indirect reflected sounds arriving after being reflected off the room surfaces.

In more general terms, an impulse response (IR) completely characterizes a linear time-invariant system between two points [59]. An IR is the linear system response to an impulsive input signal. The convolution of an input signal with the IR yields a specific output signal that includes the effect of the linear system. In electro-acoustics, the convolution of an RIR with an input signal results in an auralization, or a simulation of how sound behaves in a given acoustic space.

Such simulations of external sounds are realistic. However, for realistic auralization of one's own voice, the binaural aspects of hearing must be accounted for. In the present work, the simulation of one's own voice in different acoustic environments requires the use of oral-binaural room impulse responses (OBRIRs). These actually comprise two RIRs with the acoustic source (the mouth) positioned very close to two receivers (the ears). An example of the two types of sound paths from the mouth to the ears in a reflective space is shown in Fig. 3.2. OBRIRs

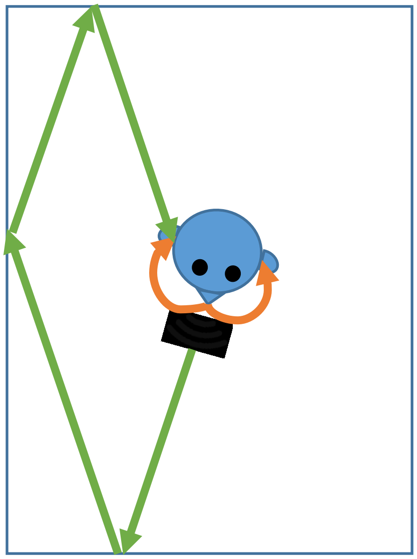


Figure 3.2. An example of sound paths from a mouth to the ears. Sound is emitted from the talker's mouth. The orange ray indicates the sound that is diffracted around the head and arrives at the ears first. The green rays represent the sound emitted from the mouth that then reflects off the surfaces in the room before arriving at the ears. A more complete diagram for an OBRIR would show many rays indicating the talker's directivity pattern, and the many directions from which the sound arrives at the ears.

contain sound from both types of sound paths and characterize an acoustic space from the point of view of a talker/listener.

3.2 OBRIR Measurements by Cabrera and Yadav

The measurement and manipulation of an OBRIR for use in a RTCS is not an insignificant matter. The method used by Cabrera et al. [20] is discussed in this section. It is compared with the method used in the present work in Sec. 3.3. To establish an OBRIR, Cabrera et al. used a Brüel and Kjær 4128C HATS with a Brüel and Kjær Type 4939 ¼" microphone at the "mouth reference point," 6 mm in front of the face plane and 25 mm from the "center of lip" point. They placed Brüel and Kjær type 4101 microphones at the entrance to the HATS ear canal simulators to avoid measuring the ear canal resonances. The OBRIR measurement was made by sending a swept-sine signal (50 Hz to 15 kHz with a logarithmically constant sweep rate and 15 s duration) to the HATS mouth simulator and recording it at each of the mouth and ear microphones. Four signals were sent to a recording device: "a signal suitable for deconvolving the IR from the sweep" and the three signals recorded by the three microphones. This yielded the IRs from the signal generator to each of the three microphones. Subsequently, the IRs from the mouth microphone to each of the ear microphones were obtained.

The IRs were obtained via the following process. First, the mouth microphone IR was zero padded to be twice the length of the desired IR (for an anechoic environment, the total window length was 2^{16}, and for a reverberant environment of mid-frequency RT = 2.5 s, a window length of 2^{18} was used). The direct sound in this IR was identified by the peak maximum absolute value. Data from −2 to +2 ms around this peak was used with a Tukey window applied (50% fade in/out, 50% constant). The ear microphone IRs were also zero padded on the second half. The Fourier transform of each of these modified IRs was performed. The frequency response function (FRF) was then computed by dividing the cross-spectrum between the mouth IR and the ear IR by the autospectrum of the mouth IR. This FRF was then filtered to be within 100 Hz to 10 kHz. The result was then inverse Fourier transformed and truncated to discard the latter half. The resultant IR was subsequently multiplied by the ratio of rms values of a calibration signal recorded on each channel to compensate for differences in gain between channels of the recording system.

Cabrera et al. additionally investigated OBRIR measurements using human talkers. In these measurements, the IRs were not immediately available. Instead, the average cross-spectrum and average auto-spectrum were computed from 10 minutes of continuous speech. Again, the OBRIR as filtered to be within the range 100 Hz to 10 kHz, justified by the poor signal-to-noise ratio above 10 kHz, where not much speech energy is available. The reliability of the FRF was estimated with the associated coherence function, computed from the average cross-spectra and autospectra, but the results of the reliability estimates were not shown in their published work.

3.3 OBRIR Measurements for this Work

The OBRIRs used in the real-time convolution system (RTCS) of the present work were measured either in a room using a KEMAR mannequin or produced using the acoustical simulation package EASE. This section provides information about the KEMAR mannequin and why it was used to perform OBRIR measurements. In addition, an explanation is given of how the directivity and HRTF of the KEMAR mannequin was used in EASE to produce the simulated room OBRIRs, thus making the measured and simulated OBRIRs comparable.

3.3.1 KEMAR Properties

KEMAR is a HATS currently produced by G.R.A.S. that meets the requirements of ANSI S3.36/ASA58-1985 and IEC 60318-7:2011 and is based on ITU-T P.58 [60,61]. It was constructed from the anatomical averages of 5,000 U.S. Air Force men and women from a 1950 survey and is meant to simulate the way an average adult head influences a sound field [62]. This unique development implies that the KEMAR HATS has the same acoustical properties as an average human, including facial features. The head and torso dimensions of KEMAR are all within 4% of the male and female median values for those dimensions [62]. The concha dimensions for KEMAR were derived from averages of 12 males and 12 females.

Burkhard noted that the pressure at an ear canal entrance exhibited similar dependence on the concha and sound diffraction around the head and torso for both KEMAR and actual people [62]. Although the use of an anatomically-averaged HATS neglects the unique features of individuals, KEMAR is a practical tool to ensure reasonable simulation of acoustical diffraction and other properties similar to those encountered by live talkers.

The spectral bandwidths of the KEMAR mouth and ear simulators also mimic human features. The KEMAR mouth simulator can produce a signal up to 100 dB re. 20 μPa. The mouth simulator can be equalized over the range 100 Hz to 10,000 Hz. The KEMAR ear simulator was not used here; instead, ear microphones were coupled directly to the pinnae at the ear-canal openings for the OBRIR measurements. This was similar to making measurements at the entrance of a blocked ear canal of a human, as discussed by Moller [15]. The microphones used in the KEMAR ears were G.R.A.S. ½” type 40 AO, which were flat over the range of 5 Hz to 12.5 kHz [61].

3.3.2 Measurement Procedure

A software package was used with an RME Fireface digital-audio interface to measure the OBRIRs of the KEMAR mannequin. The Electronic and Acoustic System Evaluation and Response Analysis (EASERA) software created by the Ahnert Feistel Media Group (AFMG) is a package for data acquisition for electronic and acoustic systems [63]. EASERA is especially useful for acoustic IR measurements. The purpose of these OBRIR measurements was to characterize acoustic spaces for use in the RTCS. Specifically, they assessed the IRs between the signal sent to the KEMAR mouth simulator and the signals from the KEMAR left and right ear microphones. The IR between the signal sent to the KEMAR mouth simulator and the signal out of the head-worn microphone were also measured. The signal used to drive the room was a pink-weighted swept sine between 10 Hz to 24 kHz, repeated 10 times, with the first sweep serving as a “presend” to excite the room, and the remaining nine allowing averaging. The length of the sweeps varied depending on the anticipated reverberation time of the room in which measurements were performed. The sampling rate was 48,000 Hz.

The process may be characterized in terms of a single-input, dual-output system, as depicted in the block diagram of Fig. 3.3. This block diagram gives the signal flow for a KEMAR OBRIR measurement in a room: $A(f)$ is the composite FRF through the EASERA software, the PreSonus FireFace, a Crown D-45 power amplifier, and the KEMAR mouth simulator. The radiated acoustic pressure signal at the point near the mouth simulator is $\hat{a}_k(f)$. The FRFs for the diffraction paths of the signal around the mannequin head to the left and right ear canal entrances are $D_{K,L}(f)$ and $D_{K,R}(f)$, respectively. The FRF $M(f)$ represents the transduction and signal-conditioning path of the head-worn microphone, and $\hat{m}_s(f)$ is the recorded signal. The FRF (i.e., Fourier transform corresponding to the RIR) from the point close

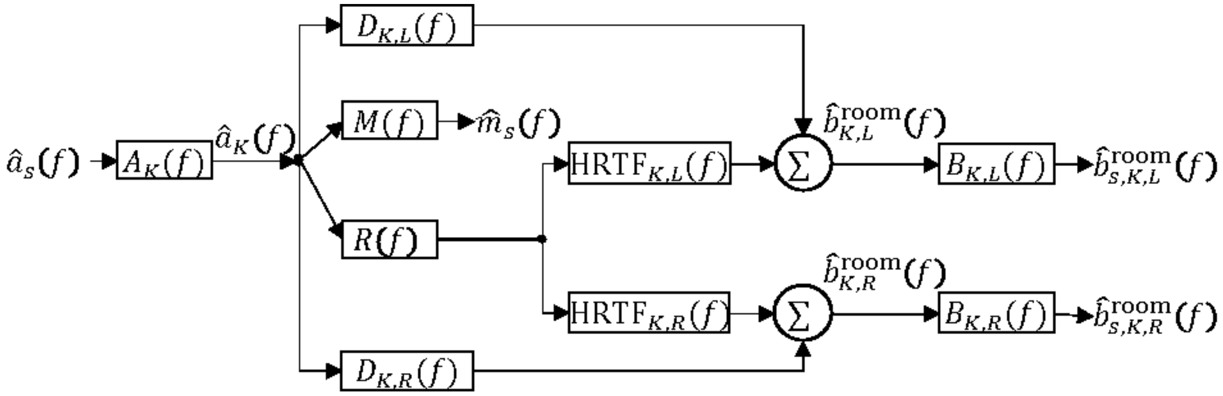


Figure 3.3. Block diagram of the signal flow for a room measurement with KEMAR. The digital input signal $\hat{a}_s(f)$ is modified by the KEMAR mouth simulator before arriving at a hypothetical point near the mouth, where it is identified as the signal $\hat{a}_K(f)$. The signals at the entrance to the ear canals, $\hat{b}_{K,L}(f)$ and $\hat{b}_{K,R}(f)$, are further modified by the ear microphones and recording hardware before being recorded as digital waveforms, $\hat{b}_{s,L}(f)$ and $\hat{b}_{s,R}(f)$. The dual output signals are the sums of the diffracted sound and the room response with HRTFs. The subscript K indicates the transfer function dependencies on the KEMAR anatomy, as opposed to that of a live talker. In addition, the signal $\hat{m}_s(f)$ represents the signal recorded by a head-worn microphone near the corner of the KEMAR mouth simulator.

to the mouth simulator and the unobstructed central head point is $R(f)$, while $\mathrm{HRTF}_{\mathrm{K,L}}(f)$ and $\mathrm{HRTF}_{K,R}(f)$ are the KEMAR HRTFs for the left and right ears, respectively. The acoustic pressure signals at the ear canal openings are $\hat{b}_{K,L}(f)$ and $\hat{b}_{K,R}(f)$, respectively. In addition, $B_{K,L}(f)$ and $B_{K,R}(f)$ are the composite FRFs from the ear canal openings through the left and right KEMAR ear microphones, the corresponding phantom-to-ICP power converters, the FireFace preamplifiers and A/D converters, and to the EASERA software. The recorded signal for the left ear is $\hat{b}^{\mathrm{room}}_{s,K,L}(f)$ and that for the right ear is $\hat{b}^{\mathrm{room}}_{s,K,R}(f)$.

EASERA does not give the recorded signals, but instead gives the computed IRs corresponding to $\mathrm{IFFT}[\hat{b}^{\mathrm{room}}_{s,K,L}(f)/\hat{a}_s(f)]$ and $\mathrm{IFFT}[\hat{b}^{\mathrm{room}}_{s,K,R}(f)/\hat{a}_s(f)]$. The Fourier transforms of these IRs show the relationship of the signals in the frequency domain:

$$H^{\mathrm{room}}_{K,L}(f) = \frac{\hat{b}^{\mathrm{room}}_{s,K,L}(f)}{\hat{a}_s(f)} = A(f)\big[\, D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f)\big]\, B_L(f), \tag{3.1a}$$

$$H_{K,R}^{\text{room}}(f) = \frac{\hat{b}_{s,K,R}^{\text{room}}(f)}{\hat{a}_s(f)} = A(f)\big[D_{K,R}(f) + R(f)\text{HRTF}_{K,R}(f)\big]B_R(f), \tag{3.1b}$$

$$\frac{\hat{m}_s(f)}{\hat{a}_s(f)} = A(f)M(f). \tag{3.1c}$$

3.4 OBRIR Modeling with EASE

The part of the RTCS most crucial to producing realistic auralizations is constructing appropriate RIRs to be used within the convolution software. This section discusses the details of creating a RIR for a modeled room in the simulation package EASE, which enables easy manipulation to create desired responses, even unrealistic responses, whereas RIRs measured from actual rooms are fixed. Once a room has been modeled in EASE with the appropriate dimensions and surface properties (including absorption and scattering coefficients), the RIR may be generated for a given source and receiver by inserting a “speaker” with appropriate orientation at the source location, and a “listener seat” with appropriate orientation at the receiver location. The application of these RIRs to create specific OBRIRs used in the RTCS are discussed in Section 3.5, along with the architectural parameters of interest in both modeled and actual (measured) rooms.

EASE uses a hybrid computation method to generate the IR between the source and receiver. The earliest reflections are calculated using image-source methods. The later reflections are computed using ray tracing. Only the latest reflections are represented by a diffuse tail. These two methods rely on the far-field directivity of the source. The rays are traced until they intersect with a virtual bubble of a defined diameter around the receiver position (the unobstructed center of the head) or they exceed the allowed number of reflections in the computation. The resulting response file (*.rsp, representing an RIR) is then convolved with the HRTFs to create a binaural response file (*.bir, representing a BRIR). The latter are often used in the EASE EARS module to better simulate how a person would experience the virtual room. The EARS module allows for the creation of auralizations [12]. In it, the BRIR is convolved with desired program material and played to a listener via headphones.

A block diagram for an auralization made with an EASE-computed RIR is shown in Fig. 3.4. Here, $\hat{a}(f)$ is a dry input signal that is convolved with a *.bir file computed by EASE. As

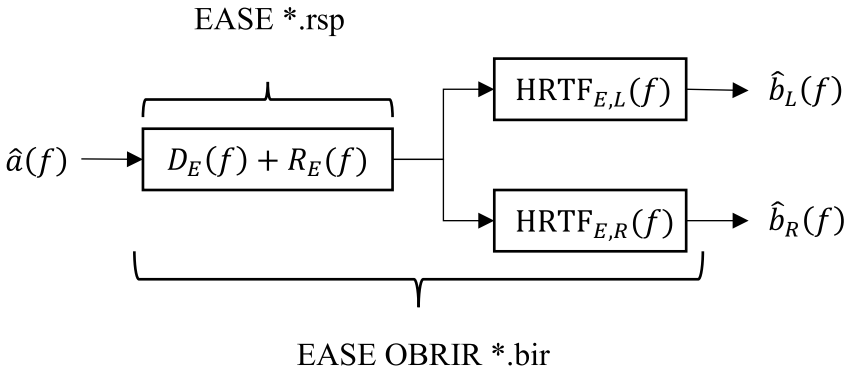


Figure 3.4. Block diagram of an OBRIR as created in EASE. A dry, monaural input signal $\hat{a}(f)$ is convolved with a BIR (*.bir) to create a stereo output signal $\left(\hat{b}_L(f)\ and\ \hat{b}_R(f)\right)$. A *.bir file is made by convolving an RIR file (*.rir) with a desired HRTF.

indicated earlier, the *.bir file is the result of the convolution of an RIR (*.rir file) with the desired EASE-computed HRTFs [$\text{HRTF}_{\text{E,L}}(f)$ and $\text{HRTF}_{E,R}(f)$]. The *.rsp file is the result of the computation EASE performs to characterize the modeled room. It includes both the direct path between the source and receiver [$D_E(f)$], and the reflections from the room as calculated by EASE [$R_E(f)$]. The resultant stereo signal [$\hat{b}_L(f)$ and $\hat{b}_R(f)$] is the auralization, which is ideally the sound a listener would hear if in the same position as the listener seat within the modeled room.

Of interest to this project is the creation of oral-binaural RIRs (OBRIRs). In this case, the “speaker” is the mouth of a talker (with appropriate directivity) and the “listener seat” is the effective center of the head of the same person, without the person being present. In principle, with the source and receiver appropriately placed, EASE should be able to compute the response from the mouth to the ears of a talker within a room. Auralizations made with this response are what a talker would hear if he or she were speaking in the modeled room. However, the OBRIR for the scenario of closely-located source and listener (mouth and ears) does not appropriately account for the direct diffracted sound from the mouth to the ears.

For example, consider the simple room depicted in Fig. 3.5. The speaker used in this room has the properties (sound power and directivity) of a KEMAR mannequin. The listener seat is positioned slightly behind and upward from the speaker, just as the center of the head is positioned behind and upward from the mouth. The dimensions of a KEMAR mannequin were used to position the listener seat precisely [62] with a tolerance of one centimeter.

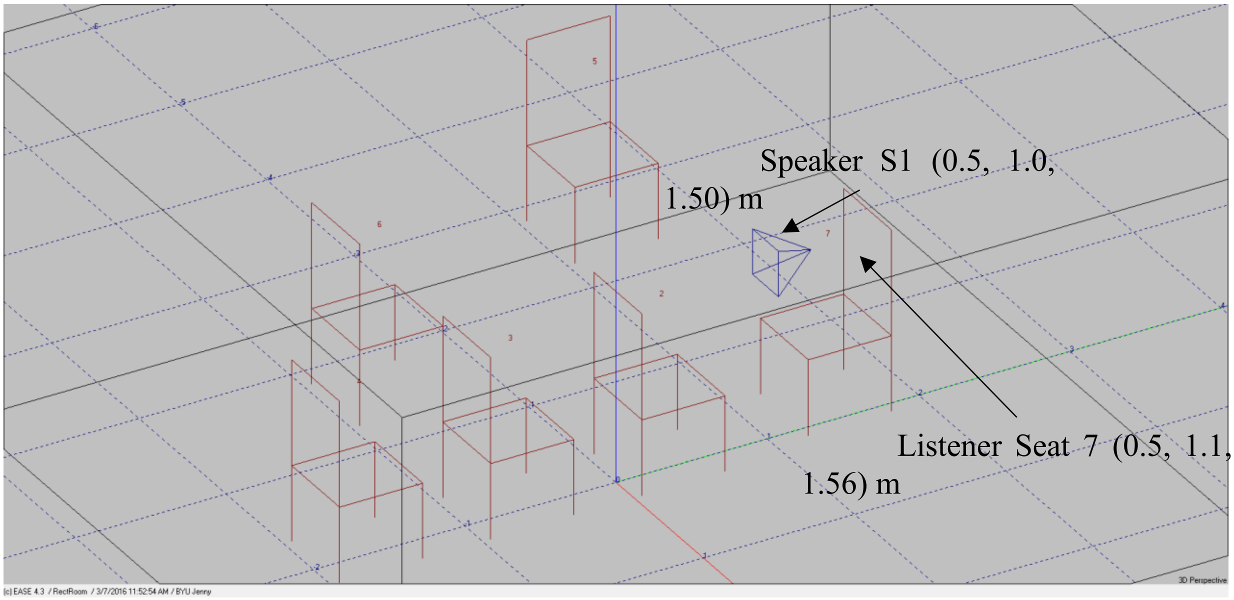

Speaker S1 (0.5, 1.0, 1.50) m
Listener Seat 7 (0.5, 1.1, 1.56) m
(c) EASE 4.3 / RectRoom / 3/7/2016 11:52:54 AM / BYU Jenny
3D Perspective

Figure 3.5. Room modeled in EASE with the speaker and listener seat locations depicted. The center of the head (listener seat) is 10 cm behind and 6 cm above the source (speaker) position. These values were chosen based on the dimensions of a KEMAR mannequin head

Some modification of the simulated reflectogram is necessary to create the OBRIR. Figure 3.6 depicts the early reflectogram made with the room, source, and receiver of Fig. 3.5. The reflectogram is a representation of the relative levels and arrival times of each of the reflections at a given frequency. The zeroth order reflection, or the direct sound between the source and receiver, is highlighted in blue at 0.339 ms. Traditional EASE calculations do not place the source and receiver this closely, so the direct sound is usually a valuable consideration. However, here, the direct sound should not be included in the calculated room response because it is nonphysical. There simply is not a way for sound from the mouth to reach the center of the head because a head obstructs the path. In any case, we are not interested in the sound at the center of the head, but at the entrances to the ear canals. Fortunately, the simulated direct sound can be removed from the RIR, and the result can be convolved with a desired head-related impulse response (HRIR, inverse Fourier transform of the HRTF) to create a binaural response file that only includes impacts from the room. (This is done by selecting the undesirable pulse

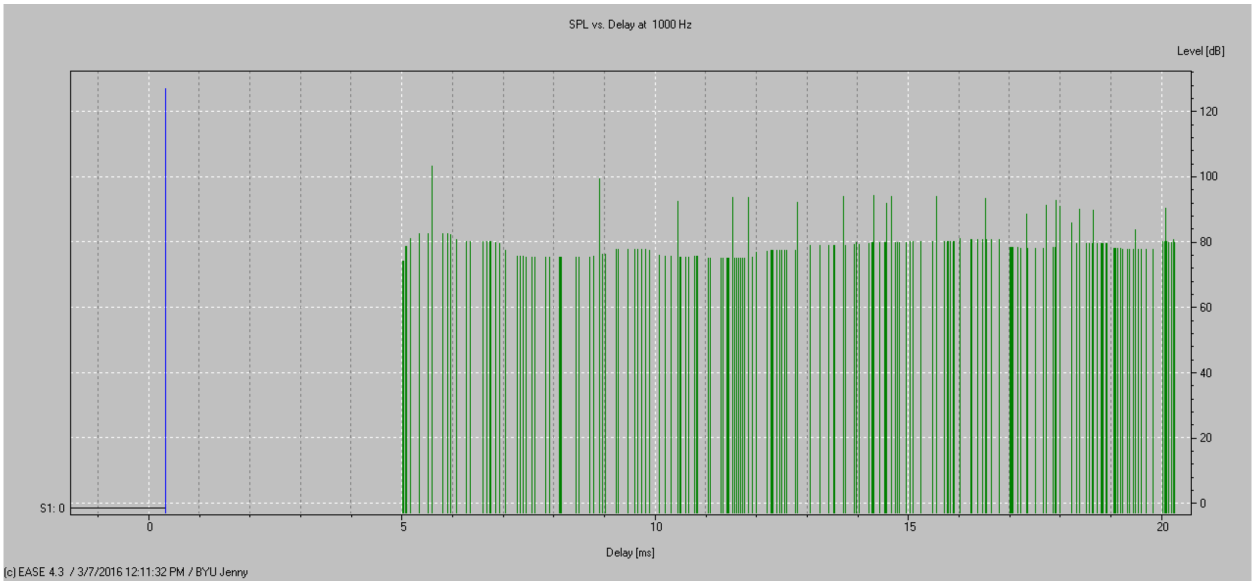


Figure 3.6. Reflectogram of EASE AURA response file from 0 to 20 ms. The direct sound pulse at 0.339 ms is highlighted in blue, while the other reflection orders are shown in green.

and deactivating it in the EASE Probe function.) The modified reflectogram is shown in Fig. 3.7. The reflectogram is then saved as a *.rsp file in EASE.

The OBRIR is calculated in the EASE EARS module wherein the response file (*.rsp), calculated for the point at the center of the head, is convolved with an HRTF. In the frequency domain, the HRTF is defined by Vorlander to be the sound pressure measured at the ear canal entrance divided by the sound pressure measured with a microphone at the center of the head, but with the head absent [11]. This is the function EASE uses. Several options for HRTFs in EASE are available, including the HRTF for a KEMAR mannequin, which was used to create the OBRIR shown in Fig. 3.8. This is a *.bir file in EASE. The binaural response is converted to *.wav format and loaded into the RTCS to simulate the experience of speaking in that simulated room. The computation parameters for the simulation are saved in tabular format. These tables

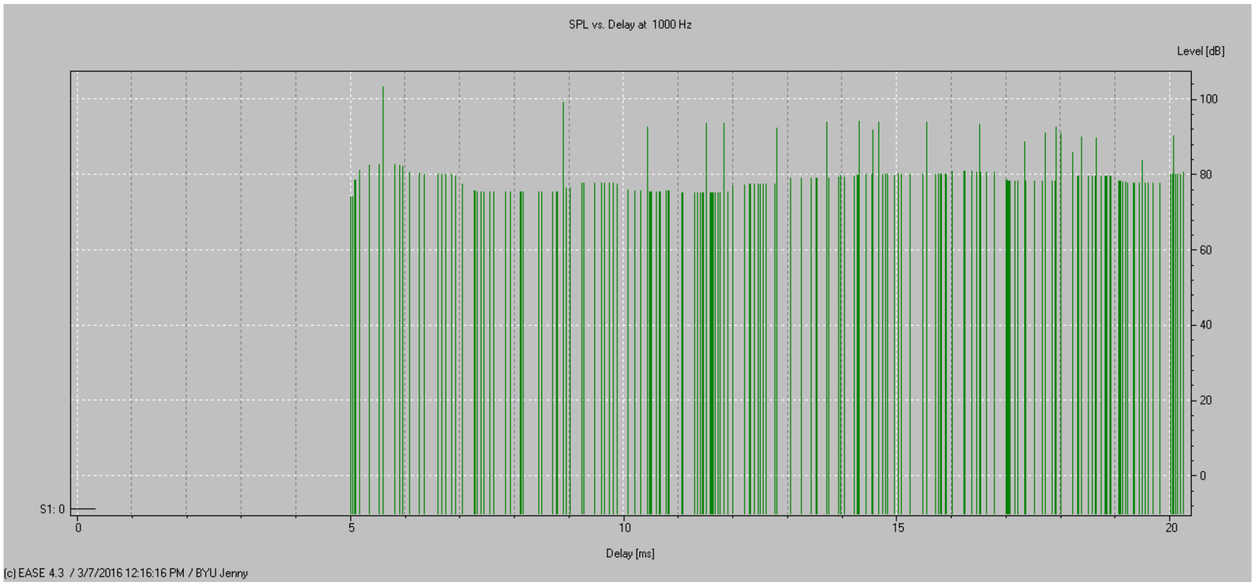


Figure 3.7. Reflectogram of the modified room response with the direct sound arrival removed.

keep track of the parameters used and provide consistency when creating multiple simulations of similar spaces.

3.5 Measured and Modeled OBRIR Characteristics

Several rooms were measured and modeled for use in the RTCS. The following sections summarize their measurements and calculations. These are: (1) several configurations of a reverberation chamber with varying amounts of added absorption, (2) two classrooms of different sizes, and (3) a large concert hall. These spaces were chosen for consideration as they

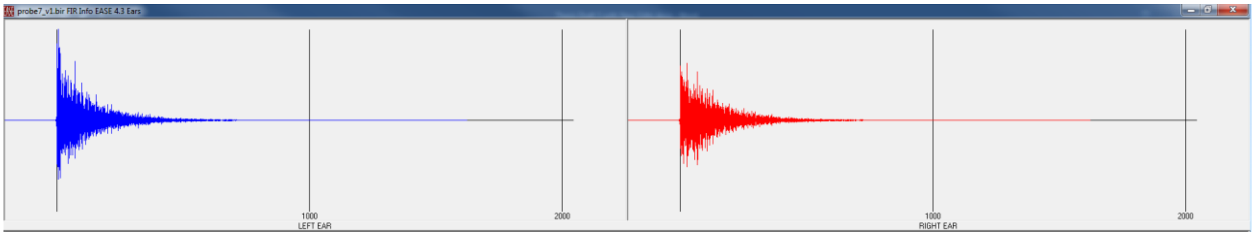


Figure 3.8. EASE binaural IR calculated for closely located source and receivers simulating the mouth and ears of a talker.

represent a wide range of acoustic spaces with both favorable speaking conditions, and detrimental speaking conditions.

3.5.1 Measured OBRIRs

3.5.1.1 *Reverberation Chamber*

The large reverberation chamber in the Eyring Science Center (ESC) at Brigham Young University (BYU) has reflective surfaces and stationary diffusers to create a nearly diffuse sound field over many audible frequencies. The room is rectangular, with a volume of 204 cubic meters [22,64]. For this work, its acoustical characteristics were altered through the addition of absorbing foam wedges to the floor of the room. Each was cut from 32 kg/m^3 open cell polyether foam rubber with a 94.5 cm overall depth, a 30.5 by 30.5 cm base and a profile similar to those suggested by Beranek and Sleeper [65]. The number of wedges introduced into the room for each configuration was 0, 2, 4, 8, 16, 24, or 32. The addition of the wedges served to lower the reverberation time and increase clarity. Traditional RIR measurements were made using a dodecahedron loudspeaker as the source, and a GRAS 40AE 12.7 mm (0.5 in) free-field microphone, with a random-incidence corrector and a Larson Davis PRM426 preamplifier, as the receiver. Table 3.1 summarizes the measured room characteristics for the various absorbing-wedge configurations in the reverberation chamber.

Table 3.1. The measured room characteristics for the various absorbing-wedge configurations in the reverberation chamber. The addition of the absorbing wedges served to reduce the reverberation time and increase clarity.

Wedges	EDT (s)	T_{10} (s)	T_{20} (s)	T_{30} (s)	C_{50} (dB)	%AL_{cons}
0	4.92	4.52	3.83	3.58	-7.57	19
2	3.96	4.04	3.67	3.36	-5.52	15
4	3.27	3.43	3.24	3.01	-4.92	14

8	2.46	2.47	2.49	2.33	-3.78	11
16	1.63	1.64	1.69	1.71	-1.47	8
24	1.36	1.42	1.38	1.40	-0.13	7
32	1.05	1.13	1.17	1.14	0.87	6

Several OBRIR measurements in the reverberation chamber were made using the KEMAR mannequin and the methods described in Sec. 3.4. In addition to the absorbing wedges, a researcher was in the room with KEMAR during the measurement. The researcher was included because the reverberation in the chamber was very sensitive to even minute additions in absorption, such as that added by the presence of a person. The KEMAR OBRIRs were meant to represent the situation of an interviewer and interviewee during a vocal effort study, so a representation of each person was present during the OBRIR measurements. Figure 3.9 gives an example of one configuration of absorbing wedges, and the positioning of KEMAR and the researcher during the OBRIR measurement with several absorptive wedges.

3.5.1.2 *Classrooms: ESC C215 and C261*

OBRIR measurements were made in two classrooms in the Eyring Science Center (ESC). C215 is a mid-sized lecture hall that seats 167 people. The OBRIR was made with KEMAR sitting in a seat to the side and front of the room. Figure 3.10 shows the positioning of the head-worn microphone at the corner of its mouth simulator. The room was empty aside from the KEMAR mannequin and the researcher performing the OBRIR measurement.

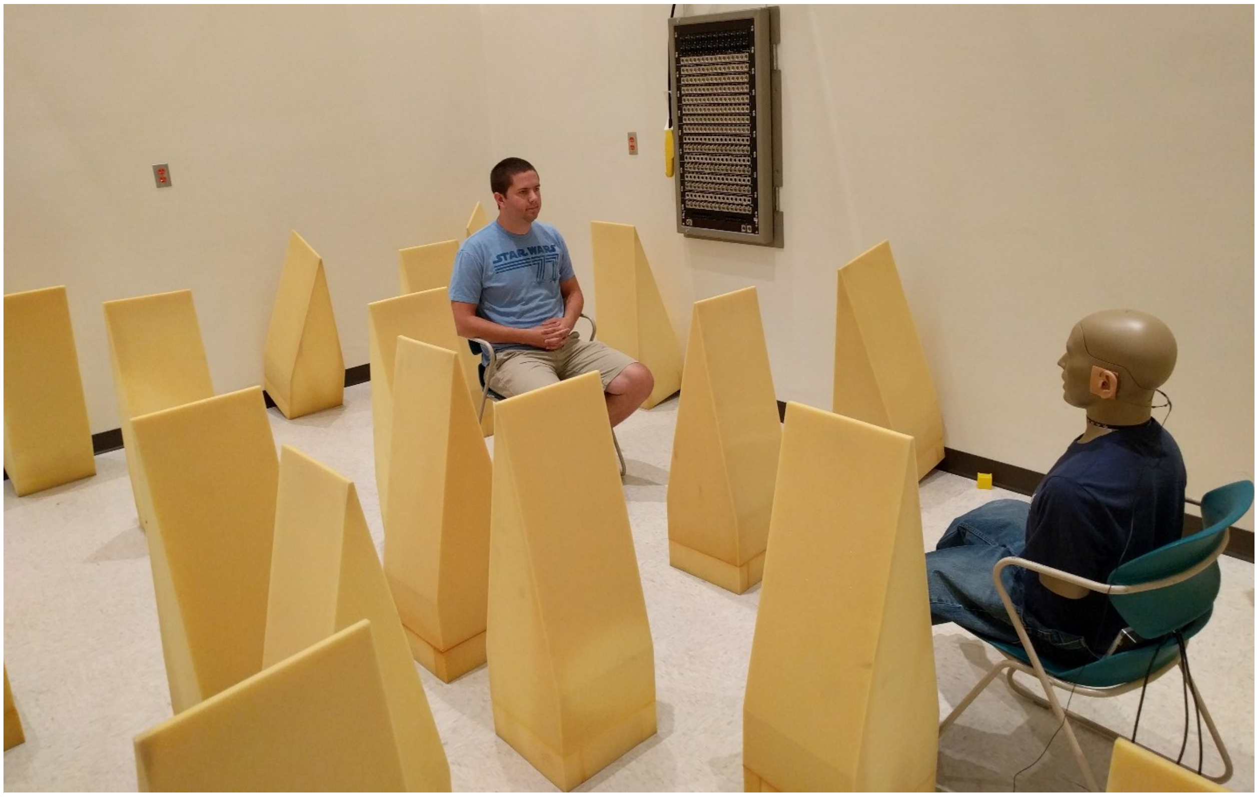


Figure 3.9. KEMAR and researcher positions during an OBRIR measurement in the reverberation chamber. The presence of 32 absorbing wedges in the room reduces the reverberation time and affects the resulting OBRIR.

Room C261 is a smaller classroom that seats about 40 people. For this case, the OBRIR was measured with KEMAR in a teaching position at the front of the room, facing the desks, as shown in Fig. 3.11. The room was empty aside from the KEMAR mannequin and the researcher operating the hardware.

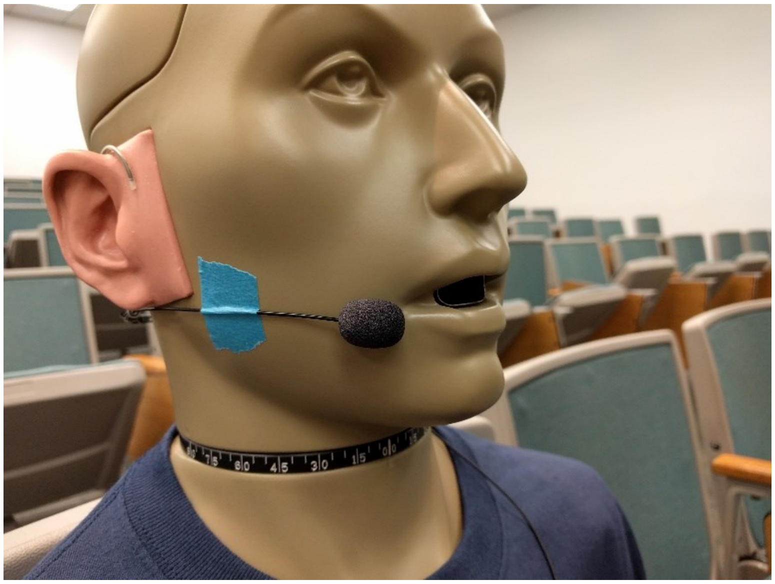


Figure 3.10. An OBRIR measurement in C215. The positioning of the head-worn microphone on KEMAR is shown.

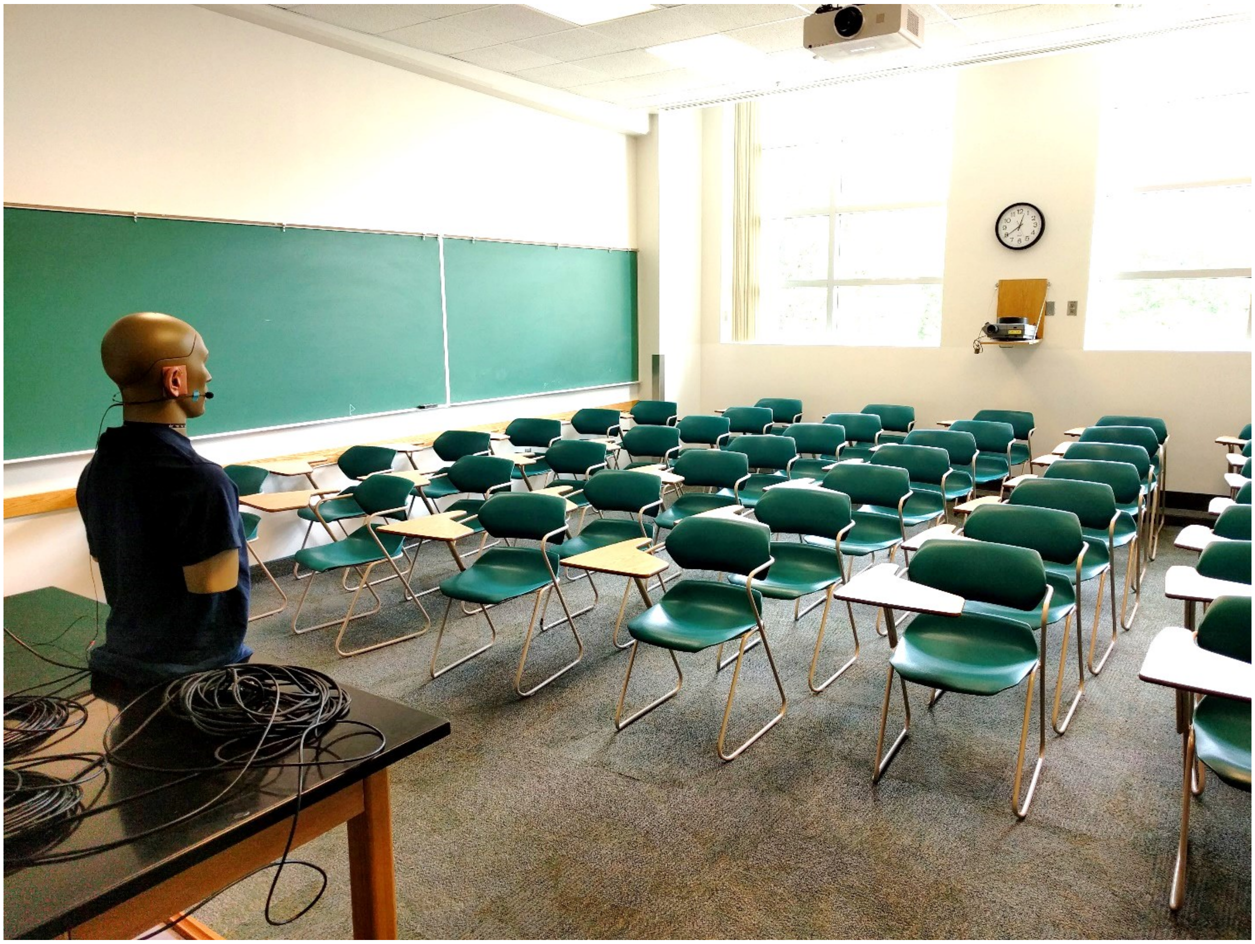


Figure 3.11. An OBRIR measurement in ESC C261. KEMAR was positioned at the front of the room, similar to where a teacher or lecturer might stand.

3.5.1.3 *de Jong Concert Hall*

The de Jong Concert Hall at BYU is a venue for musical and theatrical performances, university devotionals and forums, audio and video recordings, and other events [66]. It seats over 1200. The OBRIR measurements in this space were made with KEMAR in place of a performer, near the front of the stage. Figure 3.12 shows the hall from the perspective of the KEMAR mannequin, and Fig. 3.13 depicts the positioning of KEMAR on the stage.

3.5.2 Simulated OBRIRs

The simulated OBRIRs were created to provide contrast to the measured OBRIRs and demonstrate that both type of OBRIRs could be incorporated in the RTCS. By simulating some

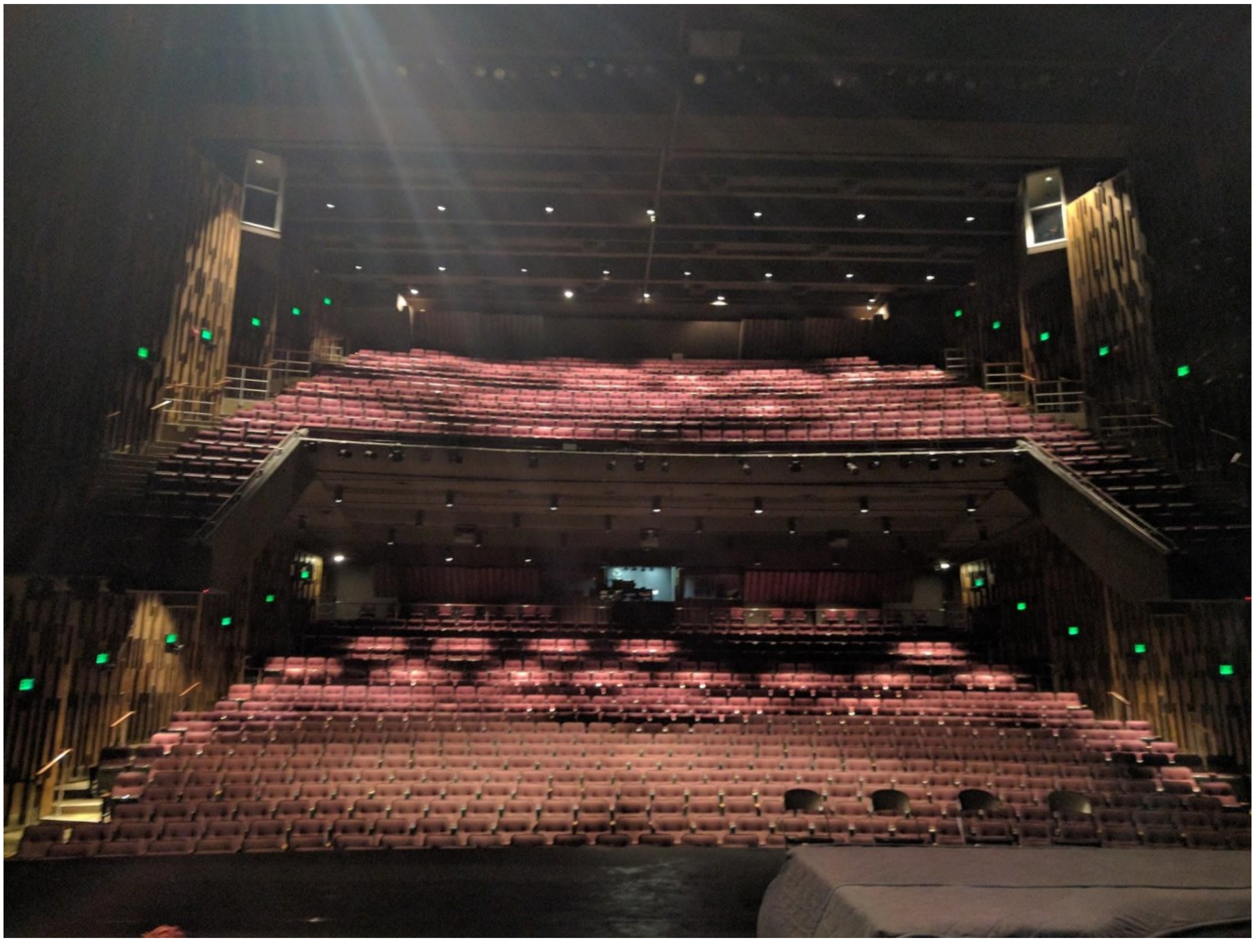


Figure 3.12. View of the de Jong Concert Hall from the front of the stage.

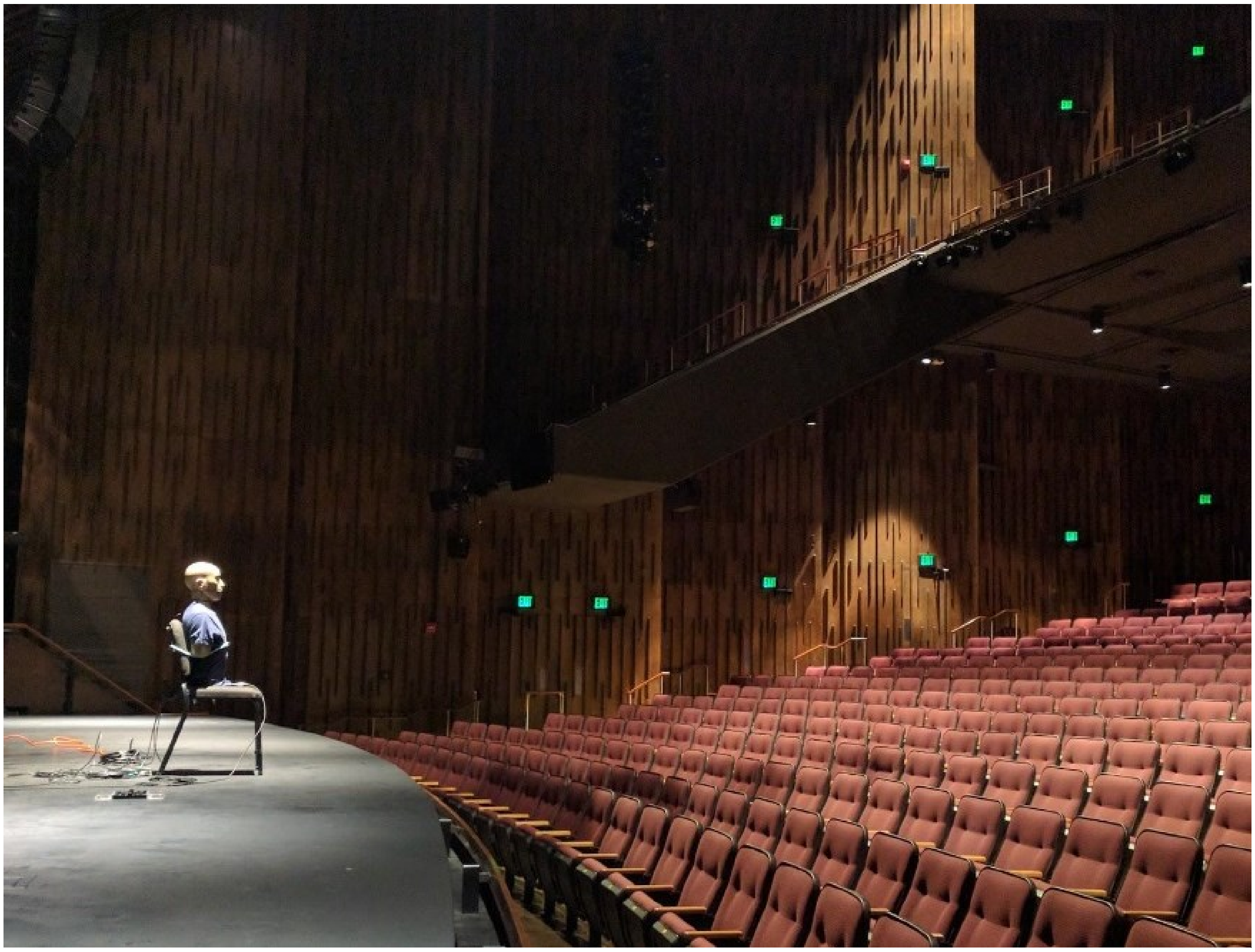


Figure 3.13. KEMAR mannequin positioned near the front of the stage in the de Jong Concert Hall for an OBRIR measurement.

of the rooms that also had OBRIR measurements, they could be directly compared and assessed for simulation accuracy.

3.5.2.1 *de Jong Concert Hall*

A geometric model of the de Jong Concert Hall was made in EASE, following CAD drawings for the Harris Fine Arts Center, in which it is housed. Appropriate acoustic absorption coefficients were applied to the surfaces of the room and a high-resolution simulation of an OBRIR was performed. Figures 3.14 and 3.15 show the different views of the geometric model of the de Jong Concert Hall. A table of the simulation parameters is included in Appendix B.

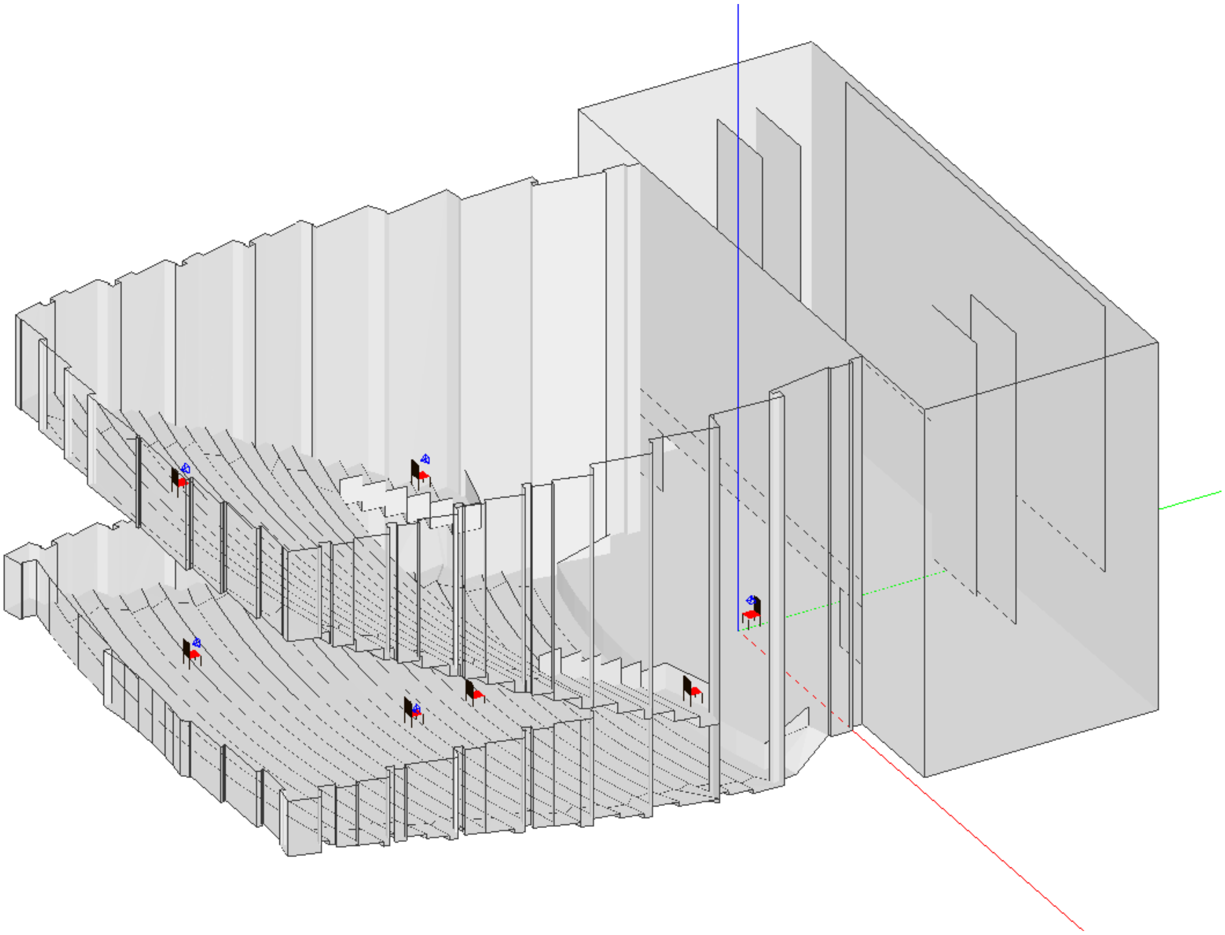
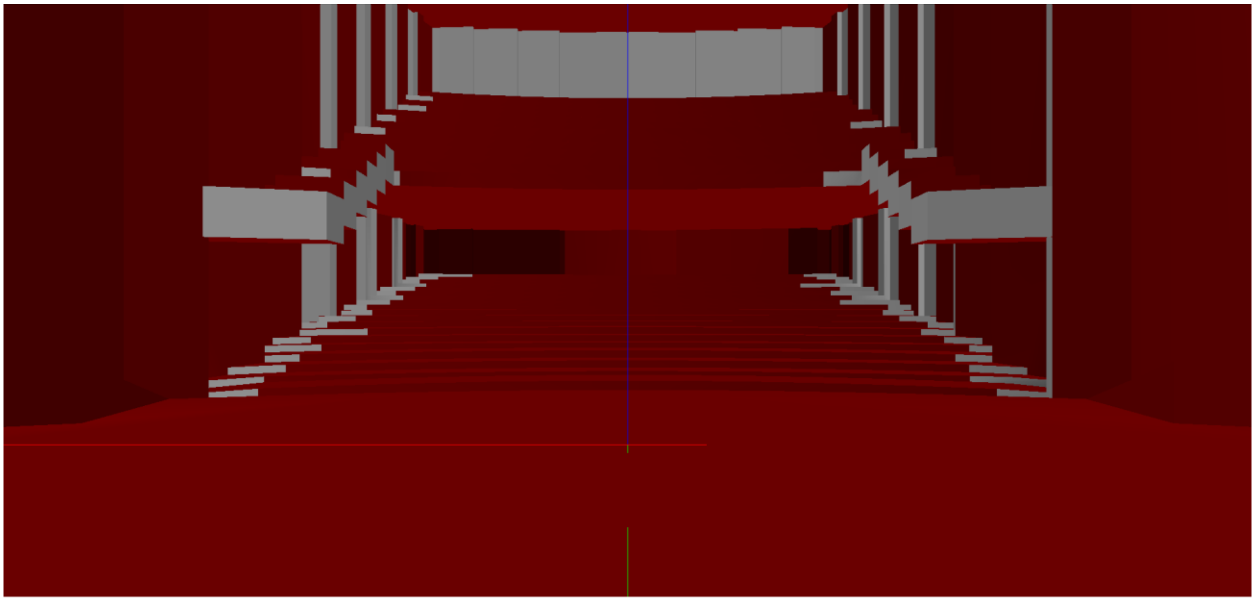


Figure 3.14. Geometric Model of de Jong Concert Hall in EASE.

3.5.2.2 *ESC C261*

An EASE model of classroom C261 in the ESC was constructed to closely replicate the features of the physical classroom. The speaker and receiver in this model were positioned closely, similar to the positioning of the KEMAR mannequin for the OBRIR measurement in the physical classroom. The geometric model is shown in Fig. 3.16.

A simulated OBRIR of C261 was also created with some modifications. So as not to affect the reverberation time in the room, the five earliest reflections were removed from the



(c) EASE 4.3 / De Jong Detailed Version from CAD floorplans / 1/16/2018 12:49:30 PM / BYU Jenny

Figure 3.15. Perspective view of the de Jong Concert Hall from the simulated stage speaker in EASE. This position is the same as that in the OBRIR measurement of the physical de Jong Concert Hall (see Fig. 3.12).

simulated OBRIR. This in effect simulated having the earliest reflections absorbed by the room boundaries. Figure 3.17 depicts the difference between the OBRIRs for the regular and absorptive cases.

3.6 OBRIR Characterization

Traditional measures computed from RIRs do not adequately characterize the space from the perspective of talkers and their OBRIRs, due mainly to the close spacing of the source and receivers. According to standards for traditional room measurements, the source and receivers are placed at least 1 m apart [67]. The close spacing of source and receivers in OBRIR measurements leads to short early decay times and reverberation time estimates from the early parts of the IR, where the decay is influenced strongly by the early diffracted sound around the HATS before room reflections arrive. For example, Fig. 3.18 shows a comparison of a traditional RIR measurement made in the reverberation chamber with eight absorbing wedges, and an

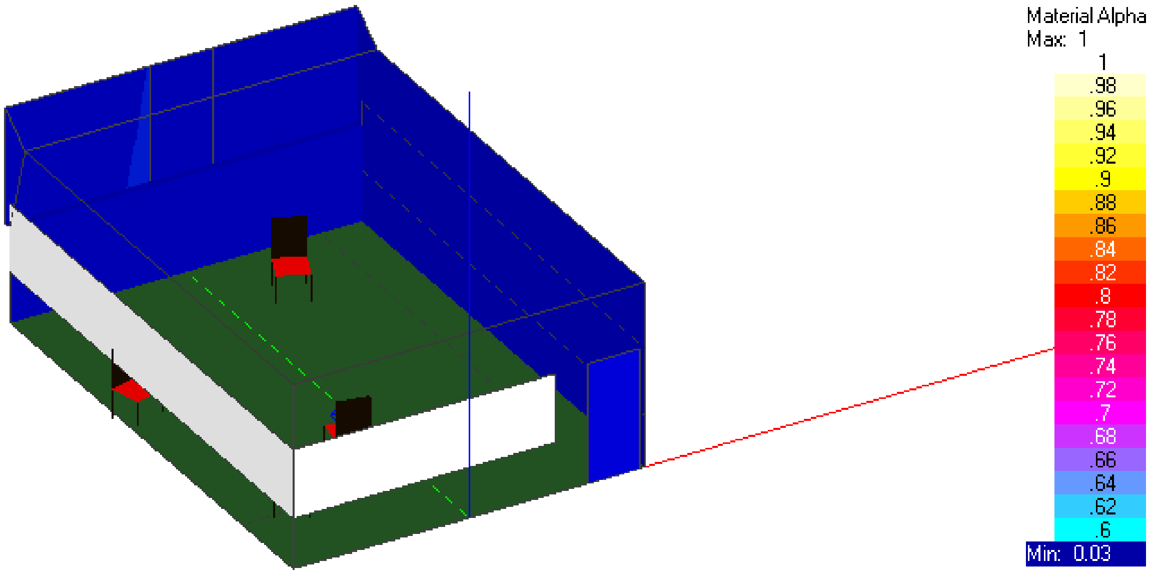


Figure 3.16. EASE Model of the classroom ESC C261. The colors on the room surfaces indicate varying absorption coefficients at 1 kHz. The chairs in the room are not physical parts of the model, but rather represent specific receiver locations. Only the receiver collocated with the speaker closest to the door was used in the OBRIR simulation.

OBRIR measurement in the same space. The Schroeder curves [68], from which estimates of reverberation time and other architectural acoustic parameters are dervied, were quite different for the two situations. In the traditional measurement, the spacing between the source (dodecahedron loudspeaker) and receiver (free-field microphone) was set at 1.83 m. This measurement characterized the room for the talker as a listener to the (omnidirectional) interviewer. The direct sound from the source, shown as the earliest arrival at the receiver, was prominent at the beginning of the IR, but did not appreciably change the slope of the Schroeder curve at that point. On the other hand, in the OBRIR measurement, the earliest diffracted arrivals were much stronger than the earliest room-reflected arrivals so that the Schroeder curve was significantly influenced. The slope of the Schroeder curve at the beginning of the IR was much different than the slope for the later portion. The estimates of T20, EDT, etc. were made by using the Schroeder curve and were thus influenced by its behavior of at the initial part of the IR.

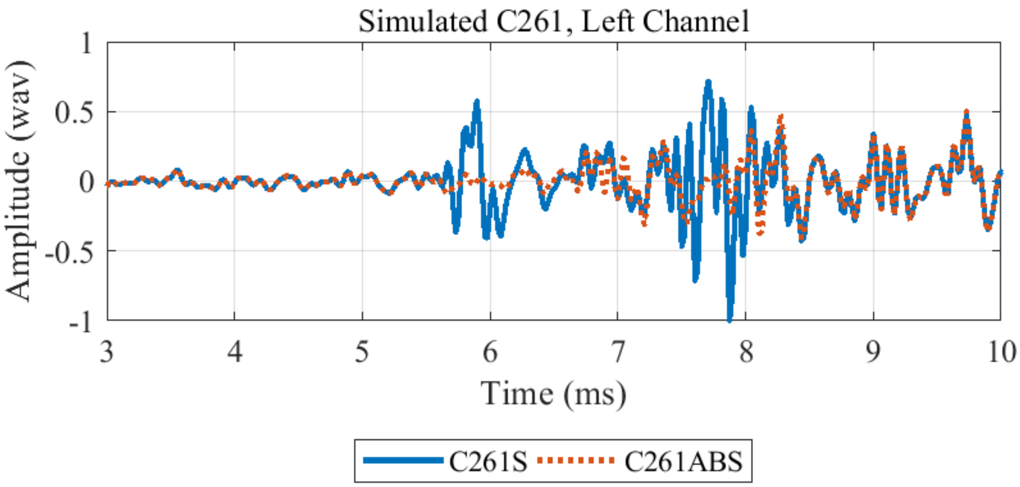


Figure 3.17. Simulated OBRIR of C261, Left Channel. In C261ABS, the five earliest reflections in the OBRIR were deleted, similar to the method used to delete the initial, direct sound part of the simulated OBRIR. The modified case is lower in amplitude during the first few milliseconds, but aligns with the original simulated OBRIR at around 8.5 ms and later.

Three parameters were chosen to better characterize the acoustic spaces through OBRIRs. The first was room gain, defined earlier by Brunskog et al. and used in the work of Pelegrin-Garcia [27,32]. This compares the energy in an OBRIR of a room to that of an OBRIR in an anechoic environment as follows:

$$RG = 10\log_{10}\left[\int_0^\infty h_{room}^2(t)dt \Big/ \int_0^\infty h_{anch}^2(t)dt\right] \qquad (3.2)$$

By definition, the room gain of an anechoic environment is then 0 dB.

The second was the 30 dB binaural decay time (BDT30), a novel extension of the traditional early decay time (EDT). The EDT was computed as the time for the Schroeder curve to decay from 0 dB to –10 dB and has been shown to correlate well with subjective perception of room reverberance. In an OBRIR, the early diffracted sound typically decays at least 10 dB before the earliest room reflections arrives at the ears. As a result, the BDT30 was instead computed by finding the time for the OBRIR Schroeder curve to decay from 0 dB to –30 dB.

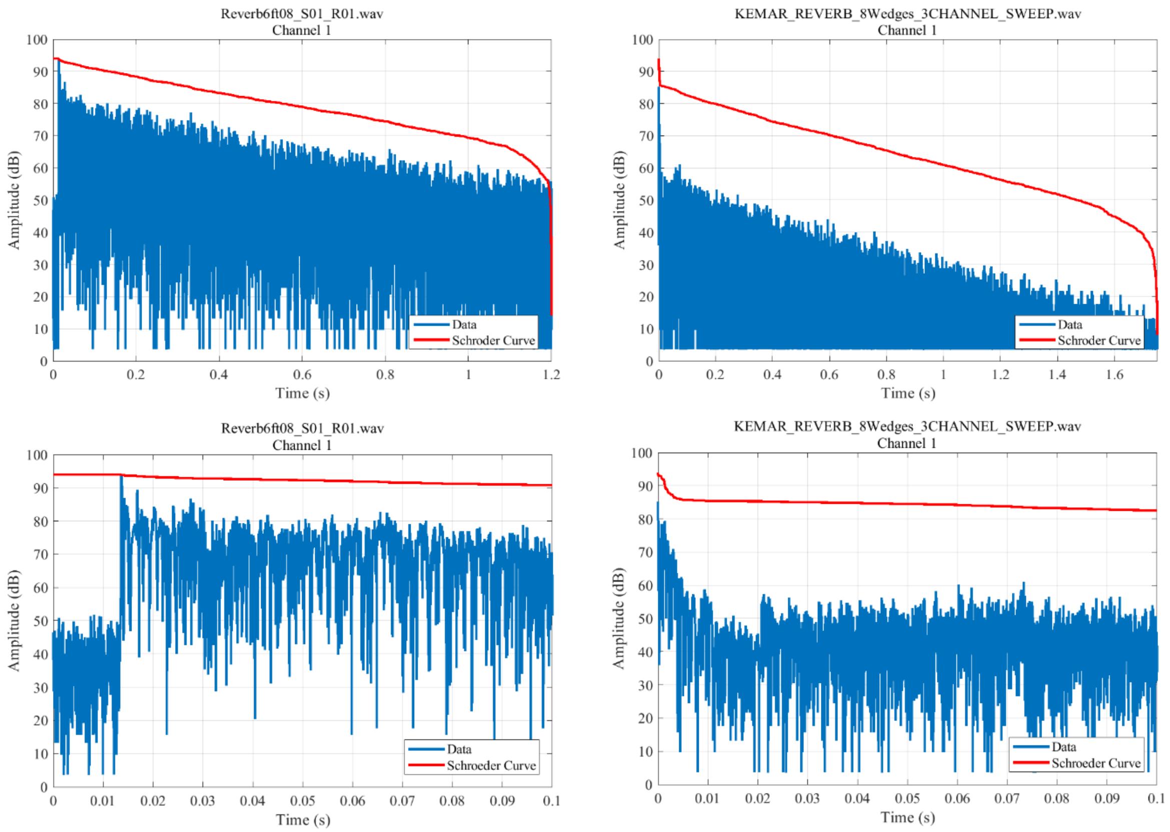


Figure 3.18. Traditional RIR measurement and OBRIR measurement for the reverberation chamber with eight absorbing wedges, left channel only. The top trace depicts the entire Schroeder integration curve while the bottom trace shows only the first 100 ms. The difference in Schroeder curves for the initial part of the OBRIR is clearly visible.

This ensured that both the diffracted sound and early reflections of the acoustic space are included in the measure.

A third parameter, diffracted-to-reflected decay-slope ratio (DRDSR) was also newly defined to compare the decay slopes for the diffracted and reflected portions of the OBRIR. They were demarcated by 7 ms, which is roughly the time required for sound to travel from the seated KEMAR mouth simulator, reflect off the floor (assumed to be the closest reflecting room surface), and arrive back at the ear microphones. The arrivals in the first 7 ms were then due to diffraction around the HATS and chair, whereas those after 7 ms contained room reflections, including the first floor reflection. Figure. 3.19 shows an example of the first 100 ms of an

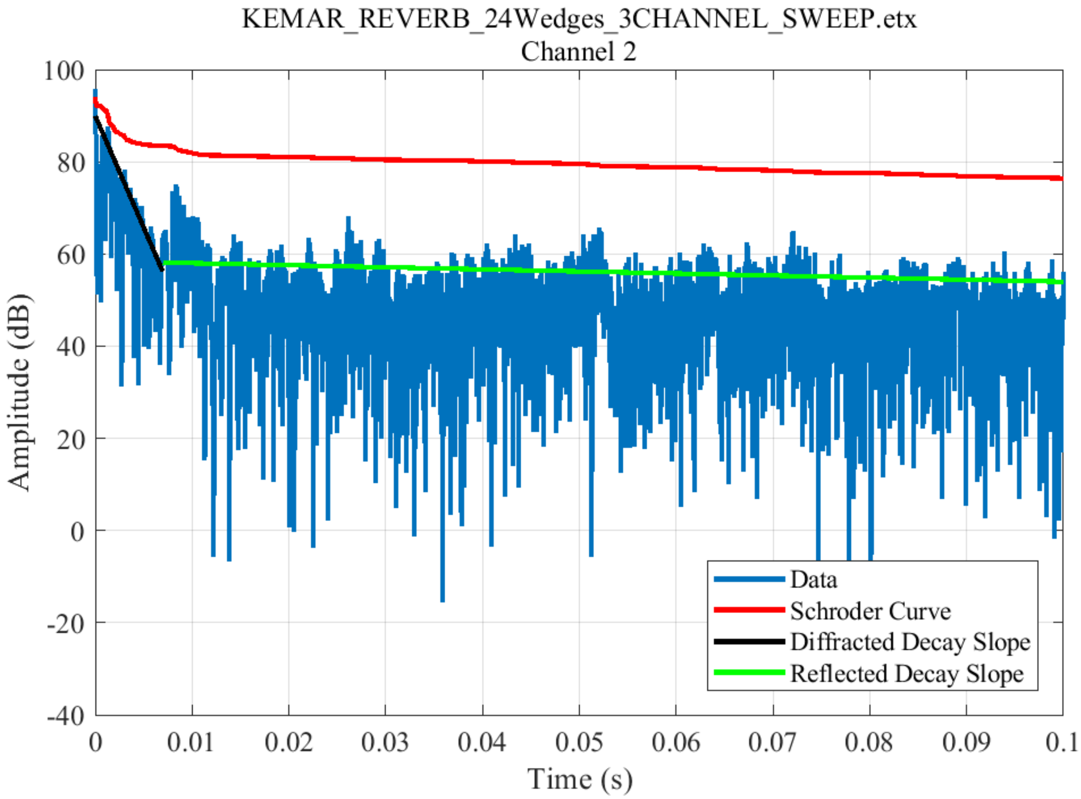


Figure 3.19. OBRIR measurement for the reverberation chamber with 24 absorbing wedges, right channel only.

OBRIR for the reverberation chamber with 24 absorbing wedges. It also includes overlays of a the Schroeder curve and diffracted and reflected sound slopes. The diffracted and reflected sound decay slopes were computed via modified Schroeder integration for each section. Instead of reverse integrating from the end of the IR to the main peak, the direct sound portion was reverse integrated from 7 ms to the main peak. The reflected sound portion was integrated from the end of the IR to 7 ms. A linear fit to these Schroeder curves resulted in the decay slopes for each portion of the OBRIR.

Figures 3.20 through 22 summarize the results of three parameters from the measured OBRIRs. They are plotted against traditional T20 measurements (computed from traditional RIR measurements made in the same spaces). The results for the OBRIRs of C261 are excluded because traditional T20 measurements were not available. In Fig. 3.20, a fitted least-means-

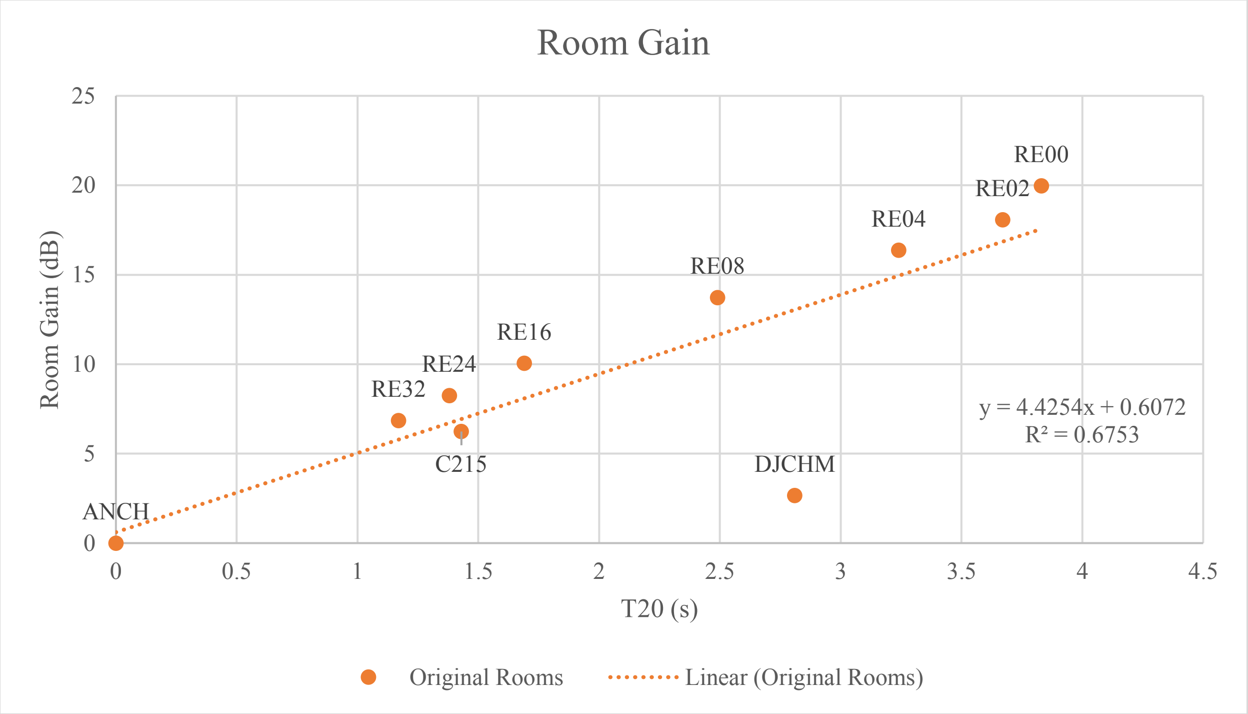


Figure 3.20. Room Gain compared to traditional T20 for acoustic spaces under consideration.

squares trend line shows a nearly linear relationship of several OBRIR-based room gains to the RIR T20s of the same spaces. An exception is the de Jong Concert Hall, which could have differed because it had a much larger volume than other rooms characterized by the OBRIRs. This would have affected the reverberation time and level, and the timing and levels of early reflections from more distant surfaces.

Figure 3.21 shows that the BDT30 values also have a nearly linear relationship with those of the traditional T20 for many rooms, but rooms with the longest T20 values seem to show an exponential trend. Another possibility is that these rooms are also outliers. The de Jong Concert Hall is again something of an outlier. The trend for DRDSR, shown in Fig. 3.22 follows that for BDT30. These new measures require additional psychoacoustical evaluation. Is BDT30 well correlated with a talker's perception of room size or speaking or listening difficulty? Is Room

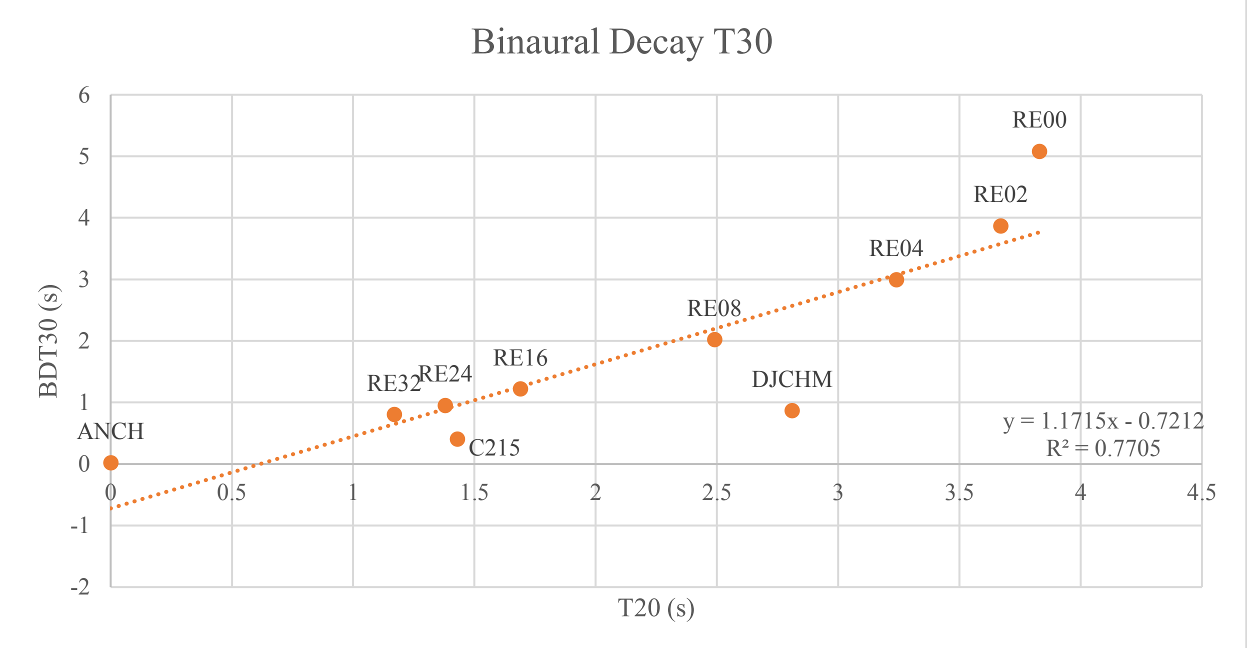


Figure 3.21. BDT30 compared to traditional T20 for acoustic spaces under consideration.

Gain an indicator of one’s perception of speaking support? These and other questions may be answered in future work.

3.7 Conclusion

A variety of OBRIRs were measured and simulated for use in the RTCS. These measurements and simulations represented a wide range of acoustic environments, as shown through traditional acoustic measurements and new acoustic parameters computed from OBRIRs. The simulated OBRIRs were created to provide contrast to the measured OBRIRs and demonstrate that both type of OBRIRs could be incorporated in the RTCS. The convolution of a dry signal with any of these OBRIRs could be used to produce an auralization with a RTCS that imitates what it would sound like if one was to actually speak in that environment. In the future,

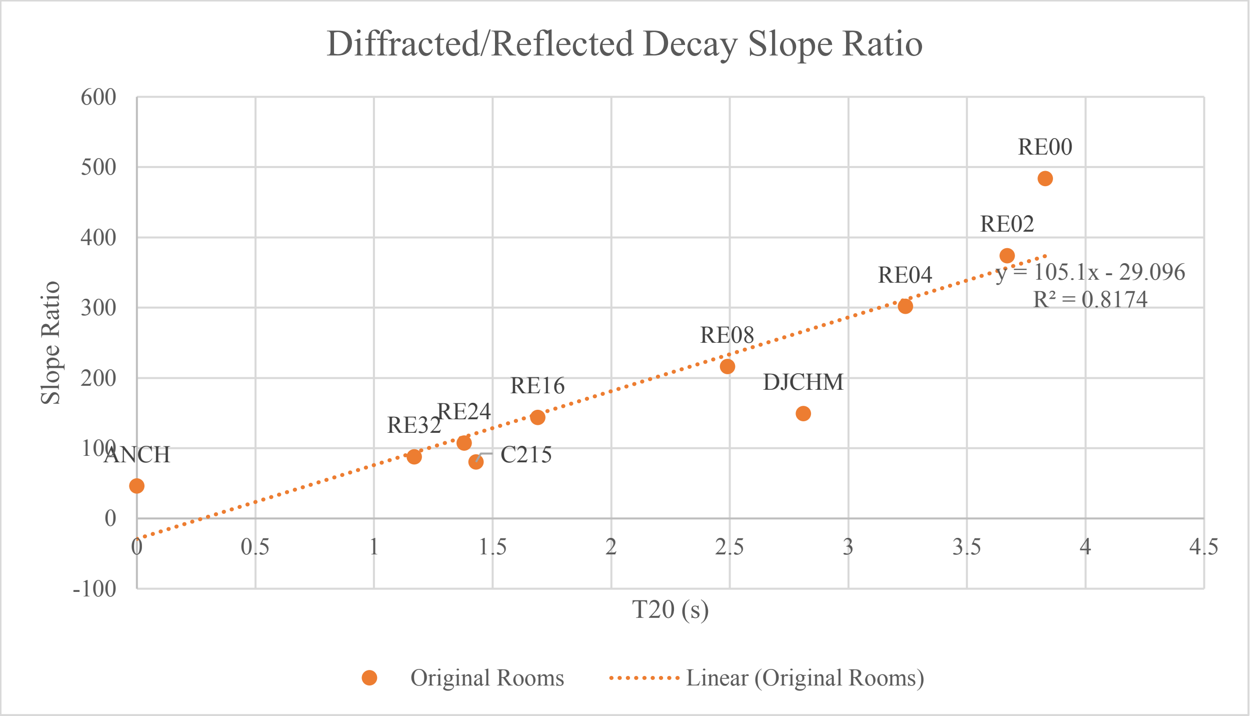


Figure 3.22. DRDSR compared to traditional T20 for acoustic spaces under consideration.

the inclusion of simulated OBRIRs that do not represent real or measured environments could further the extreme acoustic scenarios possible with the RTCS.

Chapter 4

Real-Time Convolution System Development

One of the unique aspects of this work is that it attempts to perform real-time convolution for the experience of a sound source closely located to the receivers. Most work in the realm of virtual acoustics has been done with the intent of simulating a sound source far from the receivers, as in the performance of an instrument on a stage while the listener is sitting at an audience position removed from the stage. This has been well studied and documented [16,69-73]. However, there is also value in studying what the performer experiences. The specific application of this work is the simulation of one's own voice in a virtual auditory environment. This concept could easily extend to the simulation of one's own instrument in an orchestral seating arrangement or the sound of one's own solo on the stage of a crowded concert hall, to provide just a few examples. This chapter describes the methods used to accomplish this task in previous studies and in the present research.

4.1 RTCS Background

The concept of a real-time convolution system (RTCS) is based fundamentally on the mathematical theory of convolution. A signal $x(t)$ is convolved with another signal $h(t)$, a filter or an IR, to produce an output signal $y(t)$:

$$y(t) = x(t) * h(t). \tag{4.1}$$

In the current study, $h(t)$ is a binaural IR using both left and right channels. The IR specifically pertains to a sound source (mouth) positioned close to the receivers (ears) of the subject—the OBRIR, as discussed in Ch. 3.

An auralization is produced by convolving a dry (anechoic) signal with a binaural IR to simulate how it would sound in the space the IR was measured in. Such IRs may be produced well in advance and treated like recordings in a space. Real-time auralizations then occur when an input signal $x(t)$ is sampled and passed through a digital computer to be convolved with an IR $h(t)$ and the output $y(t)$ is played back in nearly real-time.

The computational load and latency introduced by convolving even small signal buffer sizes with a long IR limits the realism of the real-time output. In practice, the convolutions are then performed indirectly via Fourier transform (Eq. 4.2). The transformed signals are multiplied in the frequency domain (Eq. 4.3a), and the result is subsequently inverse Fourier transformed back to the time domain (Eq. 4.3b).

$$X(f) = \mathrm{FFT}[x(t)], \tag{4.2a}$$

$$H(f) = \mathrm{FFT}[h(t)], \tag{4.2b}$$

$$Y(f) = X(f) \times H(f), \tag{4.3a}$$

$$y(t) = \mathrm{IFFT}[Y(f)]. \tag{4.3b}$$

This method speeds up computation and reduces the undesirable effects of convolution in the time domain. Computational advances in recent years have made real-time auralizations possible with ever decreasing latency and ever-increasing realism.

Several authors have used these principles in their research of real-time convolutions and auralizations. A summary of the work most closely related to the present study follows, along with several comments that place the latter in context.

4.1.1 The Cabrera and Yadav et al. RTCS system

The RTCS used in this work was based fundamentally on the real-time convolution system designed by Cabrera, et al [1]. They were also interested in the sound and perception of a talker's own voice in virtual acoustic environments and they implemented head tracking to more fully immerse the user into the environment. In developing their system, they faced a number of problems. First, they carried out the measurement of OBRIRs using a HATS or a live talker [20], discussed in Sec. 3.2. They also studied the variation in OBRIRs from rotations of the HATS within a room [74]. For the current work, we limited the OBRIR measurements to those of a KEMAR HATS at a single position within a room, with the assumption that the KEMAR anatomy is sufficient to provide a good OBRIR approximation for use with live talkers who each have their individual HRTFs. This is discussed further in Sec. 4.2.

The next problem they tackled was that of developing software to perform convolution in real-time [1]. They used RME ADI Quadspeed AD/DA converters and an HDSPe AES pci card with a PC and the Windows operating system. The convolution was performed in the commercially available SIR2 VST plugin housed within Max/MSP computation [75]. This present research used the same hardware and VST plugin but houses the VST plugin in a Reaper project file instead of within Max/MSP computation. The output of the convolution system was played through off-ear AKG K1000 headphones. The transducers are spaced away from the ear, allowing initial diffracted sound transmission from the talker's mouth to ears. However, the presence of the headphones does potentially affect the transmission via scattering. Yadav studied such effects and found that they did not significantly alter the FRF level across seven octave bands [76].

The Cabrera and Yadav system also implemented head tracking and selected an OBRIR for convolution, based on the position of the user’s head, from a number of previously measured OBRIRs. These corresponded to any of the possible head orientations within the closest 5-degree increment. The present work did not use head tracking, as it was more focused on the effect of a variety of rooms on a talker. However, a head-tracking component would increase realism as the sound of a room does change as one turns his or her head. Future work could include this addition.

The OBRIRs in the Cabrera and Yadav system were further altered through implementation of a headphone correction filter [1]. This filter was described as “the inversion of the transfer function from the headphones to the in-ear measurement microphones.” However, the only other details they offered were that “We used a 256-sample (sampling rate of 48 kHz) inverse filter (finite impulse response), which was combined with the OBRIR in the real-time convolver.” This filter added a 2.325 ms latency to their system. The initial part of the OBRIR was accordingly truncated by the total system latency so as to remove its direct-sound component and ensure that the simulated room reflections would arrive at the ears correctly delayed. The removal of the initial part of the OBRIR was also applied in the present work.

Their next step involved adjusting the gain of the simulation system such that the relationship between the direct and reflected sound that existed in the original OBRIR measurement was reproduced in the simulation with a HATS user. The OBRIR measurement process described above was carried out with the HATS using the RTCS with a reverberant-room OBRIR. A swept-sine signal was used to measure the IR between the mouth and ear microphones. The OBRIR obtained from the simulation system measurement was compared to the OBRIR measured in the reverberant room, and good agreement was found. Some deviation in the early part of the OBRIR was found due to the presence of the headphones, but the authors argued that the discrepancy would be masked by the direct sound. A formal listening test to compare the two OBRIRs had not been conducted, and a metric to quantify the differences between the two OBRIRS was not developed [1].

Since developing their system, Yadav and Cabrera have used it to carry out a study to investigate the perceived size of a room based on auditory stimuli [77]. It provided a key feature of the experiment: the removal of visual cues. Thus, participants in the study relied only on the "mixed-reality" experience of hearing their own voices and simulated room reflections. Because substantial initial portions of the OBRIRs were truncated by the system latency, near-field reflections were not included in the simulation-system output. To remedy this removal, a carpeted wooden floor was added to the anechoic chamber in which the participants used the simulation system. The head-tracking feature of the system was essential for the participants to fully explore the room acoustically by incorporating exploratory head movements.

In another study, these authors with their simulation system focused on talkers' voice-level regulation [78]. The talkers heard their own voices through the simulation system while addressing a mannequin seated five meters away. Using room gain as the metric of interest, the measured OBRIRs were modified to affect the reverberation time heard by the talkers. Vowel data was extracted from the speech of the talkers for statistical analysis.

Cabrera, Yadav, and their colleagues also used the system to study stage acoustics for singers [79]. The OBRIRs used in this study were not authentic measurements, but were computationally created to control early reflections, and then included a recorded reverberant tail. The authors detail the effect of computation latency on the system. With their system containing a latency of 7 ms, the earliest simulated reflections occurred 2.04 meters away from the user in the virtual acoustic environment.

This research group has published several papers on the variation in OBRIR measurements of the same room for varying horizontal rotations of the HATS [74,80]. They were most interested in metrics like room gain and interaural parameters, such as interaural level difference (ILD) and interaural cross correlation coefficient (IACC). In addition, they investigated variations between OBRIR measurements in a real room and those of a computer-modelled room, for varying horizontal rotations of the HATS [76]. They compared the measured and simulated values of EDT, C80, and IACC. For the zero-degree HATS position, all parameters were within a just noticeable difference (JND) between the measurements and simulations, but there was considerable variation for other rotations. These variations suggest that a person listening to his or her own voice would notice the differences amongst different acoustic spaces, whether modelled or simulated. This is promising news for simulation systems that rely on modelled room acoustics.

The development of the simulation system and OBRIR measurement techniques allowed Cabrera, Yadav, and their colleagues to begin to quantify how talker-listeners interact with their auditory environment, especially in the case of singers on stages, or talkers in environments intended for speaking or singing. For the present work, a similar simulation system allowed for the research of vocal effort in a wide variety of acoustic environments. The studies performed by Cabrera and Yadav thus served as a starting point for the design of experiments involving talkers and their responses to their own voices.

4.1.2 Work by Sato et al.

Sato was involved in several of the Yadav and Cabrera experiments on measuring OBRIRs using a HATS. While his report on listening, talking, and conversational speech difficulty was written in Japanese [81], Yadav provided a brief English summary [20]. In Sato's experiment, a microphone 0.1 m from the talker's mouth fed the speech signal to a convolver, which simulated room reflections and presented the result at the talker's ears via AKG K1000 headphones. The room reflections were simulated with parametric control of the simulated reverberation. The key result of this study showed that talking and conversing difficulty was more sensitive to room clarity (C50) than was listening difficulty. Yadav and Cabrera used similar methods to those of Sato, in terms of the simulation system setup, but they were interested in the acoustics of real rooms, not simulated reverberation.

4.1.3 Research of Porschmann et al.

Porschmann and colleagues focused on the psychoacoustic perception of one's own voice and identified three chief components: bone conduction, direct sound transmission around the head, and reflections from the acoustic environment [2]. In 1949, Bekesy estimated that the perceived loudnesses of bone-conducted sound and air-conducted sound were on the same order of magnitude [82]. Porschmann outlined models and measurements for each of the components and found that Bekesy's estimation was correct: both bone conduction and air conduction contribute at the same order of magnitude, but he found that bone conduction dominates a person's perception of his or her own voice at frequencies between 700 and 1200 Hz [2]. At higher frequencies, bone conduction is not as influential as the air-conducted sound.

Porschmann concluded that for the presentation of one's own voice in an auditory virtual environment, care must be taken to ensure that proper models are employed. For example, in simulating the reflections from a person’s voice, the directivity of the talker must be considered. If headphones are used to relay the virtual-acoustic signal, the insertion loss of the headphones must be determined and compensated for, especially if the headphones occlude the natural diffraction around the person’s head. With proper attention, such a system can deliver a virtual environment for talkers as well as listeners. Porschmann outlined the architecture of a headphone-based auditory virtual environment that included the presentation of one’s own voice in a 1998 German paper [83]. The present work was simpler in that it did not attempt to reproduce the initial diffracted sound, but only the reflections from the acoustic environment.

More recently, Porschmann was involved in the SCATIS system development. SCATIS is a multimodal virtual environment that can present one’s own voice in real time. Blauert et al. [84] described the system’s architecture and implementation, and Djelani et al. [85] performed a psychoacoustic evaluation of the system. The SCATIS uses a parametric reverberation algorithm that allows frequency-dependent control of the simulated reverberation time. It underwent several psychoacoustic tests to determine if the implemented model created a natural impression of one’s own voice in the virtual environment and used this information to determine if the presentation of one’s own voice increased the sense of presence in the auditory virtual environment [86]. A feedback filter was necessary to compensate for the headphone insertion loss but was susceptible to delays that affected the naturalness of the auditory presentation. In addition, the sense of presence was not found to dominate the sound of one’s own voice in the virtual auditory environment. Porschmann suggested that consideration be given to the influence of other contributing factors to the auditory virtual environment, and that perhaps the presentation of one’s own voice was not as important in some applications if a comparable effect could be achieved. In the present work, the issue of headphone occlusion was reduced significantly by the use of off-ear headphones. The reverberation of the environment was also that of a measured or simulated room instead of generic reverberation, and the presentation of one's own voice was made as natural as possible within in the virtual auditory environment.

4.1.4 Work of Pelegrin-Garcia et al.

The work of Pelegrin-Garcia et al. was most concerned with classroom acoustics from the point of view of the talker rather than the listener. His motivations were very similar to those of the present work, in attempting to develop a virtual auditory system that reproduced the sound of one's own voice in an effort to improve classroom acoustics. He worked with several room-acoustic simulation systems in his research, including loudspeaker and headphone-based systems.

Similar to the system developed by Yadav and Cabrera, only the air-conducted sound was reproduced in Pelegrin-Garcia's system [87]. Open headphones (Sennheiser HD570) were used, so the insertion loss due to the headphones was compensated for, as in the system developed by Porschmann. The OBRIRs were simulated using Catt Acoustic software with a 15° azimuthal angle resolution for orientation in the simulated room, selected using head tracking through Max MSP. The OBRIRs had their initial parts truncated due to the system latency. The system had no headphone equalization and no individualized HRTFs. It was used as part of an echolocation study, where the users of the system had no visual cues and had to navigate a space using only acoustic cues.

Another study performed by Pelegrin-Garcia et al. [32] investigated preferred acoustic conditions for speaking in classrooms and the usefulness of a parameter known as the room effect in understanding the interaction between talkers and their acoustic environments. The voice levels of talkers and the results from questionnaires were used in the investigation. The development of the system used in the study was described in Ref. [88].

Pelegrin-Garcia also investigated parameters linked to vocal effort and vocal comfort, as calculated from OBRIRs, viz.,- voice support and decay time [89]. The present work did not attempt to find such parameters to describe vocal effort directly from OBRIRs, but rather described vocal effort from parameters derived from speech signals, and then found correlation between vocal effort-related parameters and architectural-acoustics parameters that describe the conditions in which the subject was speaking. However, it is interesting to note that the OBRIRs themselves may have information that would be indicative of vocal effort in such an auditory environment.

4.2 RTCS Development and Implementation at BYU

The RTCS developed for the present study was meant to simulate the auditory experience of being in a room while a talker-listener was physically in the free-field environment of an anechoic chamber. The following sections first describe the audio hardware used to process speech signals from the live talkers, then the manipulation of OBRIRs preparatory to their inclusion in the RTCS. The derivation of a theoretical equalization filter follows, and subsequently details on the computation of that equalization filter.

4.2.1 RTCS Hardware

The hardware used in the RTCS included RME ADI Quadspeed AD/DA converters and HDSPe AES pci card, combined with a PC and the Windows operating system. The convolution was performed in the commercially available SIR2 VST plugin [75]. This research thus used the same hardware and VST plugin as was used in the system developed by Cabrera and Yadav [1], but housed the VST plugin in a Reaper project file instead of the Max/MSP computation environment. As in their system, the output of the convolution system was played through off-ear AKG K1000 headphones to allow initial diffracted sound transmission.

The signal flow for convolving one's own speech through the convolution system is shown in Fig. 4.1. In the block diagram, $\hat{a}'(f)$ is the Fourier-transformed signal produced by the talker at a hypothetical point in space near his or her mouth in the presence of the RTCS microphone and headphones. The FRFs $D_L'(f)$ and $D_R'(f)$ represent the propagation paths of the signal from this point around the head and the RTCS microphone and headphones to the left (L) and right (R) ears, respectively. [The primed variables indicate modifications to signals and FRFs caused by the presence (e.g., scattering) of the RTCS hardware.] In addition, $M(f)$ is the FRF of the propagation path from the hypothetical point near the mouth to the head-worn microphone at the corner of the talker's mouth, and the signal $\hat{m}_s(f)$ is the signal acquired by the head-worn microphone. The RTCS processing FRFs include $C_1(f)$, representing its input hardware components (RME QuadMic II preamplifier, RME ADI-8 QS A/D converter, and HDSPe AES pci card) and $C_{2,L}(f)$ and $C_{2,R}(f)$, representing two channels of output hardware components (RME ADI-8 QS D/A converter and Crown D-75 amplifier). The FRFs $R(f)$,

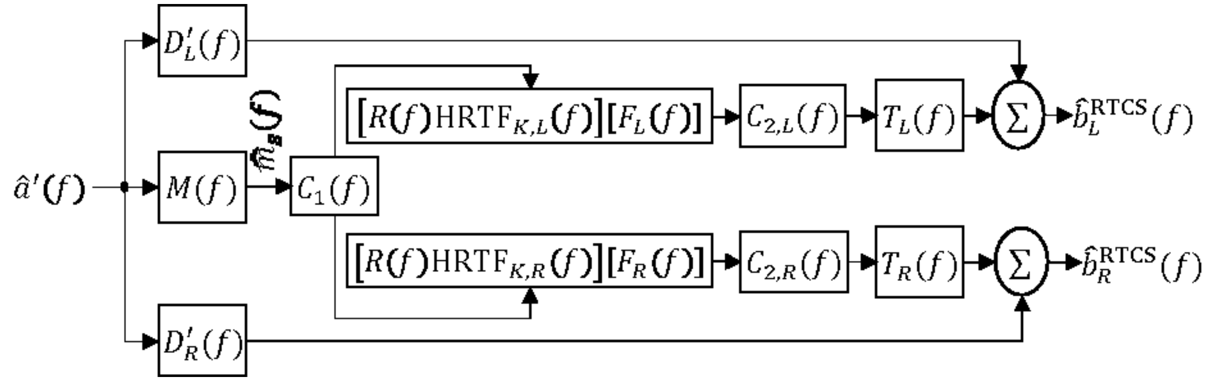

$D_L'(f)$
$[R(f)\text{HRTF}_{K,L}(f)][F_L(f)]$
$C_{2,L}(f)$
$T_L(f)$
Σ
$\hat{b}_L^{\text{RTCS}}(f)$
$\hat{m}_s(f)$
$\hat{a}'(f)$
$M(f)$
$C_1(f)$
$[R(f)\text{HRTF}_{K,R}(f)][F_R(f)]$
$C_{2,R}(f)$
$T_R(f)$
Σ
$\hat{b}_R^{\text{RTCS}}(f)$
$D_R'(f)$

Figure 4.1. Block diagram of the signal path of a talker with the RTCS. The input signal is modified by diffraction around the head and headphones, and the signal processing of the RTCS.

$\mathrm{HRTF}_{K,L}(f)$, and $\mathrm{HRTF}_{K,R}(f)$ within the square brackets correspond to OBRIRs, after direct-sound arrivals have been removed, that are loaded into the SIR2 plugin within a track in the RTCS Reaper project file. The subscript K indicates that the HRTF pertains to a KEMAR mannequin. Modification of OBRIRs to remove the direct sound arrivals are further discussed in Sec. 4.2.2. The additional FRFs $F_L(f)$ and $F_R(f)$ in square brackets represent equalization filters designed to flatten the frequency response of the RTCS. Development of these filters is detailed in Secs. 4.2.3.1.3 and 4.2.3.2.2.

The remaining blocks in Fig. 4.1 correspond to physical FRFs. Finally, $T_L(f)$ and $T_R(f)$ are the FRFs of the left and right AKG K1000 headphone transducers respectively. They include the propagation paths from the headphones to the entrances of the blocked ear canals. The output signals at the left and right ear canal entrances are $\hat{b}_L^{\mathrm{RTCS}}(f)$ and $\hat{b}_R^{\mathrm{RTCS}}(f)$, respectively. Mathematically, the relationships of the output signals to the input signal are given by the algebraic expressions

$$\begin{aligned}\hat{b}_L^{\mathrm{RTCS}}(f) &= \hat{a}'(f)D_L'(f) \\ &\quad + \hat{a}'(f)M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)C_{2,L}(f)T_L(f) \\ &= \hat{a}'(f)\left[D_L'(f) + M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)C_{2,L}(f)T_L(f)\right],\end{aligned} \tag{4.4a}$$

$$\begin{aligned}\hat{b}_R^{\mathrm{RTCS}}(f) &= \hat{a}'(f)D_R'(f) \\ &\quad + \hat{a}'(f)M(f)C_1(f)R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)C_{2,R}(f)T_R(f) \\ &= \hat{a}'(f)\left[D_R'(f) + M(f)C_1(f)R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)C_{2,R}(f)T_R(f)\right].\end{aligned} \tag{4.4b}$$

The composite FRFs for a talker using the RTCS are then

$$\begin{aligned}H_L^{\mathrm{RTCS}}(f) &= \frac{\hat{b}_L^{\mathrm{RTCS}}(f)}{\hat{a}'(f)} \\ &= D_L'(f) + M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)C_{2,L}(f)T_L(f),\end{aligned} \tag{4.5a}$$

$$\begin{aligned} H_R^{\mathrm{RTCS}}(f) &= \frac{\hat{b}_R^{\mathrm{RTCS}}(f)}{\hat{a}'(f)} \\ &= D_R'(f) + M(f)C_1(f)R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)C_{2,R}(f)T_R(f). \end{aligned} \tag{4.5b}$$

The signal produced by the talker is affected by each of the FRFs it passes through before it is presented at the ears. The composite FRFs summarize all of the alterations to the signal as it passes through the RTCS, and relates the output signal to the input signal.

4.2.2 OBRIR Manipulation

The FRFs $R(f)$, $\mathrm{HRTF}_{K,L}(f)$, and $\mathrm{HRTF}_{K,R}(f)$ within the square brackets correspond to OBRIRs housed within the RTCS. As mentioned earlier, the OBRIR measurements were modified before they were implemented in the RTCS. One major modification involved the removal of the initial diffracted sound arrival from the OBRIR, because this was inherently present (with only minor modifications) when the talker spoke with the RTCS hardware on his or her head. There was therefore no need to reproduce this arrival through RTCS. Because the OBRIR was measured with the KEMAR mannequin, the response of its mouth simulator also needed to be removed because it was not flat over the frequency range of interest (80 Hz to 10,000 Hz). The processing algorithm followed the method of Cabrera and Yadav [20], with a MATLAB routine performing the computations (see Appendix C).

To remove the mouth simulator effects, the IR of the signal acquired by the head-worn microphone relative to the signal produced by KEMAR was used. However, instead of using the KEMAR room OBRIR, a KEMAR anechoic OBRIR was used. This had the advantage of including no room reflections. The anechoic IR therefore involved only the sound directly from the mouth simulator and the near-field scattering from the KEMAR head and torso, which was similar to that experienced by subjects using the RTCS in the anechoic environment.

The IR from KEMAR's mouth simulator input to the head-worn microphone output, $\mathrm{IFFT}[\frac{\hat{m}_s(f)}{\hat{a}_s(f)}]$, was zero padded to be twice its original length. The IRs to the ear microphones, $\mathrm{IFFT}[H_{K,L}^{\mathrm{room}}(f)]$ and $\mathrm{IFFT}[H_{K,R}^{\mathrm{room}}(f)]$, were also zero padded over their second halves. The Fourier transforms of each of these modified IRs were then performed to bring the computation to the frequency domain. The RTCS FRFs were subsequently computed by dividing the cross-spectra of the mouth and ear microphone IRs by the auto-spectrum of the mouth IR. The equivalent expressions, according to the block diagram of Fig. 3.3 are

$$\frac{\hat{b}_{s,K,L}^{\mathrm{room}}(f)}{\hat{m}_s(f)} = \left[D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f)\right]\frac{B_{K,L}(f)}{M(f)} \tag{4.6a}$$

$$\frac{\hat{b}_{s,K,R}^{room}(f)}{\hat{m}_s(f)} = \left[D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f)\right]\frac{B_{K,R}(f)}{M(f)} \tag{4.6b}$$

Additional signal conditioning was needed to remove potential errors in the resultant FRFs at the extremes of the audible range. While frequency content outside the range 60 Hz to 10.5 kHz is not as important for speech production, it is still audible, as it falls in the range of 20 Hz to 20 kHz. Accordingly, these FRFs, $\frac{\hat{b}_{s,K,L}^{\mathrm{room}}(f)}{\hat{m}_s(f)}$ and $\frac{\hat{b}_{s,K,R}^{\mathrm{room}}(f)}{\hat{m}_s(f)}$ were bandpass filtered from 60 Hz to 10.5 kHz. This range may make fricatives sound fuzzier, as they do include high-frequency energy. However, hardware limitations inhibited extending the range of the bandpass filter. Future work may extend the upper limit. Assuming the ear microphones and head-worn microphone responses were flat in this frequency range, they could be dropped from the expression in Eq. 4.6. The bandpass filtered FRFs are given by

$$\left[\frac{\hat{b}_{s,K,L}^{\mathrm{room}}(f)}{\hat{m}_s(f)}\right]_{\mathrm{filt}} = D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f) \tag{4.7a}$$

and

$$\left[\frac{\hat{b}_{s,K,R}^{\mathrm{room}}(f)}{\hat{m}_s(f)}\right]_{\mathrm{filt}} = D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f). \tag{4.7b}$$

The results of Eq. 4.7 were then inverse Fourier transformed and truncated to discard their latter halves to yield time-domain IRs.

After the frequency-domain modifications were performed, an additional modification in the time-domain was needed. To minimize effects of the measurement noise floor that might be audible in the OBRIR tails, the OBRIRs were manually fitted with an exponentially decaying time window. This had the effect of removing extraneous and unhelpful parts of the IR from the RTCS convolution procedure, resulting in more realistic simulations.

Finally, the initial parts of the OBRIRs were truncated by the 6 ms RTCS latency. This involved use of a front-half Tukey window with a ramp-up time of 1 ms. It ensured that there was no convolution with the direct-sound portion of the IR, and that the simulated room reflections arrived at the user ears after being appropriately delayed. As a result, the system FRFs simplified to

$$\left[\frac{\hat{b}_{s,K,L}^{\mathrm{room}}(f)}{\hat{m}_s(f)}\right]_{\mathrm{filt,\,trunc}} = R(f)\mathrm{HRTF}_{K,L}(f) \tag{4.8a}$$

$$\left[\frac{\hat{b}_{s,K,R}^{\mathrm{room}}(f)}{\hat{m}_s(f)}\right]_{\mathrm{filt,\,trunc}} = R(f)\mathrm{HRTF}_{K,R}(f) \tag{4.8b}$$

As a final step, the truncated and filtered OBRIRs from Eq. (4.8) were bandpass filtered again from 60 Hz to 10.5 kHz to remove artifacts at the extremes of the frequency spectrum introduced by the various modifications. The OBRIRs were then normalized to avoid clipping and to maximize signal levels. They were subsequently saved as *.wav files and loaded into the SIR2 plugin.

4.2.3 Theoretical Equalization Filter Derivation

The RTCS is meant to create an auditory experience for talkers in a simulated room while they are physically located in the free field of an anechoic chamber. The RTCS is considered validated if it accurately portrays the acoustic effects of a room such that the IR measured through the RTCS appropriately matches that of the IR measured in a room. To do this, an equalization filter must be implemented to compensate for the components of the RTCS that detrimentally affect the auditory experience.

The following sections detail the theory and application of developing an equalization filter for the RTCS, given by $F_L(f)$ and $F_R(f)$. First, some theory is introduced describing the signal paths of sound emitted from a talker's mouth to the talker's ears. The signal paths for equivalent situations with a KEMAR mannequin instead of a live talker are then given. These show that with the proper filtering, the RTCS can theoretically simulate identical signal paths, thus representing a room. Secondly, IRs measured using KEMAR in rooms and in RTCS simulations of those rooms are compared. A metric quantifying the differences between the measurements is derived and used in Chapter 5.

4.2.3.1 *Live Talker Signal Paths*

When a talker produces speech in either an actual room or while using the RTCS, he or she hears it with modifications corresponding to the acoustic environment. These modifications depend on certain FRFs unique to components of the room or the RTCS. This section describes the acoustic and electric signal paths for the two scenarios.

4.2.3.1.1 Speaking in a Room

When speaking in an actual room, a talker hears his or her own speech modified by (1) initial head diffraction, (2) the RIR, and (3) the HRIR. Neglecting bone conduction and other internal responses, such as those of the vocal tract and ear canals unique to each talker [2,15], the convolution of these responses form the OBRIR. This differs from a BRIR because it includes the sound source and head diffraction of one's own voice rather than an external sound source.

To clarify, definitions of each response are repeated here. Initial head diffraction is here defined as the propagation of sound from an arbitrary point directly in front of the mouth, around the head, and to the entrance of a given ear canal. The RIR characterizes the various reflected paths of sound from this point, about the room, and to a position coinciding with the effective center of the head, without the talker being present. For a given room, it varies according to the positions of these two points, speech directivity, the geometry of the room, and the absorption and scattering of the room surfaces. As mentioned earlier, the HRIR is the inverse Fourier transform of the HRTF. Both describe the filtering of sound from any angle to the entrance of a left or right ear canal with the talker present [90]. The HRTF is sometimes defined as the sound pressure measured at the ear canal entrance divided by that measured by an omnidirectional microphone at a position coinciding with the center of the head, with the head absent [91]. As such, the HRTF depends upon the unique anatomy of each talker.

Figure 4.2 represents a rough depiction of what happens when a person speaks in a room. The talker produces speech at a point near his or her mouth. This speech arrives at the talker's ears via several paths, the shortest of which is the initial sound that has been diffracted around the head. In addition to this diffracted sound, some sound arrives later, after being reflected off the room surfaces and modified by the talker's HRTF.

In more detail, the talker produces a complex frequency-dependent signal $\hat{a}(f)$ (the Fourier transform of the time signal) at a hypothetical point near his or her mouth. The FRF $D_L(f)$ represents the diffraction of the signal from this point, around the head, and to the left ear, while $D_R(f)$ represents similar diffraction to the right ear. The FRF $R(f)$ is the Fourier transform of the RIR to the unobstructed central head position, while $\mathrm{HRTF}_L(f)$ is the talker HRTF for the left ear and $\mathrm{HRTF}_R(f)$ is that for the right ear. Finally, the signal $\hat{b}_L^{\mathrm{room}}(f)$ is the signal at the entrance to the blocked left ear canal and $\hat{b}_R^{\mathrm{room}}(f)$ is that at the entrance to the blocked right ear canal. Blocked ear canals are considered here to avoid additional unique resonances of each individual's ear canals. The signal at the blocked ear canal is analogous to a Thevenin-equivalent pressure driving the ear canal.

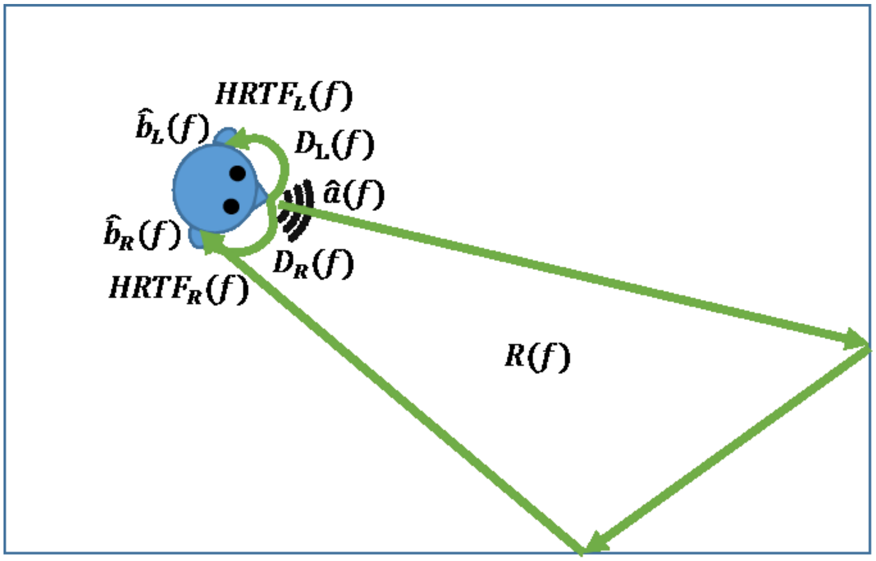


Figure 4.2. Depiction of a talker speaking in a room. After sound emits from the mouth, a portion diffracts around the head and arrives at the ears first. Some travels across the room, reflects off surfaces one or more times, and arrives at the talker's ears later. Both processes contribute to the signal the talker hears. A more complete representation would include many rays emitted from the mouth, with the relative strength of each ray being determined by the directivity of the talker and the absorption and scattering properties of the room surfaces.

The overall process may be characterized in terms of a single-input, dual-output system, as depicted in the block diagram of Fig.4.3. Mathematically, the relationships of the output signals to the input signal are

$$\begin{aligned}\hat{b}_L^{\mathrm{room}}(f) &= \hat{a}(f)D_L(f) + \hat{a}(f)R(f)\mathrm{HRTF}_L(f) \\ &= \hat{a}(f)[D_L(f) + R(f)\mathrm{HRTF}_L(f)],\end{aligned} \tag{4.9a}$$

$$\begin{aligned}\hat{b}_R^{\mathrm{room}}(f) &= \hat{a}(f)D_R(f) + \hat{a}(f)R(f)\mathrm{HRTF}_R(f) \\ &= \hat{a}(f)[D_R(f) + R(f)\mathrm{HRTF}_R(f)].\end{aligned} \tag{4.9b}$$

The composite FRFs for a talker speaking in a room are then

$$H_L^{\mathrm{room}}(f) = \frac{\hat{b}_L^{\mathrm{room}}(f)}{\hat{a}(f)} = D_L(f) + R(f)\mathrm{HRTF}_L(f), \tag{4.10a}$$

$$H_R^{\mathrm{room}}(f) = \frac{\hat{b}_R^{\mathrm{room}}(f)}{\hat{a}(f)} = D_R(f) + R(f)\mathrm{HRTF}_R(f). \tag{4.10b}$$

In the time-domain, these FRFs coincide with the OBRIR mentioned previously. (In the frequency domain, they might then be described as the OBRFRF.) An OBRIR is thus the relationship between the signal at an arbitrary point in space near the mouth and those at the entrances to the ear canals. One OBRIR is distinct from others because of the uniqueness of the talker's head geometry, which affects both the initial diffracted portion of the sound and the

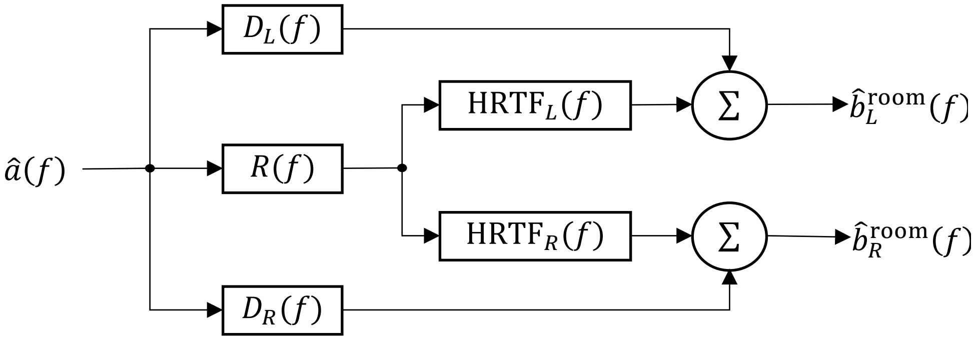


Figure 4.3. Block diagram for a talker in a room.

HRTF. It is also distinct because of the uniqueness of the RIR, which is affected by not only the room geometry and materials, but also the locations of the mouth and ears within the room.

4.2.3.1.2 Speaking with the RTCS

When using the RTCS, a different process occurs for the talker. He or she again hears initial diffracted speech first, but it is modified by the RTCS hardware. A speech signal detected by a head-worn microphone is also processed by electronic equipment to produce a real-time convolution of the speech signal with an OBRIR incorporating measured or simulated room reflections and HRTFs that are then played through the AKG K1000 headphones, which are offset from the ears, as shown in Fig. 4.4 with a RTCS user. This essentially replaces the room reflections of an actual room with the signal played through the headphones. Section 4.2.1 described the signal flow of the speech signal through the RTCS and arriving at the talker's ears.

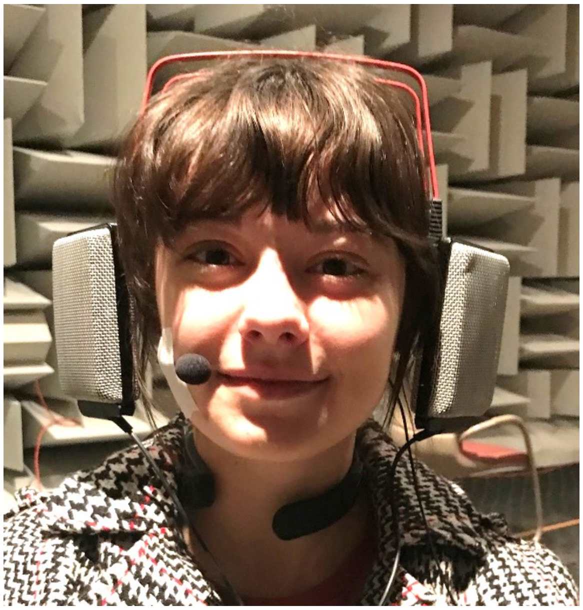


Figure 4.4. A female participant wearing AKG K1000 headphones and a head-worn DPA 4060 microphone. The headphones are offset from the ear, allowing for less disruption of the head diffracted sound. The transducers are angle-adjustable, but may be locked into position

4.2.3.1.3 Filter Derivation

The two scenarios: in a room, depicted in the block diagram of Fig. 3.3, and in the RTCS, depicted in the block diagram of 4.1, have many differences. Since so many components in the RTCS contribute to the fact that the output signals $\hat{b}_L^{\mathrm{RTCS}}(f)$ and $\hat{b}_L^{\mathrm{room}}(f)$, and $\hat{b}_R^{\mathrm{RTCS}}(f)$ and $\hat{b}_R^{\mathrm{room}}(f)$ differ, an equalization filter must be included to eliminate these unwanted effects. One method is to derive the filters $F_L(f)$ and $F_R(f)$ from the composite FRFs of the two scenarios. Since we want $H_L^{\mathrm{room}}(f) = H_L^{\mathrm{RTCS}}(f)$ and $H_R^{\mathrm{room}}(f) = H_R^{\mathrm{RTCS}}(f)$, we can solve for the filters $F_L(f)$ and $F_R(f)$ that make this possible. Equating Eqs. (4.10) and (4.5) yields

$$D_L(f) + R(f)\mathrm{HRTF}_L(f) = D_L'(f) + M(f)C_1(f)[R(f)\mathrm{HRTF_L}(f)]C_{2,L}(f)F_L(f)T_L(f), \tag{4.11a}$$

$$D_R(f) + R(f)\mathrm{HRTF}_R(f) = D_R'(f) + M(f)C_1(f)[R(f)\mathrm{HRTF}_R(f)]C_{2,R}(f)F_R(f)T_R(f), \tag{4.11b}$$

or

$$F_L(f) = \frac{D_L(f) + R(f)\mathrm{HRTF}_L(f) - D_L'(f)}{M(f)C_1(f)[R(f)\mathrm{HRTF_L}(f)]C_{2,L}(f)T_L(f)}, \tag{4.12a}$$

$$F_R(f) = \frac{D_R(f) + R(f)\mathrm{HRTF}_R(f) - D_R'(f)}{M(f)C_1(f)[R(f)\mathrm{HRTF}_R(f)]C_{2,R}(f)T_R(f)}. \tag{4.12b}$$

These filters could be included in the RTCS processing to filter out its non-ideal characteristics. In principle, the talker would then experience the effects of talking in a desired room while actually using the RTCS in the anechoic chamber. Inserting the filters into the RTCS signal flow yields the following results:

$$\begin{aligned}
\hat{b}_L^{\mathrm{RTCS}}(f) &= \hat{a}'(f)H_L^{\mathrm{RTCS}}(f) \\
&= \hat{a}'(f)D_L'(f) + \hat{a}'(f)M(f)C_1(f)[R(f)\mathrm{HRTF}_L(f)]C_{2,L}(f)F_L(f)T_L(f) \\
&= \hat{a}'(f)D_L'(f) \\
&\quad + \hat{a}'(f)M(f)C_1(f)[R(f)\mathrm{HRTF}_L(f)]C_{2,L}(f)\left[\frac{D_L(f) + R(f)\mathrm{HRTF}_L(f) - D_L'(f)}{M(f)C_1(f)[R(f)\mathrm{HRTF}_L(f)]C_{2,L}(f)T_L(f)}\right]T_L(f) \\
&= \hat{a}'(f)D_L'(f) + \hat{a}'(f)[D_L(f) + R(f)\mathrm{HRTF}_L(f) - D_L'(f)] \\
&= \hat{a}'(f)[D_L(f) + R(f)\mathrm{HRTF}_L(f)],
\end{aligned} \tag{4.13a}$$

$$\begin{aligned}
\hat{b}_R^{\mathrm{RTCS}}(f) &= \hat{a}'(f)H_R^{\mathrm{RTCS}}(f) \\
&= \hat{a}'(f)D_R'(f) + \hat{a}'(f)M(f)C_1(f)[R(f)\mathrm{HRTF}_R(f)]C_{2,R}(f)F_R(f)T_R(f) \\
&= \hat{a}'(f)D_R'(f) \\
&\quad + \hat{a}'(f)M(f)C_1(f)[R(f)\mathrm{HRTF}_R(f)]C_{2,R}(f)\left[\frac{D_R(f) + R(f)\mathrm{HRTF}_R(f) - D_R'(f)}{M(f)C_1(f)[R(f)\mathrm{HRTF}_R(f)]C_{2,R}(f)T_R(f)}\right]T_R(f) \\
&= \hat{a}'(f)D_R'(f) + \hat{a}'(f)[D_R(f) + R(f)\mathrm{HRTF}_R(f) - D_R'(f)] \\
&= \hat{a}'(f)[D_R(f) + R(f)\mathrm{HRTF}_R(f)].
\end{aligned} \tag{4.13b}$$

If the scattering caused by the RTCS microphone and headphones, denoted with primed variables, causes negligible impact on the acoustic pressure at the hypothetical point in front of the mouth, $\hat{a}'(f) = \hat{a}(f)$ and from Eq. (4.9), $\hat{b}_L^{\mathrm{RTCS}}(f) = \hat{b}_L^{\mathrm{room}}(f)$ and $\hat{b}_R^{\mathrm{RTCS}}(f) = \hat{b}_R^{\mathrm{room}}(f)$. In this case, the filtered RTCS system thus reproduces output signals identical to those produced by an actual room.

In order to calculate the filters, one needs to know $H_L^{\mathrm{room}}(f)$, $H_R^{\mathrm{room}}(f)$, $H_L^{\mathrm{RTCS}}(f)$, $H_R^{\mathrm{RTCS}}(f)$, $D_L'(f)$, and $D_R'(f)$. The first four can be obtained from the measurement framework associated with the signal flows outlined in Secs. 3.3.2 and 4.2.1. In addition, as seen from Eq. (4.4), $D_L'(f)$ and $D_R'(f)$ can be measured with a talker wearing the RTCS microphone and headphones in an anechoic chamber, but with the RTCS signal processing path turned off, such that $C_1(f) = C_{2,L}(f) = C_{2,R}(f) = 0$.

If we further assume that $D_L(f)$ is negligibly different from $D_L'(f)$ and $D_R(f)$ is negligibly different from $D_R'(f)$, then the expression for the filter can be further simplified. Equating Eqs. (4.10) and (4.5) again with this assumption (denoted with superscript ASM) yields

$$D_L(f) + R(f)\mathrm{HRTF}_L(f) = D_L'(f) + M(f)C_1(f)[R(f)\mathrm{HRTF}_L(f)]C_{2,L}(f)F_L^{\mathrm{ASM}}(f)T_L(f), \quad (4.14a)$$

$$D_R(f) + R(f)\mathrm{HRTF}_R(f) = D_R'(f) + M(f)C_1(f)[R(f)\mathrm{HRTF}_R(f)]C_{2,R}(f)F_R^{\mathrm{ASM}}(f)T_R(f), \quad (4.14b)$$

or

$$F_L^{\mathrm{ASM}}(f) = \frac{1}{M(f)C_1(f)C_{2,L}(f)T_L(f)}, \quad (4.15a)$$

$$F_R^{\mathrm{ASM}}(f) = \frac{1}{M(f)C_1(f)C_{2,R}(f)T_R(f)}. \quad (4.15b)$$

These filter expressions show explicitly that only the components related to the hardware of the RTCS are being compensated, or equalized. However, they involve three assumptions in their derivations: (1) the signal at the hypothetical point in space near the talker's mouth is unaffected by the presence of the RTCS hardware, (2) the signal that diffracts around the talker's head is unaffected by the RTCS hardware, and (3) the Fourier-transformed OBRIR [$R(f)HRTF_L(f)$ and $R(f)HRTF_R(f)$] housed within the RTCS is a faithful reproduction of the OBRIR the talker would hear when speaking in a given room. The expressions in Eq. (4.12) were derived with none of these assumptions. Therefore, the expressions in Eq. (4.12) account for both the responses of the RTCS hardware components, and the potential modifications to signal propagation due to the physical presence of the RTCS headgear.

4.2.3.2 *KEMAR Signal Paths in a Room and with the RTCS*

Because of the difficulty of measuring $H_L^{\mathrm{room}}(f)$, $H_R^{\mathrm{room}}(f)$, $H_L^{\mathrm{RTCS}}(f)$, $H_R^{\mathrm{RTCS}}(f)$, $D_L(f)$, $D_R(f)$, $D_L'(f)$, and $D_R'(f)$ for multiple talkers using the RTCS, a KEMAR

HATS was used to take measurements in both types of environments to measure reasonable estimates of the FRFs. These estimates were then used to compute the equalization filters introduced in Sec. 4.2.3.1.3. As suggested earlier, this mannequin was chosen for its unique average anatomical properties and other features (see Sec 3.3.1). To differentiate between FRFs for a live talker and those for KEMAR, the variables $H_{K,L}^{\text{room}}(f)$, $H_{K,R}^{\text{room}}(f)$, $H_{K,L}^{\text{RTCS}}(f)$, $H_{K,R}^{\text{RTCS}}(f)$, $D_{K,L}(f)$, $D_{K,R}(f)$, $D'_{K,L}(f)$, and $D'_{K,R}(f)$, are used to denote the latter. A few additional components to the KEMAR-related block diagrams contribute to these FRFs. However, the principles in assessing the FRFs remain the same.

Instead of defining the input signal as that of the radiated signal at an arbitrary point in space near the talker mouth, one might instead define it as the waveform $\hat{a}_s(f)$ used to drive the KEMAR mouth simulator. The output signals in the various measurement situations [$\hat{b}_{K,L}^{\text{room}}(f)$, $\hat{b}_{K,R}^{\text{room}}(f)$, $\hat{b}_{K,L}^{\text{RTCS}}(f)$, and $\hat{b}_{K,R}^{\text{RTCS}}(f)$] were also recorded using the KEMAR ear microphones at the entrances to the ear canals. Section 3.3.2 discussed the signal paths for an OBRIR measurement using KEMAR in a room. The following sections discuss the block diagrams for the KEMAR measurement system in the anechoic chamber, utilizing the RTCS system. They also address the associated mathematical formulations and the creation of equalization filters based on the resulting FRFs.

4.2.3.2.1 KEMAR using the RTCS

Figure 4.5 shows the block diagram for the signal flow of a KEMAR measurement while using the RTCS. In addition to the definitions given earlier, $D'_{K,L}(f)$ and $D'_{K,R}(f)$ are the FRFs

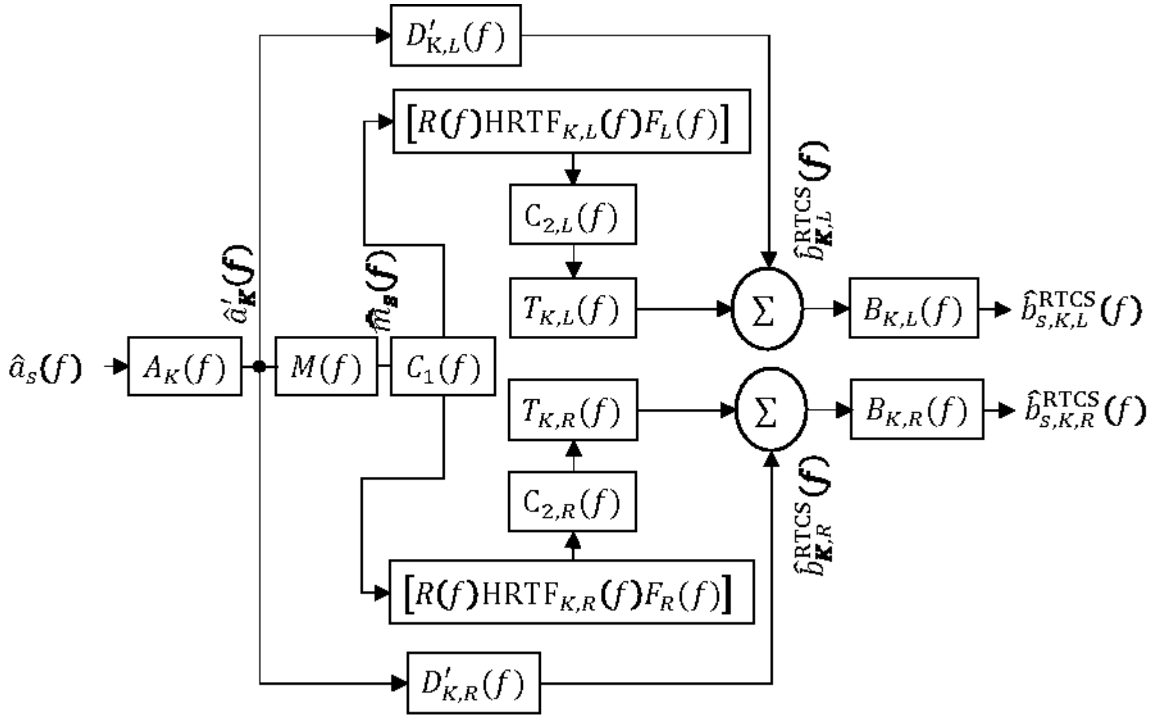

$D'_{K,L}(f)$
$[R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)]$
$C_{2,L}(f)$
$T_{K,L}(f)$
$\hat{b}^{\mathrm{RTCS}}_{\mathbf{K},L}(f)$
Σ
$B_{K,L}(f)$
$\hat{b}^{\mathrm{RTCS}}_{s,K,L}(f)$
$\hat{a}'_{\mathbf{K}}(f)$
$\hat{m}_s(f)$
$\hat{a}_s(f)$
$A_K(f)$
$M(f)$
$C_1(f)$
$T_{K,R}(f)$
Σ
$B_{K,R}(f)$
$\hat{b}^{\mathrm{RTCS}}_{s,K,R}(f)$
$C_{2,R}(f)$
$\hat{b}^{\mathrm{RTCS}}_{\mathbf{K},R}(f)$
$[R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)]$
$D'_{K,R}(f)$

Figure 4.5. Block diagram of the signal flow for a KEMAR measurement with the RTCS. The input signal is modified by diffraction around the head and headphones, and the signal processing of the RTCS. The signal flow between $\hat{a}_K(f)$ and $\hat{b}'_{K,L}(f)$ and $\hat{b}'_{K,R}(f)$ is the same as that for a live talker, as shown in Fig. 4.1 for the signal path of a talker with the RTCS. The input signal is modified by diffraction around the head and headphones, and the signal processing of the RTCS, except that the diffraction paths $D'_{K,L}(f)$ and $D'_{K,R}(f)$ are specific to the KEMAR anatomy.

for the initial diffraction paths of the potentially modified signal $\hat{a}'_K(f)$ around the KEMAR head and RTCS hardware to the blocked left and right ear canal openings. The FRF $M(f)$ represents the propagation of $\hat{a}'_K(f)$ to a head-worn microphone at the corner of the KEMAR mouth-simulator opening and through the microphone transduction system. The resulting microphone output signal $\hat{m}_s(f)$ is again processed via $C_1(f)$, $C_{2,L}(f)$, $C_{2,R}(f)$ (defined in Sec. 4.2.3.1.2), and the room responses loaded into the SIR2 plugin $[R(f)\mathrm{HRTF}_{K,L}(f)]$ and $[R(f)\mathrm{HRTF}_{K,R}(f)]$. The presence of a compensation filter is indicated with $F_L(f)$ and $F_R(f)$. The results of the RTCS convolutions are played to the left and right KEMAR ear simulators via the headphone transducer FRFs $T_{K,L}(f)$ and $T_{K,R}(f)$, respectively. The acoustic signals at the entrances to the left and right ear canals are then $\hat{b}_{K,L}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{K,R}^{\mathrm{RTCS}}(f)$, and the corresponding recorded output signals are $\hat{b}_{s,K,L}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{s,K,R}^{\mathrm{RTCS}}(f)$, respectively.

The relationships of these output signals to the input signal are described as follows:

$$\begin{gathered}\hat{b}_{s,K,L}^{\mathrm{RTCS}}(f) = \hat{a}_s(f)A_K(f)D'_{K,L}(f)B_{K,L}(f) + \\ \hat{a}_s(f)A_K(f)M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)\big]C_{2,L}(f)T_{K,L}(f)B_{K,L}(f),\end{gathered} \tag{4.16a}$$

$$\begin{gathered}\hat{b}_{s,K,R}^{\mathrm{RTCS}}(f) = \hat{a}_s(f)A_K(f)D'_{K,R}(f)B_{K,L}(f) + \\ \hat{a}_s(f)A_K(f)M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)\big]C_{2,R}(f)T_{K,R}(f)B_{K,R}(f).\end{gathered} \tag{4.16b}$$

The composite FRFs are then

$$\begin{gathered}H_{K,L}^{\mathrm{RTCS}}(f) = \frac{\hat{b}_{s,K,L}^{\mathrm{RTCS}}(f)}{\hat{a}_s(f)} \\ = A_K(f)\big[D'_{K,L}(f) + M(f)C_1(f)\big\{R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)\big\}C_{2,L}(f)T_{K,L}(f)\big]B_{K,L}(f),\end{gathered} \tag{4.17a}$$

$$\begin{gathered}H_{K,R}^{\mathrm{RTCS}}(f) = \frac{\hat{b}_{s,K,R}^{\mathrm{RTCS}}(f)}{\hat{a}_s(f)} \\ = A_K(f)\big[D'_{K,R}(f) + M(f)C_1(f)\big\{R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)\big\}C_{2,R}(f)T_{K,R}(f)\big]B_{K,R}(f)\end{gathered} \tag{4.17b}$$

4.2.3.2.2 Filter Derivation for KEMAR OBRIR Measurements

Similar to the case for a live talker, filters may be derived to account for the differences between $\hat{b}_{s,K,L}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{s,K,L}^{\mathrm{room}}(f)$, and $\hat{b}_{s,K,R}^{\mathrm{RTCS}}(f)$ and $\hat{b}_{s,K,R}^{\mathrm{room}}(f)$, respectively. They may then be included as part of the RTCS processing, in the SIR2 plugin. Since we want $H_{K,L}^{\mathrm{room}}(f) = H_{K,L}^{\mathrm{RTCS}}(f)$ and $H_{K,R}^{\mathrm{room}}(f) = H_{K,R}^{\mathrm{RTCS}}(f)$, we solve for the filters that make this possible. Equating Eqs. (3.1) and (4.17) yields

$$\begin{gathered}D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f) = D'_{K,L}(f) + \\ M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,L}(f)\big][F_L(f)]C_{2,L}(f)T_{K,L}(f),\end{gathered} \tag{4.18a}$$

$$\begin{gathered}D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f) = \ D'_{K,R}(f) + \\ M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,R}(f)\big][F_R(f)]C_{2,R}(f)T_{K,R}(f).\end{gathered} \tag{4.18b}$$

The filter FRFs are then

$$F_L(f) = \frac{D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f) - D'_{K,L}(f)}{M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,L}(f)\big]C_{2,L}(f)T_{K,L}(f)}, \tag{4.19a}$$

$$F_R(f) = \frac{D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f) - D'_{K,R}(f)}{M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,R}(f)\big]C_{2,R}(f)T_{K,R}(f)}. \tag{4.19b}$$

If these filters are included in the RTCS processing to filter out its nonideal characteristics, the following results should occur:

$$\begin{aligned}
&\hat{b}_{s,L}^{\mathrm{RTCS}}(f) = \hat{a}_s(f)H_{K,L}^{\mathrm{RTCS}}(f) \\
&\qquad =\hat{a}_s(f)\ A_K(f)\big[D'_{K,L}(f) + \\
&M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)F_L(f)C_{2,L}(f)T_{K,L}(f)\big]B_{K,L}(f) \\
&= \hat{a}_s(f)\ A_K(f)\big[D'_{K,L}(f) + M(f)C_1(f)\big[R(f)HRTF_{K,L}(f)F_L(f)\big]C_{2,L}(f)T_{K,L}(f)\big]B_{K,L}(f) \\
&= \hat{a}_s(f)A_K(f)B_{K,L}(f)D'_{K,L}(f) \\
&+ \hat{a}_s(f)A_K(f)B_{K,L}(f)M(f)C_1(f)R(f)HRTF_{K,L}(f)\frac{D_{K,L}(f) + R(f)HRTF_{K,L}(f) - D'_{K,L}(f)}{M(f)C_1(f)[R(f)HRTF_{K,L}(f)]C_{2,L}(f)T_{K,L}(f)} \\
&= \hat{a}_s(f)A_K(f)B_{K,L}(f)\Big\{D'_{K,L}(f) \\
&+ M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)\left[\frac{D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f) - D'_{K,L}(f)}{M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)\ C_{2,L}(f)T_{K,L}(f)}\right]C_{2,L}(f)T_{K,L}(f)\Big\} \\
&= \hat{a}_s(f)A_K(f)B_{K,L}(f)\{D'_{K,L}(f) + \big[D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f) - D'_{K,L}(f)\big]\} \\
&= \ \hat{a}_s(f)A_K(f)B_{K,L}(f)\big[D_{K,L}(f) + R(f)\mathrm{HRTF}_{K,L}(f)\big] \\
&= \hat{b}_{s,L}^{\mathrm{room}}(f),
\end{aligned} \tag{4.20a}$$

and

$$\hat{b}_{s,R}^{\mathrm{RTCS}}(f) = \hat{a}_s(f) H_{K,R}^{\mathrm{RTCS}}(f)$$

$$=\hat{a}_s(f)\ A_K(f)\big[D'_{K,R}(f) +$$

$$M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)\big]C_{2,R}(f)T_{K,R}(f)\big]B_{K,R}(f)$$

$$= \hat{a}_s(f)\ A_K(f)\big[D'_{K,R}(f)$$

$$+ M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,R}(f)F_R(f)\big]C_{2,R}(f)T_{K,R}(f)\big]B_{K,R}(f)$$

$$= \hat{a}_s(f)A_K(f)B_{K,R}(f)D'_{K,R}(f)$$

$$+ \hat{a}_s(f)A_K(f)B_{K,R}(f)M(f)C_1(f)R(f)\mathrm{HRTF}_{K,R}(f)\frac{D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f) -}{M(f)C_1(f)\big[R(f)\mathrm{HRTF}_{K,R}(f)\big]C_{2,R}} \tag{4.20b}$$

$$= \hat{a}_s(f)A_K(f)B_{K,R}(f)\left\{D'_{K,R}(f)\right.$$

$$+ M(f)C_1(f)R(f)\mathrm{HRTF}_{K,R}(f)\left[\frac{D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f) - D'_{K,R}(f)}{M(f)C_1(f)R(f)\mathrm{HRTF}_{K,R}(f)\ C_{2,R}(f)T_{K,R}(f)}\right]C_{2,L}(f)T$$

$$= \hat{a}_s(f)A_K(f)B_{K,R}(f)\{D'_{K,R}(f) + \big[D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f) - D'_{K,R}(f)\big]\}$$

$$=\ \hat{a}_s(f)A_K(f)B_{K,R}(f)\big[D_{K,R}(f) + R(f)\mathrm{HRTF}_{K,R}(f)\big]$$

$$= \hat{b}_{s,R}^{\mathrm{room}}(f).$$

Comparison of Eqs. (4.12) and (4.19) reveal that the filters derived for a live talker and the KEMAR mannequin are similar insofar as the KEMAR anatomy is similar to that of the live talker, as discussed in Sec. 3.3.1.

If, as in Sec. 4.2.3.1.3, we assume that $D_{K,L}(f)$ and $D'_{K,L}(f)$, and $D_{K,R}(f)$ and $D'_{K,R}(f)$ are negligibly different, then Eq. (4.19) can be further simplified to

$$F_L^{\mathrm{ASM}}(f) = \frac{1}{M(f)C_1(f)C_{2,L}(f)T_{K,L}(f)}, \tag{4.21a}$$

$$F_R^{\mathrm{ASM}}(f) = \frac{1}{M(f)C_1(f)C_{2,R}(f)T_{K,R}(f)}. \tag{4.21b}$$

As with the simplified filters for a talker in Eq. (4.15), alternative expressions assume the RTCS headgear does not affect the propagation and diffraction of the acoustic signal from the point near the mouth simulator to the ear simulators, and that the OBRIR housed in the RTCS is a faithful reproduction of the OBRIR measured in a room.

4.2.4 Inversion Filter Computation

Sections 4.2.3.1.3 and 4.2.3.2.2 have discussed the theoretical computation of inversion filters. The IR measurement procedures, filter designs, MATLAB implementations, and their limitations are described in this section. Section 3.3.2 discussed the OBRIR measurements performed with the KEMAR mannequin in a room, and their subsequent modification for use in the RTCS. Section 4.2.3.2.1 discussed the OBRIR measurements with a filtered RTCS. However, additional measurements were needed to measure the necessary FRFs to compute the theoretical filters of Eqs. (4.19) and (4.21).

4.2.4.1 *Additional IR Measurements*

The software package EASERA was used with an RME Fireface interface and direct cabling to the KEMAR mannequin to measure the OBRIRs in the anechoic chamber. The purpose of these measurements was to extract the necessary FRFs to compute the theoretical filters derived in Sec. 4.2.3.2. The output files from the EASERA measurements are IRs between the signal sent to the KEMAR mouth simulator and those produced by the KEMAR left and right ear microphones. In addition, the IR between the signal sent to the KEMAR mouth simulator and that produced by the head-worn microphone was measured whenever KEMAR wore the microphone.

4.2.4.1.1 Equalization filter derived with assumptions

If the theoretical filter of Eq. (4.21) was desired, two OBRIR measurement cases were required with KEMAR in the anechoic chamber: (1) with the RTCS turned off, and (2) with the RTCS turned on while utilizing a delta function for the SIR2 RIR convolution. These scenarios are depicted in the block diagrams of Figs. 4.6 and 4.7 respectively. In these figures, and corresponding equations, the superscripts RTCS δ denotes a measurement in which the the RTCS was enabled and utilized a δ function, and ANCH denotes a measurement in which the RTCS was disabled.

The EASERA IR measurement in each case utilized a log sweep from 10 Hz to 24 kHz, repeated 10 times and averaged within the software. The IRs between swept-sine input signal and the ear microphone and head-worn microphone output signals were available from the

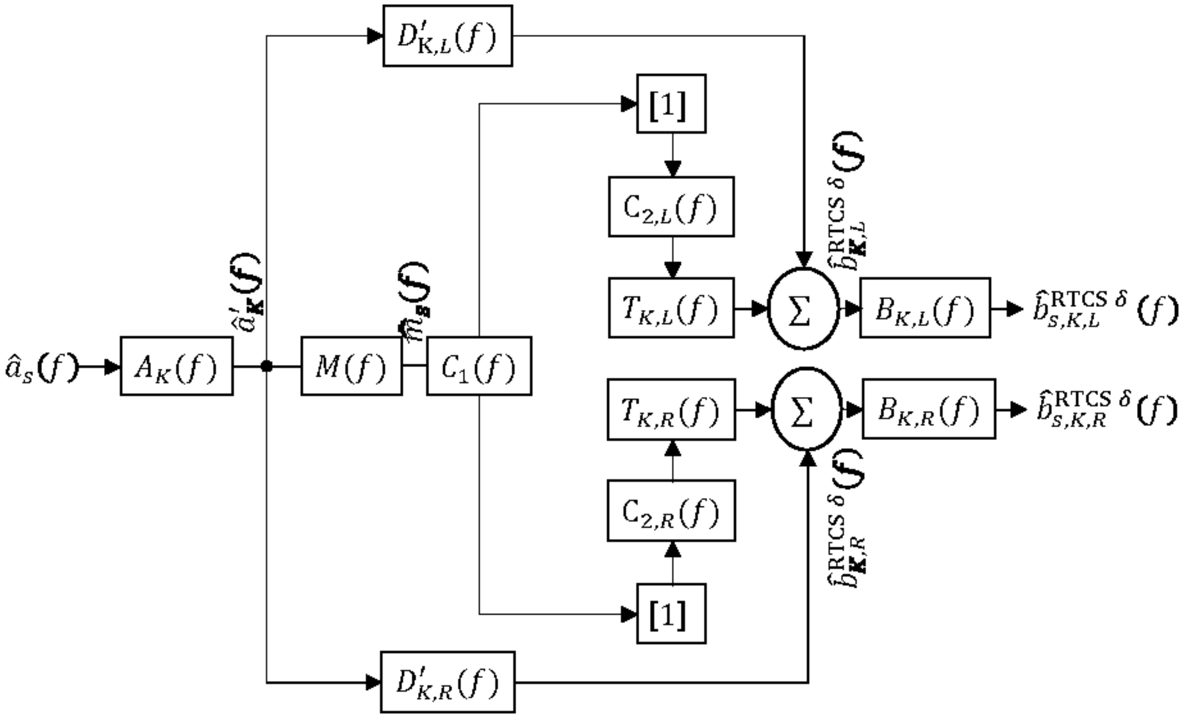

$\hat{a}_s(f)$
$A_K(f)$
$\hat{a}'_{\mathbf{K}}(f)$
$M(f)$
$\hat{m}_s(f)$
$C_1(f)$
$D'_{K,L}(f)$
[1]
$C_{2,L}(f)$
$T_{K,L}(f)$
Σ
$\hat{b}^{RTCS\ \delta}_{\mathbf{K},L}(f)$
$B_{K,L}(f)$
$\hat{b}^{RTCS\ \delta}_{s,K,L}(f)$
$T_{K,R}(f)$
Σ
$B_{K,R}(f)$
$\hat{b}^{RTCS\ \delta}_{s,K,R}(f)$
$\hat{b}^{RTCS\ \delta}_{\mathbf{K},R}(f)$
$C_{2,R}(f)$
[1]
$D'_{K,R}(f)$

Figure 4.6. Block diagram for impulse response measurements of KEMAR wearing the active RTCS and utilizing a delta function in the SIR2 plugin.

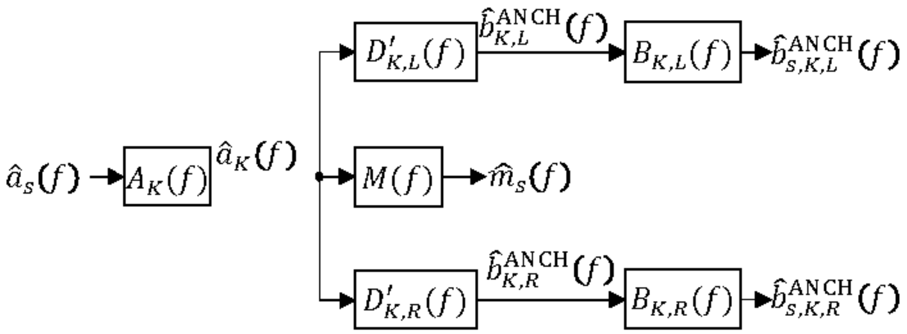

$\hat{a}_s(f)$
$A_K(f)$
$\hat{a}_K(f)$
$D'_{K,L}(f)$
$\hat{b}^{\mathrm{ANCH}}_{K,L}(f)$
$B_{K,L}(f)$
$\hat{b}^{\mathrm{ANCH}}_{s,K,L}(f)$
$M(f)$
$\hat{m}_s(f)$
$D'_{K,R}(f)$
$\hat{b}^{\mathrm{ANCH}}_{K,R}(f)$
$B_{K,R}(f)$
$\hat{b}^{\mathrm{ANCH}}_{s,K,R}(f)$

Figure 4.7. Block diagram for impulse response measurements of KEMAR wearing the RTCS headgear in the anechoic chamber with the RTCS turned off.

software as *.wav files, or as *.etx files from which textual information about the measurement could be extracted and imported into MATLAB for further processing.

In the first case, with the RTCS turned off, the Fourier transform of the IRs from EASERA gave the FRFs $\hat{b}^{\mathrm{ANCH}}_{s,K,L}(f)/\hat{a}_s(f)$, $\hat{b}^{\mathrm{ANCH}}_{s,K,R}(f)/\hat{a}_s(f)$, and $\hat{m}_s(f)/\hat{a}_s\,(f)$. However, the IRs for $\hat{b}^{\mathrm{ANCH}}_{s,K,L}(f)/\hat{m}_{s(f)}$ and $\hat{b}^{\mathrm{ANCH}}_{s,K,R}(f)/\hat{m}_s(f)$ are desired to eliminate the effects of the KEMAR mouth simulator. They are found simply in Eq. 4.22 by dividing $\hat{b}^{\mathrm{ANCH}}_{s,K,L}(f)/\hat{a}_s(f)$ and $\hat{b}^{\mathrm{ANCH}}_{s,K,R}(f)/\hat{a}_s(f)$ by $\hat{m}_s(f)/\hat{a}_s(f)$, which would correspond to a time-domain deconvolution:

$$\frac{\hat{b}^{\mathrm{ANCH}}_{s,K,L}(f)}{\hat{m}_{s(f)}} = \frac{\hat{b}^{\mathrm{ANCH}}_{s,K,L}(f)/\hat{a}_s(f)}{\hat{m}_s(f)/\hat{a}_s\,(f)} = \frac{A_K(f)D'_{K,L}(f)B_{K,L}(f)}{A_K(f)M(f)} = \frac{D'_{K,L}(f)B_{K,L}(f)}{M(f)} \tag{4.22a}$$

$$\frac{\hat{b}^{\mathrm{ANCH}}_{s,K,R}(f)}{\hat{m}_{s(f)}} = \frac{\hat{b}^{\mathrm{ANCH}}_{s,K,R}(f)/\hat{a}_s(f)}{\hat{m}_s(f)/\hat{a}_s\,(f)} = \frac{A_K(f)D'_{K,R}(f)B_{K,R}(f)}{A_K(f)M(f)} = \frac{D'_{K,R}(f)B_{K,R}(f)}{M(f)} \tag{4.22b}$$

In the second case, with the RTCS turned on and utilizing a delta function in the SIR2 convolver (denoted with the superscript RTCS δ), the FRFs $\hat{b}^{\mathrm{RTCS}\,\delta}_{s,K,L}(f)/\hat{m}_s(f)$ and

$\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)/\widehat{m}_s(f)$ are similarly found by dividing $\hat{b}_{s,K,L}^{\mathrm{RTCS}\,\delta}(f)/\hat{a}_s\,(f)$ and $\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)/\hat{a}_s(f)$ by $\widehat{m}_s(f)/\hat{a}_s\,(f)$.

$$\begin{aligned}\frac{\hat{b}_{s,K,L}^{\mathrm{RTCS}\,\delta}(f)}{\widehat{m}_{s(f)}} &= \frac{\hat{b}_{s,K,L}^{\mathrm{RTCS}\,\delta}(f)/\hat{a}_s(f)}{\widehat{m}_s(f)/\hat{a}_s\,(f)} \\ &= \frac{A_K(f)\big[D'_{K,L}(f)+M(f)C_1(f)C_{2,L}(f)T_{K,L}(f)\big]B_{K,L}(f)}{A_K(f)M(f)} \\ &= \frac{\big[D'_{K,L}(f)+M(f)C_1(f)C_{2,L}(f)T_{K,L}(f)\big]B_{K,L}(f)}{M(f)}\end{aligned} \tag{4.23a}$$

$$\begin{aligned}\frac{\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)}{\widehat{m}_{s(f)}} &= \frac{\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)/\hat{a}_s(f)}{\widehat{m}_s(f)/\hat{a}_s\,(f)} \\ &= \frac{A_K(f)\big[D'_{K,R}(f)+M(f)C_1(f)C_{2,R}(f)T_{K,R}(f)\big]B_{K,R}(f)}{A_K(f)M(f)} \\ &= \frac{\big[D'_{K,R}(f)+M(f)C_1(f)C_{2,R}(f)T_{K,R}(f)\big]B_{K,R}(f)}{M(f)}\end{aligned} \tag{4.23b}$$

Subtraction of $\hat{b}_{s,K,L}^{\mathrm{ANCH}}(f)/\widehat{m}_s(f)$ from $\hat{b}_{s,K,L}^{\mathrm{RTCS}\,\delta}(f)/\widehat{m}_s\,(f)$ and $\hat{b}_{s,K,R}^{\mathrm{ANCH}}(f)/\widehat{m}_s(f)$ from $\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)/\widehat{m}_s(f)$ removes the effects of the recording hardware used in the measurement and leaves the multiplied FRFs containing only information about the RTCS computation [$C_1(f)$, $C_{2,L}(f)$, $C_{2,R}(f)$], headphones [$T_{K,L}(f)$, $T_{K,R}(f)$], and KEMAR ear microphones [$B_{K,L}(f)$, and $B_{K,R}(f)$]:

$$\begin{aligned}
&\frac{\hat{b}_{s,K,L}^{\mathrm{RTCS}\,\delta}(f)}{\hat{m}_{s(f)}}-\frac{\hat{b}_{s,K,L}^{\mathrm{ANCH}}(f)}{\hat{m}_{s(f)}}\\
&=\frac{[D'_{K,L}(f)+M(f)C_1(f)C_{2,L}(f)T_{K,L}(f)]B_{K,L}(f)}{M(f)}-\frac{D'_{K,L}(f)B_{K,L}(f)}{M(f)}\\
&=\frac{M(f)C_1(f)C_{2,L}(f)T_{K,L}(f)B_{K,L}(f)}{M(f)}\\
&=\ C_1(f)C_{2,L}(f)T_{K,L}(f)B_{K,L}(f)
\end{aligned}\tag{4.24a}$$

$$\begin{aligned}
&\frac{\hat{b}_{s,K,R}^{\mathrm{RTCS}\,\delta}(f)}{\hat{m}_{s(f)}}-\frac{\hat{b}_{s,K,R}^{\mathrm{ANCH}}(f)}{\hat{m}_{s(f)}}\\
&=\frac{[D'_{K,R}(f)+M(f)C_1(f)C_{2,R}(f)T_{K,R}(f)]B_{K,R}(f)}{M(f)}-\frac{D'_{K,R}(f)B_{K,R}(f)}{M(f)}\\
&=\frac{M(f)C_1(f)C_{2,R}(f)T_{K,R}(f)B_{K,R}(f)}{M(f)}\\
&=\ C_1(f)C_{2,R}(f)T_{K,R}(f)B_{K,R}(f)
\end{aligned}\tag{4.24b}$$

The inverses of the expressions in Eq. (4.24) very nearly matches the theoretical filter expression of Eq. (4.21). Indeed, if $M(f)$, $B_{K,L}(f)$, and $B_{K,R}(f)$ are flat over the frequency range of interest, then these measurements should yield a valid equalization filter. Note, however, that both of these measurements in the anechoic chamber (with the RTCS disabled and with the RTCS utilizing a δ function) occured with the RTCS headphones present. This means the filters derived to equalize the RTCS components using this method do not compensate for the presence or lack of presence of the RTCS headgear and their scattering of initial diffracted sound paths, as was assumed in the theoretical derivation of Eq. 4.21. However, any compensation that would occur for this scattering would be in the early parts of the IRs—well within the latency of the RTCS (~6 ms).

Informal listening tests suggested that the presence of the RTCS headgear was not deleterious, and would not affect the final results significantly. However, if it was deemed necessary to account for the presence of the hardware, a RTCS using stereo loudspeakers instead of closely-spaced headphones could be designed and implemented. These would need to be placed a few meters away from the RTCS user, and an interaural cross-talk cancellation filter would need to be implemented [92].

4.2.4.1.2 Equalization Filter derived from room measurements

Alternatively, measurements could be made to produce the theoretical filter of Eq. (4.19), which did not assume the presence of the RTCS headgear was negligible. These required assessment of (1) the KEMAR OBRIR in a room, (2) the KEMAR OBRIR in the anechoic chamber while wearing the RTCS headgear, but with the RTCS disengaged, and (3) the KEMAR OBRIR in the anechoic chamber using the RTCS with the modified OBRIR from the room in (1) after the manner of Sec. 3.3. The Fourier transforms of these IRs give the needed FRFs for computation of the filter in Eq. (4.19). The first situation was the subject of Section 3.3.2. The second was depicted in the block diagram of Fig. 4.7. The third was depicted in the block diagram of Fig. 4.8. In this figure and corresponding equations, the superscript RTCS UFILT denotes a measurement in which the RTCS was enabled and housed a room OBRIR, but had no compensation filter in place.

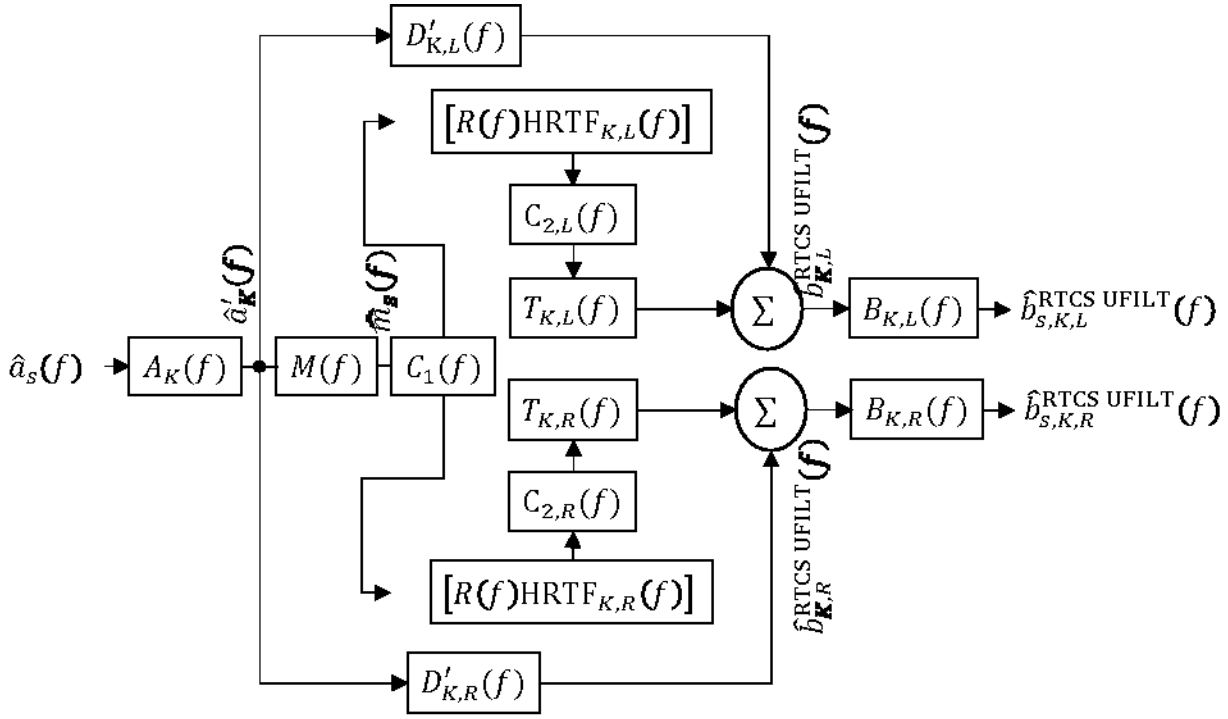

$D'_{K,L}(f)$
$[R(f)\mathrm{HRTF}_{K,L}(f)]$
$C_{2,L}(f)$
$T_{K,L}(f)$
Σ
$\hat{b}^{\text{RTCS UFILT}}_{K,L}(f)$
$B_{K,L}(f)$
$\hat{b}^{\text{RTCS UFILT}}_{s,K,L}(f)$
$\hat{a}'_{K}(f)$
$\hat{m}_{s}(f)$
$\hat{a}_s(f)$
$A_K(f)$
$M(f)$
$C_1(f)$
$T_{K,R}(f)$
Σ
$B_{K,R}(f)$
$\hat{b}^{\text{RTCS UFILT}}_{s,K,R}(f)$
$C_{2,R}(f)$
$\hat{b}^{\text{RTCS UFILT}}_{K,R}(f)$
$[R(f)\mathrm{HRTF}_{K,R}(f)]$
$D'_{K,R}(f)$

Figure 4.8. Block Diagram for KEMAR OBRIR measurement utilizing the unfiltered RTCS.

In the third case, the level of the room OBRIR $[R(f)\mathrm{HRTF}_{K,L}(f)$ and $R(f)\mathrm{HRTF}_{K,R}(f)]$ in SIR2 had to be adjusted such that the IRs corresponding to the inverse Fourier transforms of $\hat{b}^{\text{RTCS UFILT}}_{s,K,L}(f)/\hat{a}_s(f)$ and $\hat{b}^{\text{RTCS UFILT}}_{s,K,R}(f)/\hat{a}_s(f)$ had the same levels as those corresponding to $\hat{b}^{\text{room}}_{s,K,L}(f)/\hat{a}_s(f)$ and $\hat{b}^{\text{room}}_{s,K,R}(f)/\hat{a}_s(f)$ for the initial 100 ms of the IRs, respectively. This ensured that the filter would not be set at an unrealistic level so that it could be effective when implemented in the RTCS. The Fourier transforms of the IRs from EASERA for this third case are

$$\begin{aligned}\frac{\hat{b}^{\text{RTCS UFILT}}_{s,K,L}(f)}{\hat{a}_s(f)}& \\ &= A_K(f)\big[D'_{K,L}(f) \\ &+ M(f)C_1(f)R(f)\mathrm{HRTF}_{K,L}(f)C_{2,L}(f)T_{K,L}(f)\big]B_{K,L}(f),\end{aligned} \tag{4.25a}$$

$$\frac{\hat{b}_{s,K,R}^{\text{RTCS UFILT}}(f)}{\hat{a}_s(f)} \tag{4.25b}$$

$$= A_K(f)\big[D'_{K,R}(f)$$

$$+ M(f)C_1(f)R(f)\text{HRTF}_{K,R}(f)C_{2,L}(f)T_{K,R}(f)\big]B_{K,R}(f).$$

The filter expressions in Eq. (4.19) can be rewritten to include the Fourier transform of the IRs of Eqs. (4.22), (4.23), and (4.25).

$$\begin{aligned} F_L(f) &= \frac{D_{K,L}(f) + R(f)\text{HRTF}_{K,L}(f) - D'_{K,L}(f)}{M(f)C_1(f)\big[R(f)\text{HRTF}_{K,L}(f)\big]C_{2,L}(f)T_{K,L}(f)} \\ &= \frac{A_K(f)B_{K,L}(f)\big[D_{K,L}(f) + R(f)\text{HRTF}_{K,L}(f) - D'_{K,L}(f)\big]}{A_K(f)B_{K,L}(f)\big[M(f)C_1(f)R(f)\text{HRTF}_{K,L}(f)C_{2,L}(f)T_{K,L}(f)\big]} \\ &= \frac{\hat{b}_{s,K,L}^{\text{room}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,L}^{\text{ANCH}}(f)/\hat{a}_s(f)}{\hat{b}_{s,K,L}^{\text{RTCS UFILT}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,L}^{\text{ANCH}}(f)/\hat{a}_s(f)} \end{aligned} \tag{4.26a}$$

$$\begin{aligned} F_R(f) &= \frac{D_{K,R}(f) + R(f)\text{HRTF}_{K,R}(f) - D'_{K,R}(f)}{M(f)C_1(f)\big[R(f)\text{HRTF}_{K,R}(f)\big]C_{2,R}(f)T_{K,R}(f)} \\ &= \frac{A_K(f)B_{K,R}(f)\big[D_{K,R}(f) + R(f)\text{HRTF}_{K,R}(f) - D'_{K,R}(f)\big]}{A_K(f)B_{K,R}(f)\big[M(f)C_1(f)R(f)\text{HRTF}_{K,R}(f)C_{2,R}(f)T_{K,R}(f)\big]} \\ &= \frac{\hat{b}_{s,K,R}^{\text{room}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,R}^{\text{ANCH}}(f)/\hat{a}_s(f)}{\hat{b}_{s,K,R}^{\text{RTCS UFILT}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,R}^{\text{ANCH}}(f)/\hat{a}_s(f)}. \end{aligned} \tag{4.26b}$$

The modifications of $\hat{b}_{s,K,L}^{\text{room}}(f)/\hat{a}_s(f)$ and $\hat{b}_{s,K,R}^{\text{room}}(f)/\hat{a}_s(f)$ {to remove the KEMAR mouth simulator effects, to force an exponential decay to below the noise floor, to truncate the initial 6 ms, and to bandpass filter between 60 Hz and 10,500 Hz [see Eq. (4.8)]} effectively approximated the numerator of Eq.(4.26). If the same modifications are performed on the unfiltered RTCS OBRIR measurement, the denominator of Eq. (4.26) is also effectively approximated. Specifically, it is the truncation of the initial 6 ms of the IRs that serves as the substitute for the subtraction of $\hat{b}_{s,K,L}^{\text{ANCH}}(f)/\hat{a}_s(f)$ and $\hat{b}_{s,K,R}^{\text{ANCH}}(f)/\hat{a}_s(f)$, as these FRFs only involve information about the initial diffracted sound from the KEMAR mouth simulator to the KEMAR ear microphones, as determined in an anechoic chamber. The truncation of the initial 6 ms removes that information from the IRs corresponding to $\hat{b}_{s,K,L}^{\text{room}}(f)/\hat{a}_s(f)$, $\hat{b}_{s,K,R}^{\text{room}}(f)/\hat{a}_s(f)$, $\hat{b}_{s,K,L}^{\text{RTCS UFILT}}(f)/\hat{a}_s(f)$, and $\hat{b}_{s,K,R}^{\text{RTCS UFILT}}(f)/\hat{a}_s(f)$. In addition, the removal of the mouth-simulator effects from the measured OBRIRs occurs in both the numerator and denominator, so the net effect does not impact the expression for the equalization filter. Thus, Eq. (4.26) can be rewritten with the OBRIR modifications included:

$$F_L(f) = \frac{\hat{b}_{s,K,L}^{\text{room}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,L}^{\text{ANCH}}(f)/\hat{a}_s(f)}{\hat{b}_{s,K,L}^{\text{RTCS UFILT}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,L}^{\text{ANCH}}(f)/\hat{a}_s(f)}$$

$$\approx \frac{\left[\frac{\hat{b}_{s,K,L}^{\text{room}}(f)}{\widehat{m}_s(f)}\right]_{\text{filt,trunc}}}{\left[\frac{\hat{b}_{s,K,L}^{\text{RTCS UFILT}}(f)}{\widehat{m}_s(f)}\right]_{\text{filt,trunc}}} \tag{4.27a}$$

$$F_R(f) = \frac{\hat{b}_{s,K,R}^{\text{room}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,R}^{\text{ANCH}}(f)/\hat{a}_s(f)}{\hat{b}_{s,K,R}^{\text{RTCS UFILT}}(f)/\hat{a}_s(f) - \hat{b}_{s,K,R}^{\text{ANCH}}(f)/\hat{a}_s(f)}$$

$$\approx \frac{\left[\frac{\hat{b}_{s,K,R}^{\text{room}}(f)}{\widehat{m}_s(f)}\right]_{\text{filt,trunc}}}{\left[\frac{\hat{b}_{s,K,R}^{\text{RTCS UFILT}}(f)}{\widehat{m}_s(f)}\right]_{\text{filt,trunc}}}. \tag{4.26b}$$

These expressions demonstrate how the measurements outlined above result in a valid equalization filter for the RTCS.

4.2.4.2 *Filter Calculation and Implementation*

The filters discussed in the previous section were not computed as simply as described because the electro-acoustic RTCS system is a mixed-phase system [59]. A simple inversion of its FRFs would therefore result in acausal and likely unstable filter IRs [93]. The inversions were then performed after the method of Scharer and Lindau [94]. In their work, they drew strongly from the work of Mourjopoulos [95] on digital filter design to equalize for room acoustics, in which the measured system IR was decomposed into minimum- and maximum-phase components through cepstral analysis [96,97]. Each of these components was inverted separately. The filter IR derived from the inversions was also delayed to ensure the acausal part shifted into the positive part of the time domain [95,98,99].

In a basic sense, the measurement of the IR and the computation from that measurement was for a single system with the source (KEMAR mouth simulator) and receivers (KEMAR ear microphones) in time-invariant positions. However, the exact positioning of the RTCS headgear could affect the measurement and filter. This was especially true at high frequencies, wherein the geometric details of KEMAR and the RTCS headgear became significant compared to the wavelength of sound. Perfect equalization from measurements from one headphone positioning would therefore result in a mismatch for even slight variations in subsequent positionings. Therefore, several measurements were performed with minor adjustments to headphone positions to provide an average that would smooth high-frequency notches in the measured FRFs and reduce the perceptible artifacts produced by positioning changes [94]. Complex averaging reduced the depths of notches to be compensated for, thus reducing the amplitudes of the peaks in the equalization filter. Scharer and Lindau pointed out that notches in the filter are less perceptually noticeable than are peaks [94,100,101].

Code made available by Brinkmann[102] was used to compute the equalization filter for the RTCS using a least-mean-squares (LMS) algorithm to minimize the error in magnitude between the target and compensation functions [99,103]. The equalization filter was also computed to be minimum-phase, after the method of Norcross[104]. The equalization filter was then given by

$$H_{\mathrm{eq}}(f) = \frac{H_{\mathrm{target}}(f) H_{\mathrm{input}}^{*}(f)}{\left|H_{\mathrm{input}}(f)\right|^{2}} \tag{4.28}$$

where $H_{\mathrm{eq}}(f)$ is the equalization filter, $H_{\mathrm{target}}\ (f)$ is the desired target function, $H_{\mathrm{input}}(f)$ is the input FRF, and * denotes its complex conjugate. The input FRF $H_{\mathrm{input}}(f)$ is the current FRF that needs to be compensated by the equalization filter to give an end result similar to the desired target function.

The theoretical filter of Eq. (4.26) was computed by incorporating this approach. In this case, $H_{\mathrm{target}}(f)$ was the modified room OBRIR described in Eq. (4.8) and $H_{\mathrm{input}}(f)$ was the modified unfiltered RTCS OBRIR. The modifications of both were described in Sec. 3.3.2 and included (1) the removal of the mouth simulator effects, (2) the removal of the audible noise floor by the application of an exponentially-decaying window, (3) the truncation of the initial portion of the IR by 6 ms, and (4) the application of a bandpass filter with cutoff frequencies of 60 Hz and 10,500 Hz. Both the target function and the input function had separate information for left and right channels, so technically two filters were created, one for each ear. In the computation, both the input and target functions were smoothed in 6th octave bands to avoid unnecessarily sharp peaks and dips in the compensation filter. The length of the filter varied, based on the length of the RIR being considered.

4.2.4.3 *Simulation of Inversion Filter Effectiveness*

Simulations were performed to check the accuracies of the filters and attempt to predict their effectiveness in the RTCS. This was done by multiplying the modified RTCS IR measurements by the filter in the frequency domain and comparing the result to the modified

RIR in the frequency domain. Deviations of the magnitude responses of the compensated RTCS IRs from the target function (room IR magnitude response) were computed using an auditory filter bank that modeled the behavior of the human auditory system [105]. This was constructed from 40 overlapping equivalent rectangular bandwidth (ERB) filters using MakeERBFilters and ERBFilterBank from the Auditory Toolbox for MATLAB [106-108]. For each band of the filter bank, the decibel difference between the compensated RTCS and the room was calculated. Values beyond the frequency range of interest (100 Hz to 10,500 Hz) were discarded. The arithmetic mean and standard deviation across the bands of the filter bank was reported. In addition, the arithmetic mean and standard deviation of the decibel difference between the inverse Fourier transform of the compensated RTCS and the room IR was calculated. Since the IRs and filters considered here were in waveform audio file (*.wav) format, their amplitudes were constrained to waveform values (-1 to 1), which made their relative amplitudes uncertain. However, the simulations provided best-case scenarios for comparing the compensated RTCS responses to the original room responses.

An example of the simulation results for one of the rooms (the reverberation chamber with two absorbing wedges) is shown in Figs. 4.9 and 4.10. Simulation results for the additional rooms are given in Appendix D. Figure 4.9 first provides information on the inversion filter computation, while Fig. 4.10 shows the simulation results. In Fig. 4.9, the lower two traces report energy time curves (ETCs). These are not true ETCs, but simply a time-domain representation of the IR, given by $10 * \log_{10}(IR^2)$. The upper traces show the magnitude of the FRFs corresponding to the IRs. Both the time- and frequency-domain traces are useful in describing the IRs and help the researcher find anomalies such as comb filtering in the time-domain, or sharp peaks and dips in the frequency domain, neither of which are present in Fig. 4.9. In Fig.

4.10, the “raw data” referred to is the ten input OBRIRs that were smoothed and averaged to create the input function, as described in Sec. 4.2.4.2. They serve as a simulation of ten additional OBRIR measurements with the compensation filter to predict its performance. The simulation results predict that the compensation filter applied to an OBRIR nearly matches the target function, with some variation. However, a mean level error of about 1 dB and a standard deviation less than 2 dB, combined with the graphic showing that the maximum error is less than 3 dB at any frequency, indicate that the variations are small and the target function is achievable.

4.3 Conclusions

The development of the RTCS has been discussed in detail, including comparisons with the work that influenced its design and implementation at BYU. The major contributions of this chapter include its diagrammatical and theoretical derivations of system responses associated with a talker/listener speaking in a room and with a RTCS. They also include the development of required computations for creating equalization filters that compensate for undesirable RTCS effects in order to produce more faithful representations of simulated acoustic environments. The following chapter discusses the results of the implemented compensation filter.

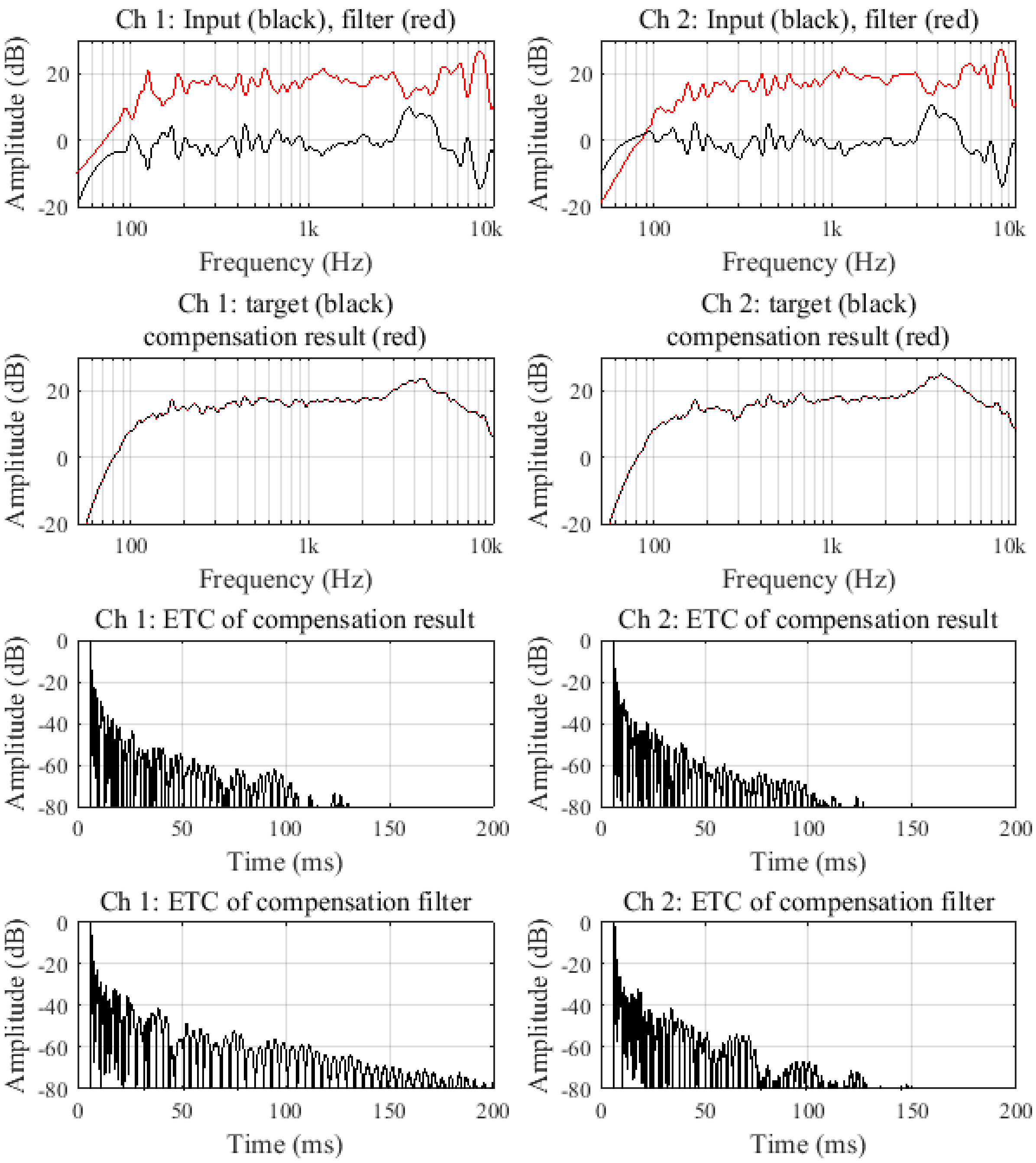


Figure 4.9. Computed compensation filter results. The top traces show the frequency-response curves for the complex-smoothed, averaged-input OBRIRs, and the inversion filters for left and right channels respectively. The second traces from the top overlay the target function (frequency response of the target room OBRIR) and the compensation result: the multiplication of the input and filter of the top trace. Note that on the decibel scale, addition is preferable to multiplication on the linear scale. The third traces from the top show the inverse Fourier transform of the compensation result, also plotted on a decibel scale. Here, ETC stands for energy time curve. The associated impulse responses should behave similarly to the impulse responses of the target OBRIR. The bottom trace shows the time-domain ETC of the compensation filter.

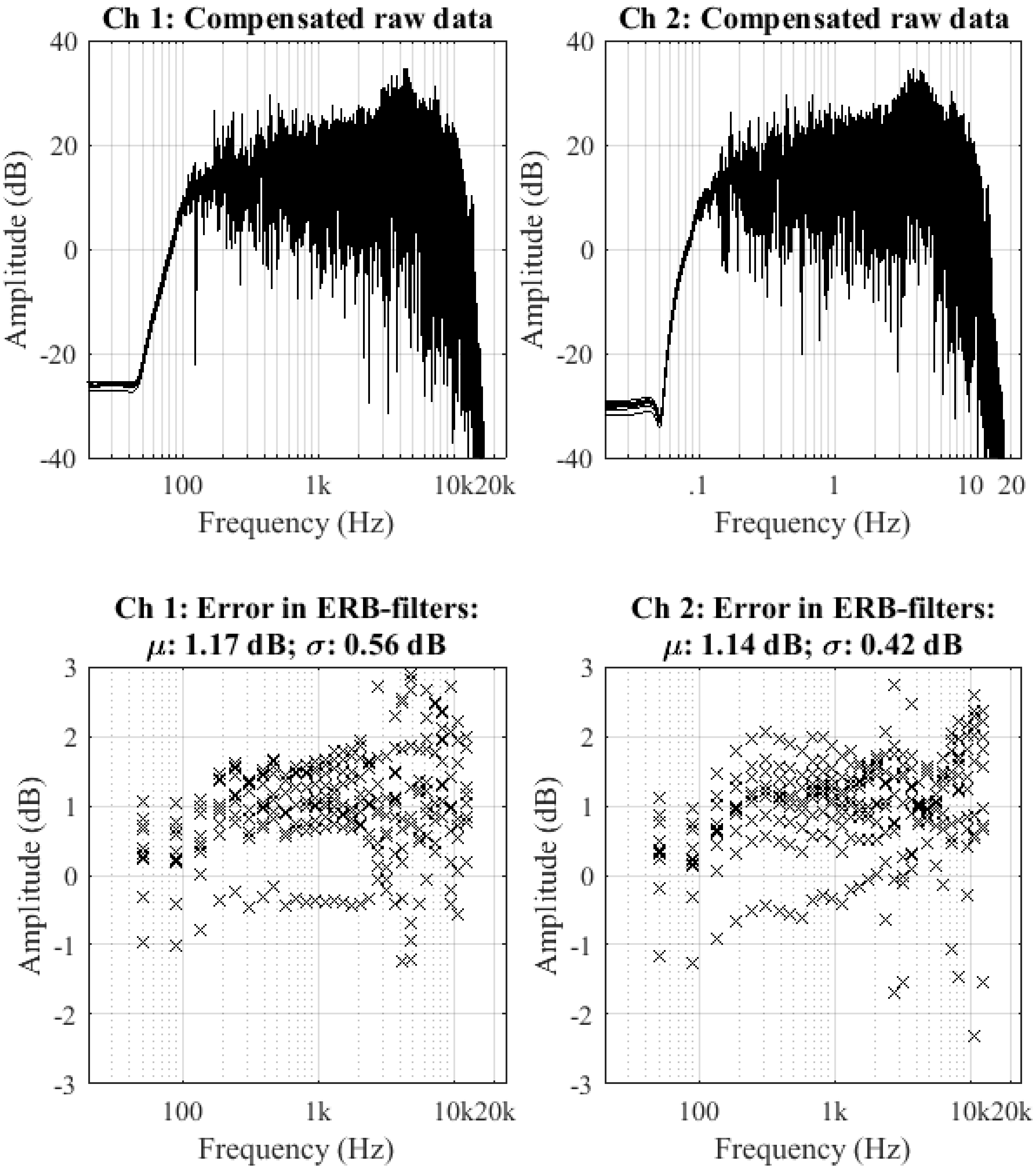


Figure 4.10. Simulation results for inversion filter computations for the reverberation chamber with two absorptive wedges. The top traces show the simulation frequency-domain results of convolving the inversion filter with each of the ten input OBRIRs, with left and right channels shown on the left and right sides of the figure. The results should be similar to the target function in Fig. 4.9, but because the input functions were not smoothed, more variation occurs shown over frequency. The bottom trace marks the errors when comparing the compensated raw data to the target function in the ERB filter bank. The mean decibel error and mean standard deviation in the error across the 10 simulations are reported.

Chapter 5

RTCS Validation

5.1 Introduction

The RTCS underwent objective and subjective tests to verify that it faithfully reproduced the effects of the OBRIR in use. Time and frequency-domain comparisons of OBRIR measurements from actual rooms and those from the RTCS were used as objective measures. Speaking-listening tests were designed and used for subjective measures. The former showed that the compensation filter computed in Ch. 4 improved the performance of the RTCS to more closely match the OBRIR it was representing. The speaking-listening tests indicated that the users considered the RTCS to be more realistic than not.

5.2 Objective Evaluation: Measurements of the Filtered RTCS

An evaluation of the RTCS performance with both the room OBRIR and the compensation filter was performed. The filter and the modified RIRs for each room condition were loaded into separate SIR2 plugins in separate tracks of the Reaper project file. A new series of OBRIR measurements was taken with both tracks enabled. This allowed the amplitude of each to be controlled separately. The advantage of this was that the level of each could be adjusted until the error between the original RIR and the new OBRIR measurements was minimized.

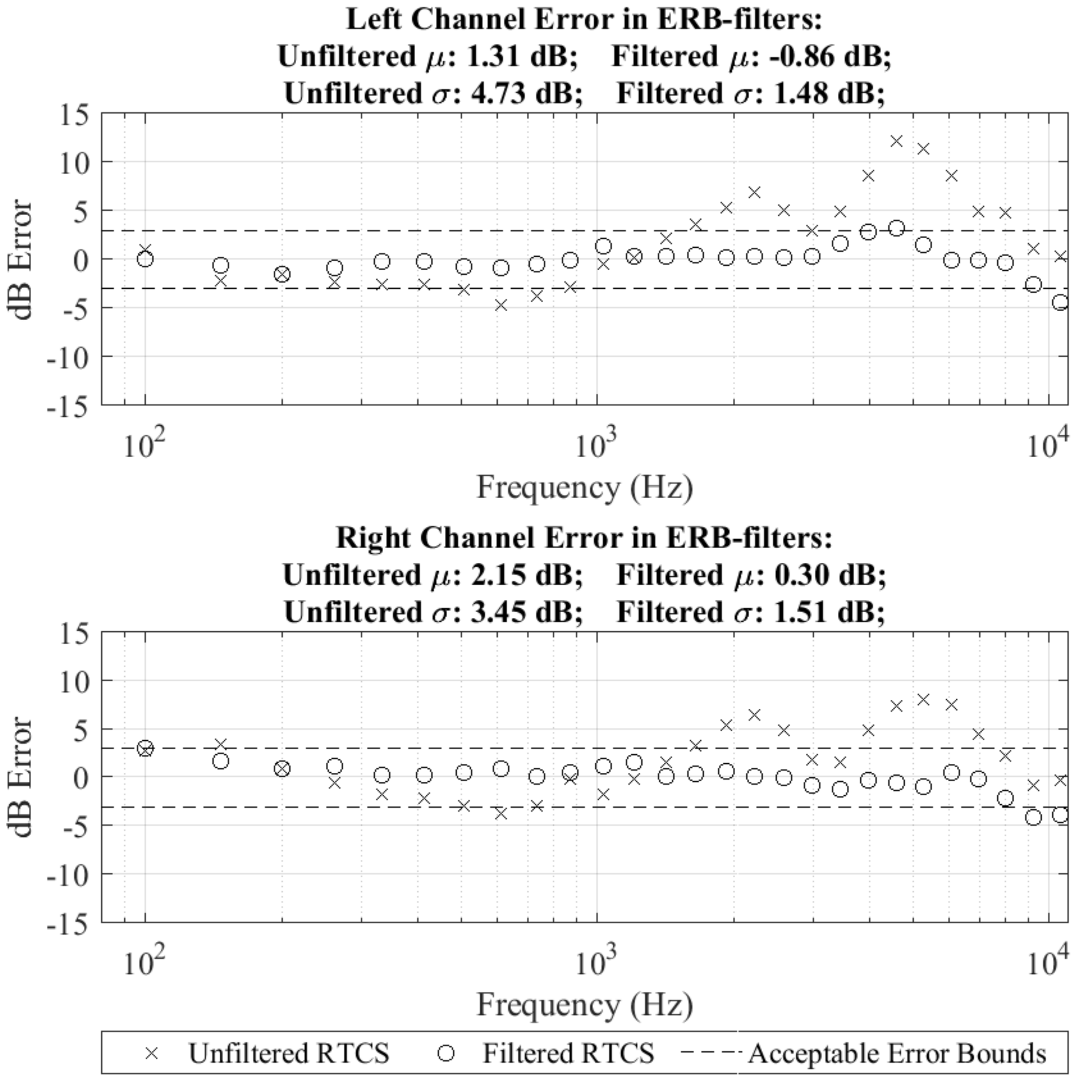
The track level of the modified RIR was previously set such that the level of the unfiltered RTCS OBRIR measurement matched that of the original RIR as closely as possible. The track level of the compensation filter was adjusted so that the level of the new (compensated RTCS) OBRIR measurements matched that of the original RIR and minimized the error in the frequency domain. This was done with the idea that the objective measures would then assess the optimal performance of the RTCS. The IR measurements of the filtered RTCS were carried out after the method described in Sec. 4.2.3.2.1 to be consistent with the measurements of the unfiltered RTCS OBRIRs and original room OBRIRs.

As with the simulation described in Sec. 4.2.4.2, the mean and standard deviation of the difference between the compensated RTCS OBRIR and the original room OBRIR was determined across the bands of the ERB filter banks. These error calculations were carried out on the modified OBRIRs to remove the effects of the KEMAR mouth simulator, which were dominant, and compute the error only on the performance of the RTCS itself. In addition, the energetic average of the level of the 1-ms moving average rms for each of the OBRIRs was computed and reported in dB. These measures in both the time- and frequency-domains provided a check whereby comparisons to the original room OBRIRs could be made.

A unique compensation filter was created for each OBRIR implemented in the RTCS, which consistently improved its performance. The results of the objective measures for all rooms are shown in Appendix E. By way of clarification, those for the reverberation chamber containing 32 wedges are explained here. Both the frequency-domain and time-domain error measurements for this room showed that the RTCS with the compensation filter was a closer match to the original OBRIR of the reverberation chamber

The OBRIR for this room condition lasted 1.2s. The OBRIRs measured with the RTCS implementing the 32-wedge reverberation chamber OBRIR likewise lasted 1.2s, but the frequency response was dissimilar to that of the original OBRIR. Thus, as mentioned in Ch. 4, the compensation filter used the modified reverberation chamber OBRIR as the target function and took a complex average of ten OBRIRs of the unfiltered RTCS implementing only the reverberation chamber OBRIR as its input data. (The ten OBRIRs followed headphone repositioning for each of the 10 measurements.) The filter length in this case was 2^{16} samples long, or 1.36 s. In SIR2, the reverberation chamber OBRIR was set to a level of -10.4 dB and the compensation filter was set to a level of -5.3 dB. The room OBRIR level was the same as in the unfiltered-RTCS OBRIR measurements. The filter track level was set using a trial-and-error approach to adjust the level over the course of five or so measurements and computed the frequency-domain level error between the newly compensated RTCS OBRIR measurement and the original room OBRIR. After investigating the results of these error computations, the setting that resulted in minimal frequency-domain error was selected and used for ten additional OBRIRs of the RTCS with headphone replacement.

An example for the frequency-domain error measurements is shown in Fig. 5.1. The error between the modified initial RTCS OBRIRs and the modified OBRIR of the reverberation chamber with 32 wedges are shown as “x” symbols, and the frequency-domain error measurements between the modified compensated RTCS OBRIRs and the modified room OBRIR are shown as “o” symbols. The filtered RTCS mean error was almost completely contained within the acceptable error bounds of ± 3 dB on both channels, except for the extreme ends of the frequency range. The mean-level errors and standard deviations also indicated that



Left Channel Error in ERB-filters:
Unfiltered μ: 1.31 dB; Filtered μ: -0.86 dB;
Unfiltered σ: 4.73 dB; Filtered σ: 1.48 dB;

Right Channel Error in ERB-filters:
Unfiltered μ: 2.15 dB; Filtered μ: 0.30 dB;
Unfiltered σ: 3.45 dB; Filtered σ: 1.51 dB;

Figure 5.1. Frequency-domain error results for RTCS representing Reverberation Chamber with 32 wedges for the left (upper) and right (lower) channels using the unfiltered (x) and filtered (o) RTCS OBRIRs.

the filtered RTCS OBRIR was a closer match to the original OBRIR of the reverberation chamber with 32 wedges.

The error bounds of ± 3 dB were chosen as visual guides to the “flatness” of the error measurement curve. While a loudness JND is frequency dependent [109], these uniform error bounds provided a constraint to the difference in amplitude across the frequency band of interest. Three decibels corresponds to a doubling or halving of sound intensity, but the ear-brain sensitivity follows the logarithmic decibel scale more than the linear intensity scale [105].

Accordingly, the deviations outside these error bounds correspond to large, audible, and noticeable differences between the original room OBRIR and the RTCS OBRIR. The errors inside the three-decibel bounds should correspond to a more realistic representation of the original room OBRIR.

As another indicator of improvement, the standard deviation of the filtered RTCS OBRIR error was closer to zero than for the unfiltered case. This metric was also reported in Fig. 5.1. The smaller standard deviation value shows that the errors were more consistent over the entire frequency band, especially since there were no large deviations.

Another measure of the performance of the filtered RTCS in the form of time-domain error measurements is shown in Fig. 5.2. Log-scale representations of each OBRIR of interest: the modified original room, the modified initial RTCS (used to generate compensation filter), and the modified filtered RTCS are shown. As a reminder, each OBRIR was modified to remove the effects of the KEMAR mouth simulator used in making the measurements, so as leave only the part of the OBRIR indicative of the RTCS performance. Each of the modified OBRIRs are presented on a log scale using a moving-average RMS window with a length of 1 ms. Only the first 500 ms are plotted. The mean level error between each of the RTCS OBRIRs and the room OBRIR was computed and reported in Fig. 5.2. The mean-level errors for the filtered RTCS are much closer to zero than for the unfiltered RTCS. A visual inspection of the modified OBRIRs also shows that while imperfect, the filtered RTCS OBRIR more closely approximates the room OBRIR than does the unfiltered RTCS OBRIR.

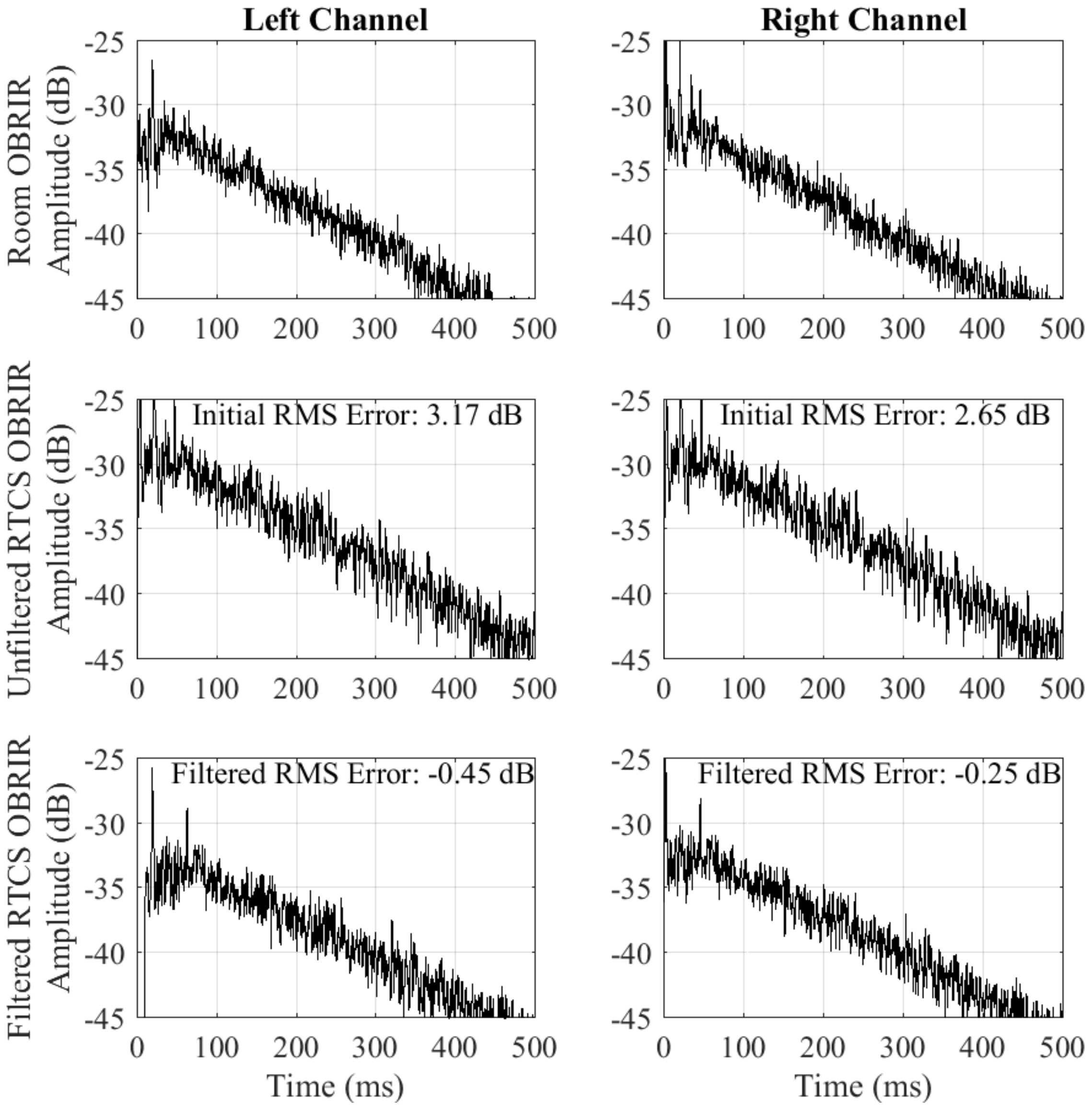


Figure 5.2. Time-domain error results for RTCS representing Reverberation Chamber with 32 wedges for left and right channels comparing room OBRIR (upper), unfiltered (middle), and filtered (lower) RTCS OBRIR.

These objective measures in the time and frequency domains show two perspectives of the differences between RTCS OBRIRs and the original room OBRIR (Figs. 5.1 and 5.2, respectively). They indicate that the compensation filter improved the performance of the RTCS, enabling it to more closely match the room OBRIR. This should correspond to improved subjective realism (introduced in Sec. 5.3).

The additional rooms presented in Appendix E had similar results, in that the inclusion of the compensation filter improved the RTCS OBRIR to more closely correspond to the original room OBRIRs in both the time and frequency domains. The room OBRIRs based on simulations required special considerations for level adjustments, as they did not include calibrated pressure data and were based on normalized *.wav files.

As a further validation, the room gain, BDT30, and DRDSR (introduced in Section 3.6) were computed for each of the acoustical conditions presented via the RTCS and compared to those from the original OBRIRs. These computations were performed pre-modification, so that the KEMAR effects were still included. The results are summarized in Figs. 5.3 to 5.5.

To determine how large the discrepancies were between the RTCS results and those of the original room OBRIRs, a percent error between the two was computed for each case. Several of the reverberation room conditions had percent errors smaller than 10%. Because the values for room gain for the de Jong Concert Hall and C215 were much smaller than those of the reverberation chamber, the percent errors appear much larger.

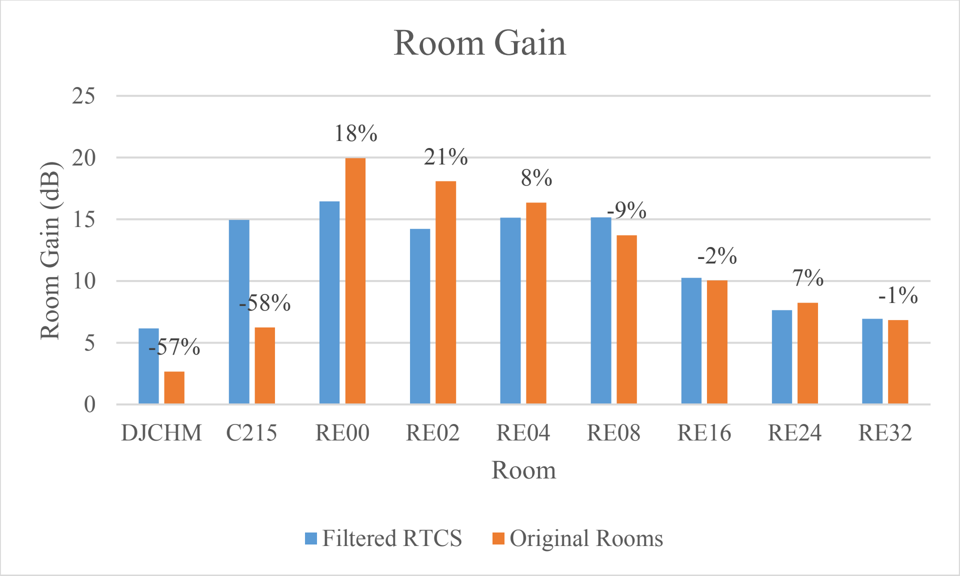


Figure 5.3. Percent Error in room gain between filtered RTCS OBRIRs and original room OBRIRs.

Figure 5.4 shows the differences in the BDT30s. The agreement between reverberation chamber conditions with added absorption is apparent, but conditions with longer decay times show larger discrepancies. Because the KEMAR effects had been removed in prior validation measures, some differences between the OBRIRs in the rooms and those based on the RTCS were not apparent as they are here. Of note is the fact that in the reverberation chamber OBRIRs, the KEMAR mannequin was positioned on a reflective plastic chair, but with stuffed pants that simulated talker legs. In the anechoic chamber, it was situated on a more absorptive lightly cushioned chair without pants. These differences may have produced differences in the earliest diffracted and reflected arrivals captured in the BDT30. Similarly, for the de Jong Concert Hall and C215 OBRIRs, KEMAR was seated in a different chair than in the anechoic chamber, or the reverberation chamber, and was without pants.

Interestingly, , the results for DRDSR show much better agreement between the room OBRIRs and the filtered RTCS OBRIRs (see Fig. 5.5). The values for the de Jong Concert Hall and reverberation chamber with 32 absorbing wedges have the largest percent errors, but those of

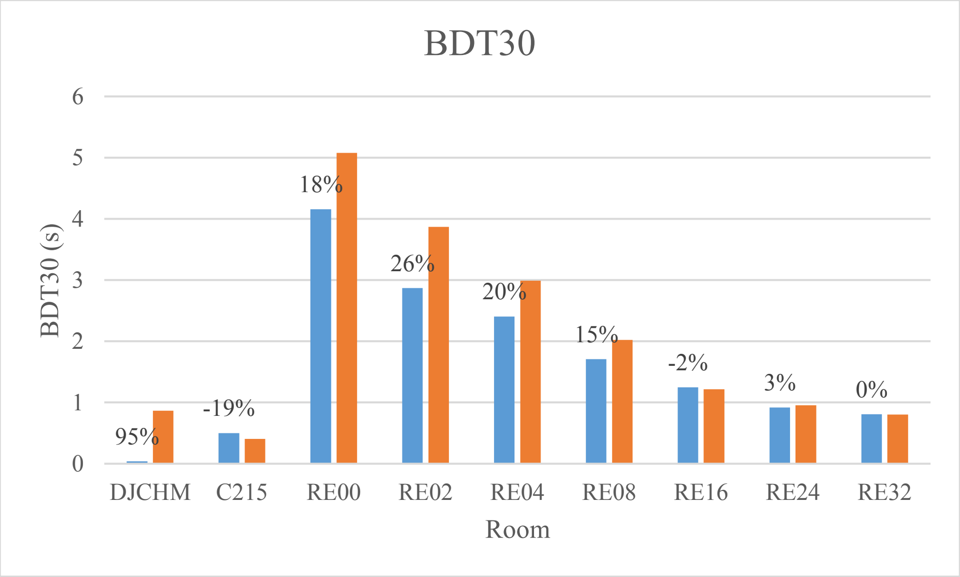


Figure 5.4. Percent error in BDT30 between filtered RTCS OBRIRs and original room OBRIRs.

the other conditions fall within 10 % error. One possible explanation for the apparent improvement is that this measure separates the diffracted and reflected sound portions of the OBRIR, whereas the other measures involve both portions together. Slight differences in the OBRIR measurement procedure, including the amplification provided to the KEMAR mouth simulator, may have affected the results of room gain and BDT30. The DRDSR, on the other hand, compares slopes of the room reflections to the earliest diffracted sound, so the differences between OBRIR measurements become less obvious. Since the RTCS was calibrated using modified OBRIRs with the diffracted sound removed, this measure appears to provide further validation that the reflections presented via the RTCS were similar to the reflections of the rooms relative to the diffracted sound of the OBRIR.

Figure 5.6 shows the first 100 ms of an OBRIR for the reverberation chamber with 24 wedges, its Schroeder curve, and the decay slopes for the diffracted and reflected sound portions of the OBRIR. By comparing it to Fig. 3.19, it becomes apparent that the levels of some of the

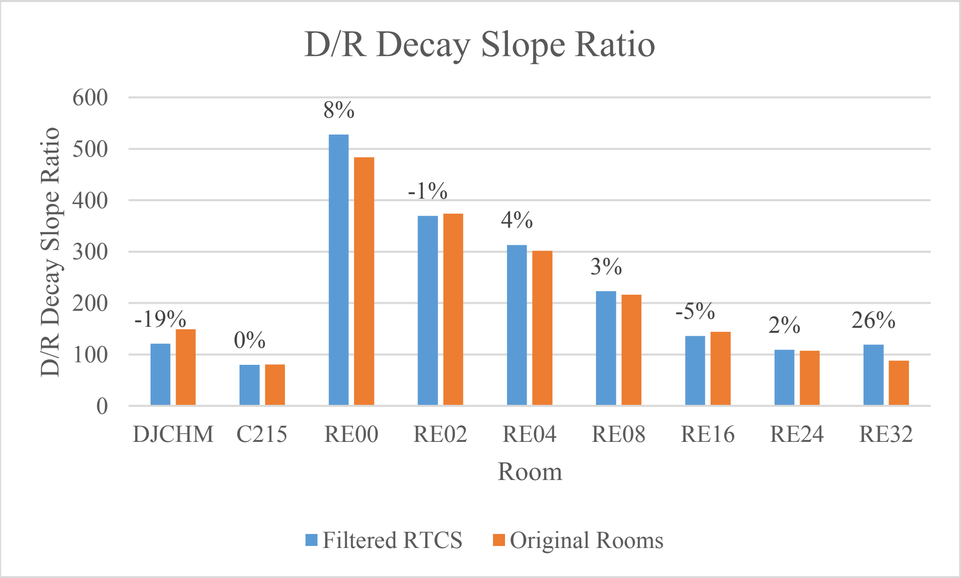


Figure 5.5. Percent error in DRDSR between filtered RTCS OBRIRs and original room OBRIRs.

prominent room reflections differ, which would cause a difference in room gain. The Schroeder curve is also slightly different for the room and the RTCS OBRIR, which would lead to a differing value for BDT30. Despite these differences, the ratio of diffracted decay slope to the reflected decay slope is quite similar. This means the DRDSR is very similar for the room and the RTCS OBRIRs.

To summarize, the RTCS was validated objectively in a number of ways. First, in investigation into the OBRIRs with the KEMAR effects removed compared the room representations in both the time and frequency domains. These measures show that the RTCS was calibrated to represent the rooms at the proper level, with the proper timing of reflection arrivals, and the proper frequency response. Secondly, the RTCS performance was evaluated with a number of new OBRIR characterization parameters. These parameters included the early

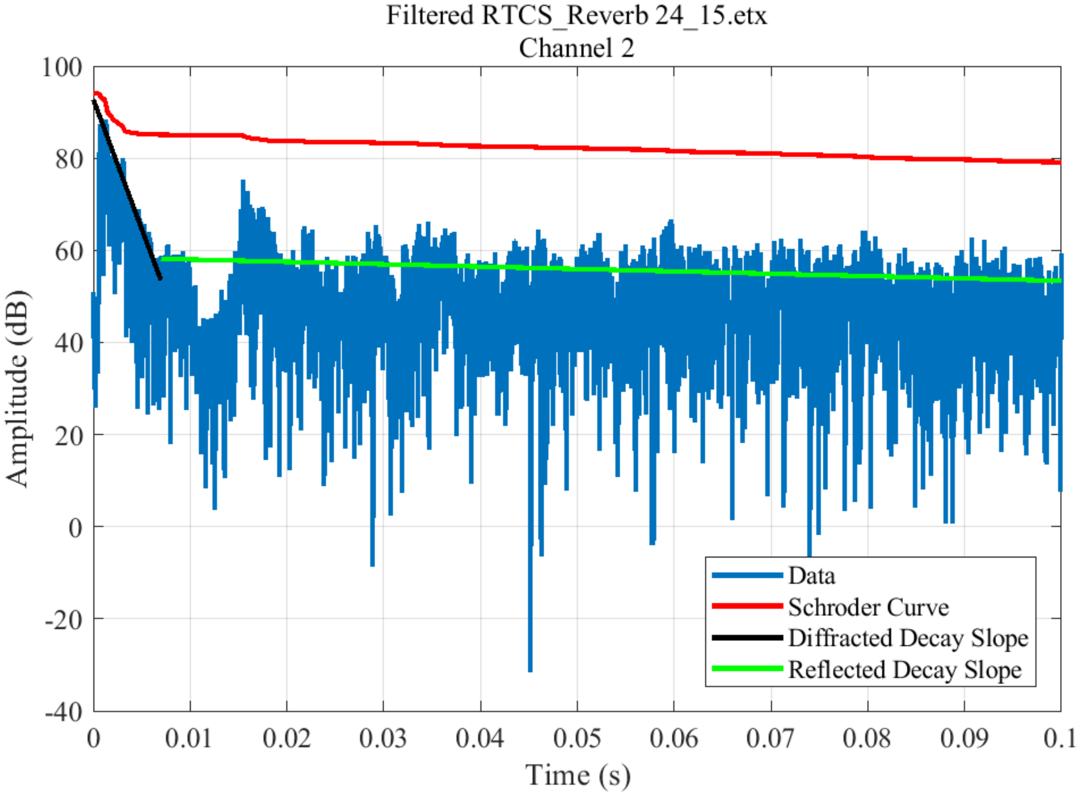


Figure 5.6. Compensated RTCS OBRIR measurement for the RTCS representing the reverberation chamber with 24 absorbing wedges, right channel only

diffracted sound and the reflected sound portions of the OBRIR. Additional study is needed to more fully understand these parameters' relationships to psychoacoustic evaluation, but they do indicate good agreement between the RTCS and the original room OBRIRs.

5.3 Subjective Evaluation: Listening and Speaking Tests

In addition to the objective measures, several subjective evaluations of the filtered RTCS were performed. These were done through a survey of untrained listeners who spent a brief time talking and listening to the simulated acoustic environments, and then rated the experience on its realism.

5.3.1 Methods

Thirty-two subjects were invited to participate in a vocal effort experiment utilizing the RTCS. In addition to providing speech and vocal effort data for the experiment (discussed in Ch. 6), the subjects answered questions describing their auditory experience while using the system. Both they and the research interviewers were blind to the acoustic environment being simulated. Each subject had about four minutes to speak and listen in each simulated environment prior to answering questions. He or she then rated, on a scale of 1 to 7, how natural or realistic the condition sounded and how believable it seemed that he or she was actually in another room or place. Each subject subsequently identified a room or place the condition most sounded like (free response, no prompt) and described the features of the condition that contributed most to unbelievability. The free response answers were categorized post-experiment, and the results were analyzed using statistical tests.

5.3.2 Results

To assist in explaining the results, abbreviations for the ten simulated room conditions used during the subjective testing are given in Table 5.1. These abbreviations were also used in the vocal effort study to identify acoustic conditions during the randomization procedure. The subjects only experienced using the RTCS with the appropriate compensation filters. They did not use the RTCS without a compensation filter for any condition. They also did not visit the rooms the RTCS was simulating. Listener responses to subjective evaluation questions are now presented.

Table 5.1. Acoustic Conditions Abbreviations

Abbreviation	Condition
RE00M	Measured OBRIR of the reverberation chamber with zero absorptive wedges.
RE02M	Measured OBRIR of the reverberation chamber with two absorptive wedges.
RE04M	Measured OBRIR of the reverberation chamber with four absorptive wedges.
RE08M	Measured OBRIR of the reverberation chamber with eight absorptive wedges.
RE16M	Measured OBRIR of the reverberation chamber with 16 absorptive wedges.
RE24M	Measured OBRIR of the reverberation chamber with 24 absorptive wedges.
RE32M	Measured OBRIR of the reverberation chamber with 32 absorptive wedges.
C215M	Measured OBRIR of the classroom C215
DJCHM	Measured OBRIR of the de Jong Concert Hall
DJCHS	Simulated OBRIR of the de Jong Concert Hall

5.3.2.1 *Realistic Rating*

The realistic ratings are the subjects' responses to the question "How realistic or natural is this acoustic space on a scale from one (obviously digitally synthesized) to seven (very realistic, agrees with every day experience)?" Figure 5.7 is a box-and-whisker plot summarizing the realistic ratings for the ten simulated room conditions. The interquartile range is shown as the box portion, and the outlying data as the whisker portion. The "x" marks the mean of the data and the horizontal bar in the box represents the median. All rooms had an upper rating of 7 (completely realistic and natural sounding). Four rooms, C215M, RE02M, RE08M, and RE16M, had a lowest rating of 1, while the others had a lowest rating of 2. All but two of the rooms had a first quartile above 3.5, with RE00M, and RE08M having lower first quartiles. Since the rating

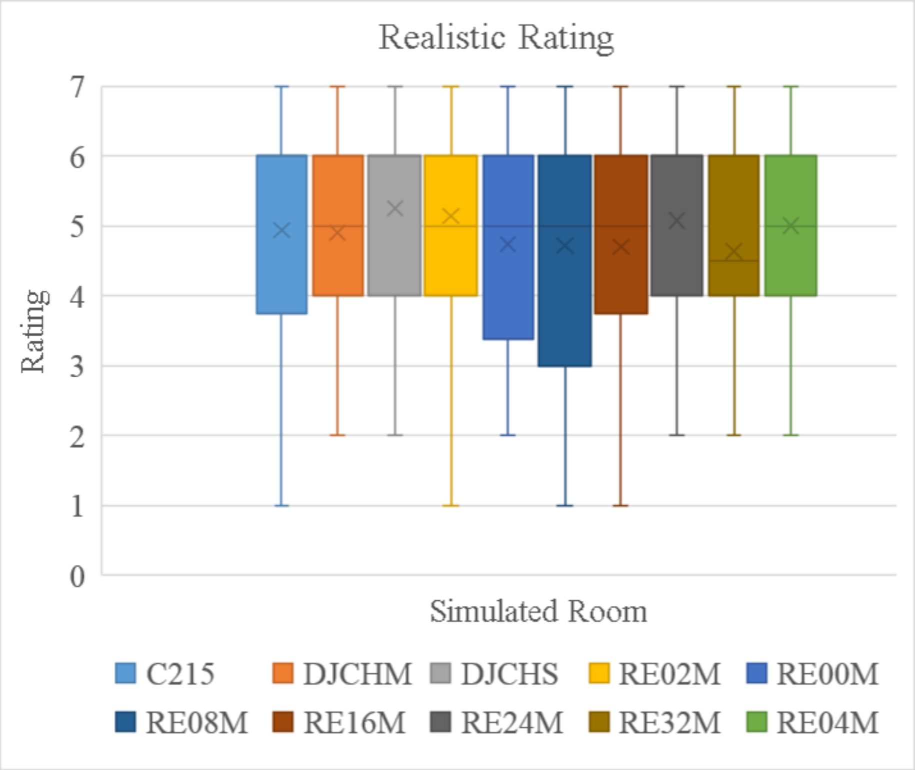


Figure 5.7. Realism ratings for the ten simulated room conditions.

scale extended from 0 to 7, the results indicated that all the rooms except RE00M and RE08M were rated as “more realistic” more than 75% of the time.

5.3.2.2 *There Rating*

The subjects were prompted to describe their suspension of disbelief with the prompt “If you closed your eyes, can you imagine yourself in this space? Tell me the degree to which you can believe you’re there between one (not at all there) and seven (definitely there).” Figure 5.8 is a box-and-whisker plot summarizing the ratings of how well the participants believed they were in another room or place for the ten simulated room conditions, despite being blind to the acoustic condition being simulated. All rooms except RE32M had a first quartile above 3.5, indicating that they could believe they “were ‘there’ more than not,” at least 75% of the time. Part of the low rating for RE32M followed from the difficulty participants had in picturing

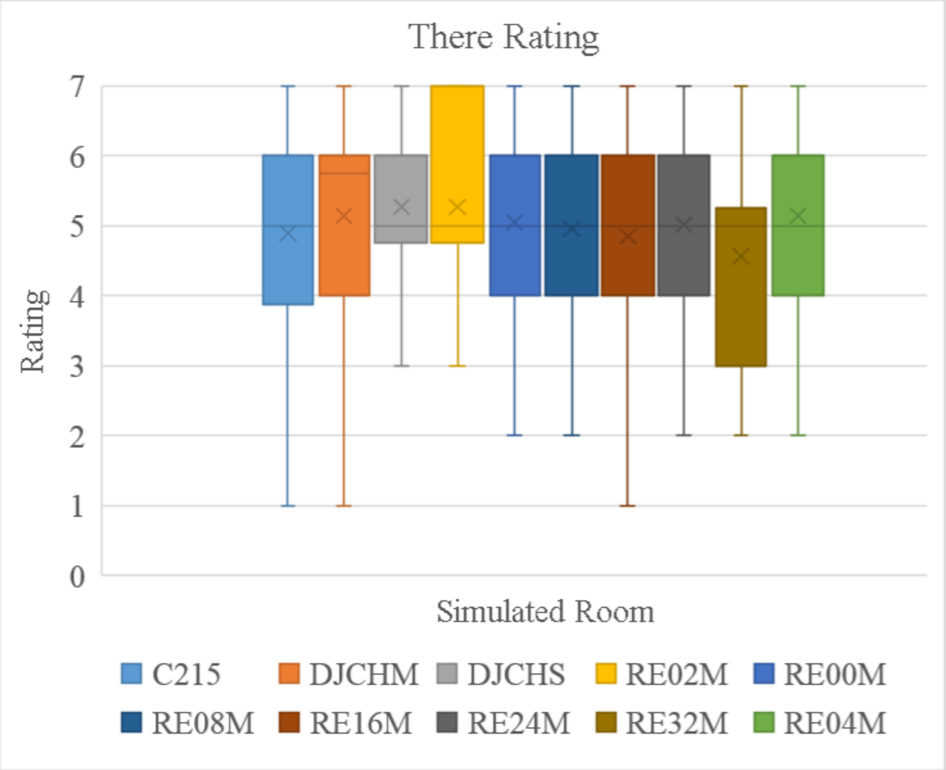


Figure 5.8. Ratings of perception that subjects were in another room or place for the ten simulated acoustic conditions.

themselves in an unusual acoustic space—a well-damped reverberation chamber—that they could not see. The other rooms may have been perceived as more traditional or agreeable with the participants' common experiences. Room RE02M especially had a third quartile rating of seven, which was the highest possible.

5.3.2.3 *Identification of Acoustic Condition*

The subjects were asked to freely identify the acoustic space they experienced. Figures 5.9 through 5.11 show the distribution of places participants identified for each simulated acoustic condition. Clear identification of another acoustic environment may be associated with a high perceptual rating of perceiving that one is actually in another space. Future work may include picture representations of simulated rooms to aid the participants in identifying the acoustic space. Room RE02M was most commonly described as a cave or a reverberation chamber. In contrast, room C215M was most often described as a classroom or a generic room. The distributions are less clear for the highly damped reverberant conditions, such as RE24M and RE32M. These conditions do not have a clear majority for their common identifiers. For readability, the abbreviations for Court as Ct., Cathedral as Cthl., Classroom as Clsm., Bathroom as Bthm., and Auditorium as Aud. are used in the plots. The category "Other" was used for identifiers that did not appear more than once from multiple participants. Future work may have a set list of identifiers for the acoustic conditions, but for this work, it was interesting to learn what the participants could identify without any prompts.

5.3.2.4 *Unrealistic Characteristics*

The participants' responses to what made the simulated acoustic conditions sound most unrealistic are included as an indicator of where the RTCS may need to improve in future

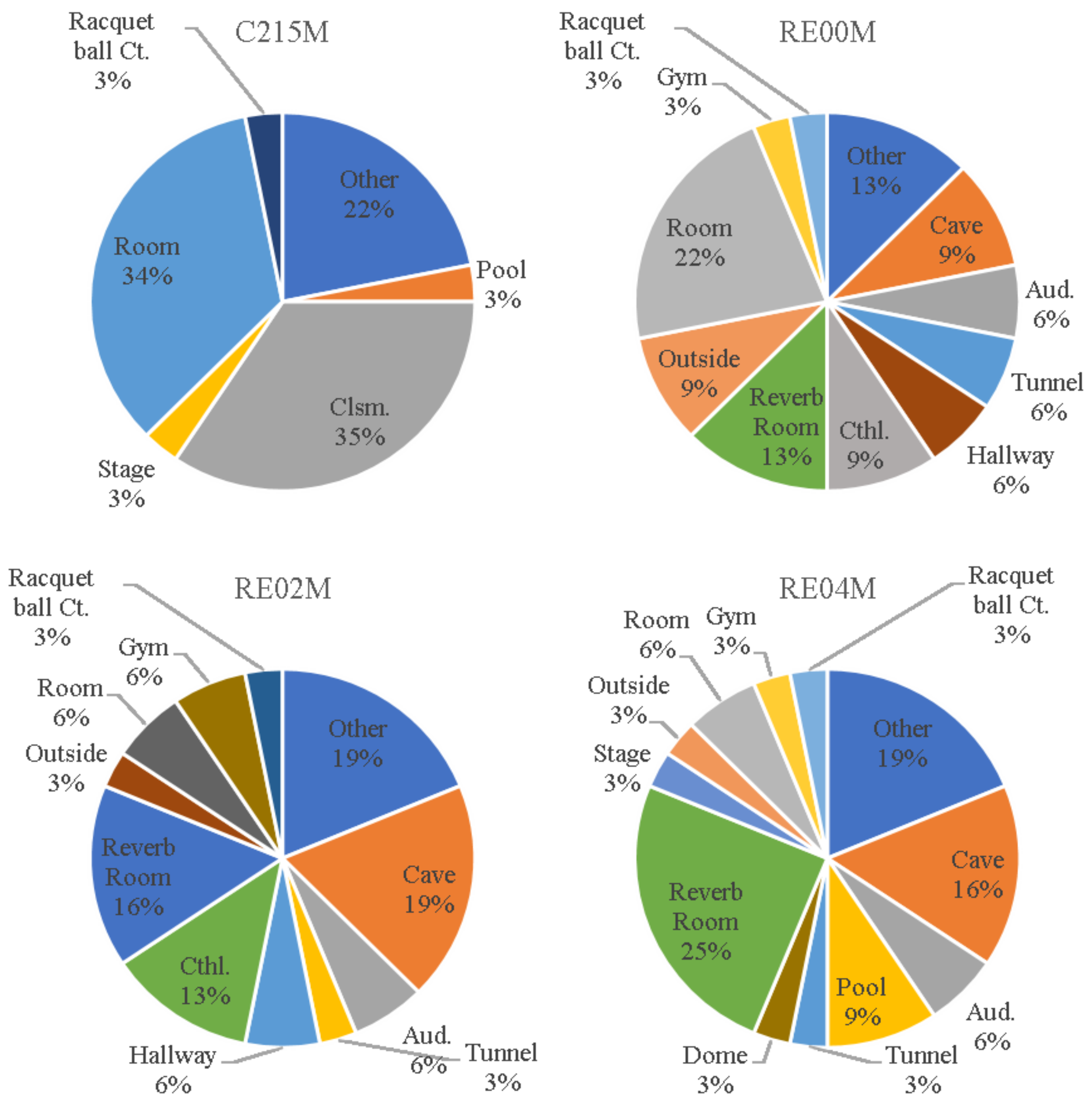


Figure 5.9. Distribution of common identifiers for simulated acoustic conditions C215M, RE00M, RE02M, and RE04M. Room C215 is overwhelmingly identified as a classroom (clsm) or a generic room. The reverberation chamber simulations were more often identified as a cave or as a reverberation chamber, but the majority is not as strong. This could be due to the reverberation chamber being an unnatural acoustic condition that most people do not commonly experience.

studies. However, subjects responded "nothing" a large portion of the time. Some of the rooms had issues that were mentioned more commonly than others were. One common response was that the lack of visual correlation with the acoustics made the simulation seem unrealistic or unbelievable. The participants rated RE00M and RE04M as having issues with the timing of the

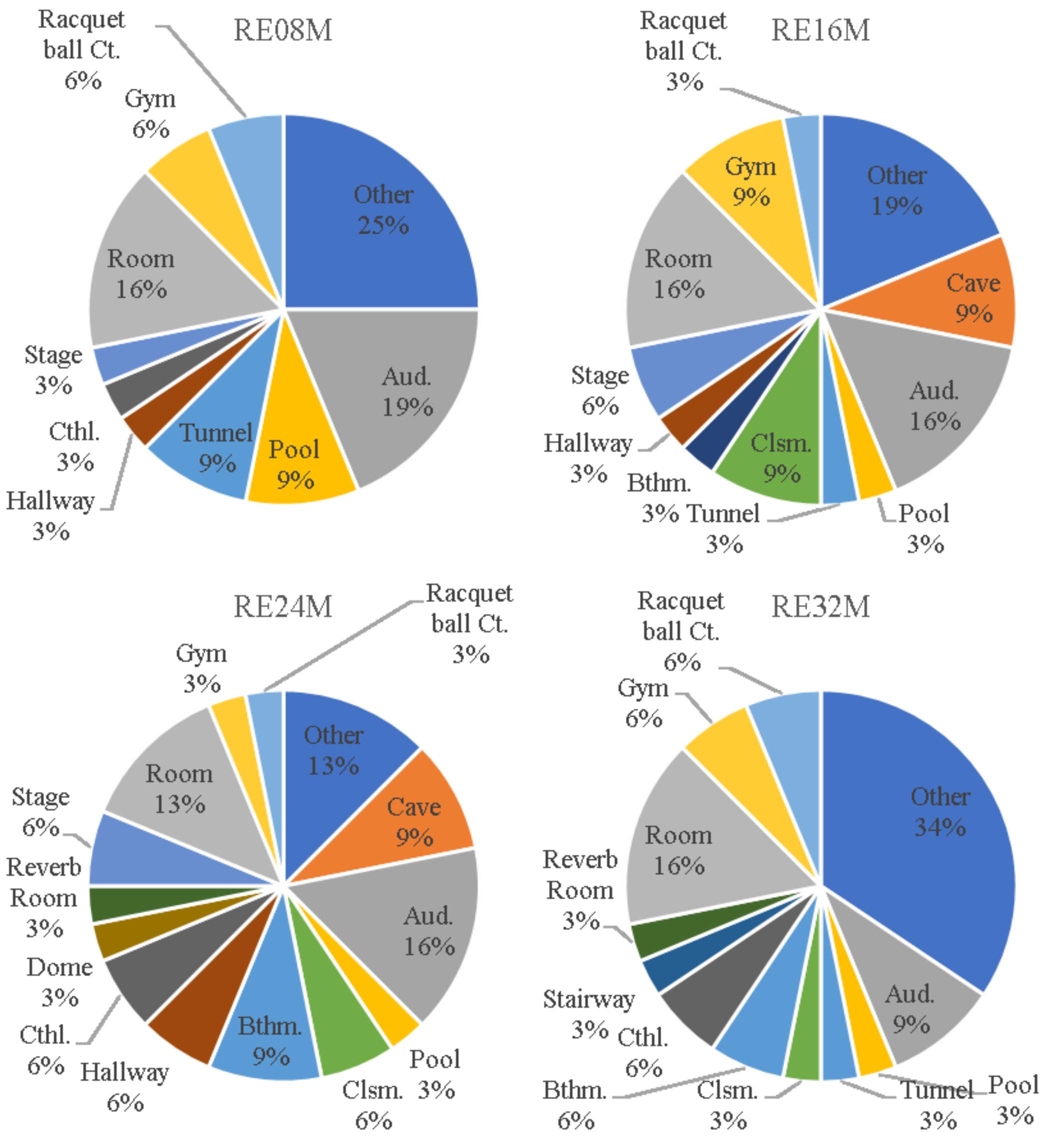


Figure 5.10. Distribution of most common identifiers for simulated acoustic conditions RE08M, RE16M, RE24M, and RE32M. These damped reverberation chamber conditions seemed to be more difficult to identify, as shown by the multiple identifiers for each room.

reverberation, either too early or too late. The de Jong Concert Hall conditions DJCHM and DJCHS were rated as having issues with the level of the reverberation, as some participants felt that the simulation was too quiet. This may have been due to the conditioning produced by the reverberation chamber simulations. In those cases, the onset of reverberation was strong and almost immediate. In the concert hall simulations, the reflective walls were relatively distant,

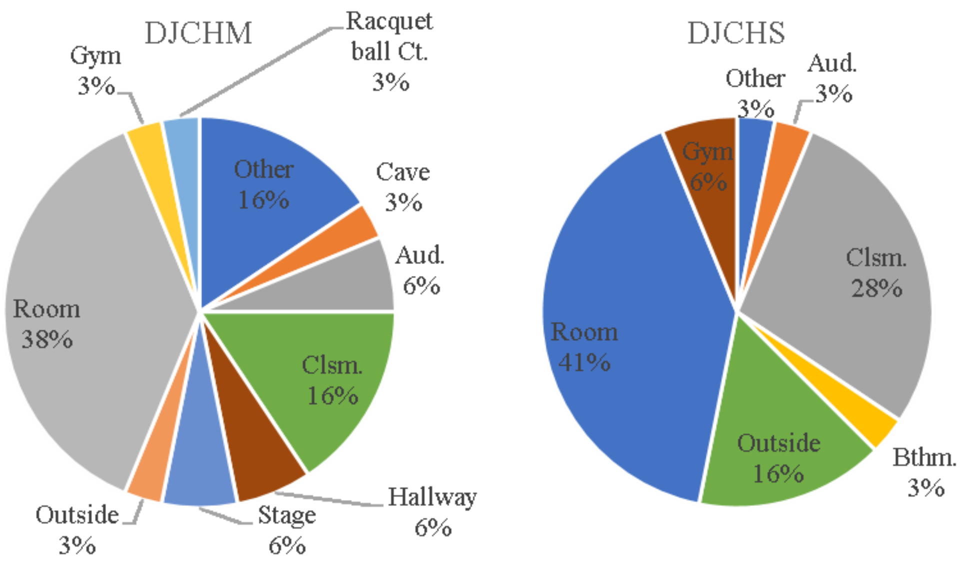


Figure 5.11. Distribution of most common identifiers for simulated acoustic conditions DJCHM and DJCHS. These conditions were more commonly identified as a room, due to the lack of perceived reverberation (for collocated source and receivers and delayed reflections from distant walls) as compared to the reverberation chamber simulations.

meaning the collocation of source (mouth) and receivers (ears) would lead to stronger “direct-to-reverberant” sound ratios. Additional comments that did not appear commonly were marked as “Other.” Figures 5.12 through 5.14 summarize the room characteristics that were considered unrealistic.

5.4 Discussion and Conclusions

The objective and subjective evaluations of the RTCS performance show that the simulated acoustical conditions were realistic and believable to the RTCS users. The objective measures show that the inclusion of a compensation filter brought the RTCS simulations closer to the original room conditions. The subjective measures show that RTCS users could identify the simulated acoustic conditions as an actual place, believe they were there more often than not, and rate it as more realistic than not most of the time. When asked about what made the simulated acoustic condition unrealistic, the most common answers were "nothing" or "a lack of visual correlation to the acoustic experience," although some of the rooms also had additional issues that were pointed out. The combination of looking at objective and subjective measurements of the RTCS performance provides both a quantitative and a qualitative evaluation of the RTCS with the inclusion of a compensation filter.

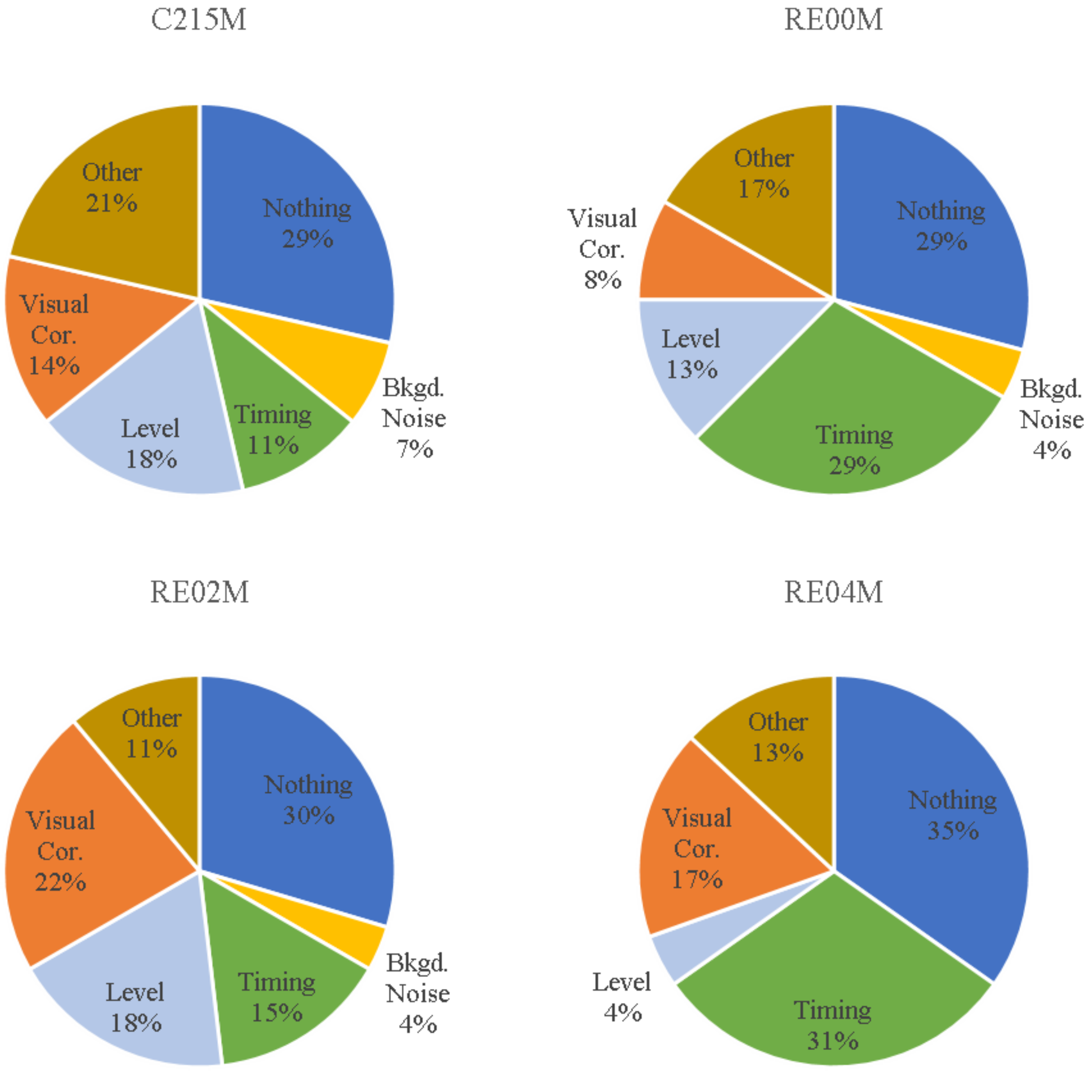


Figure 5.12. Unrealistic characteristics for simulated acoustic conditions C215M, RE00M, RE02M, and RE04M.

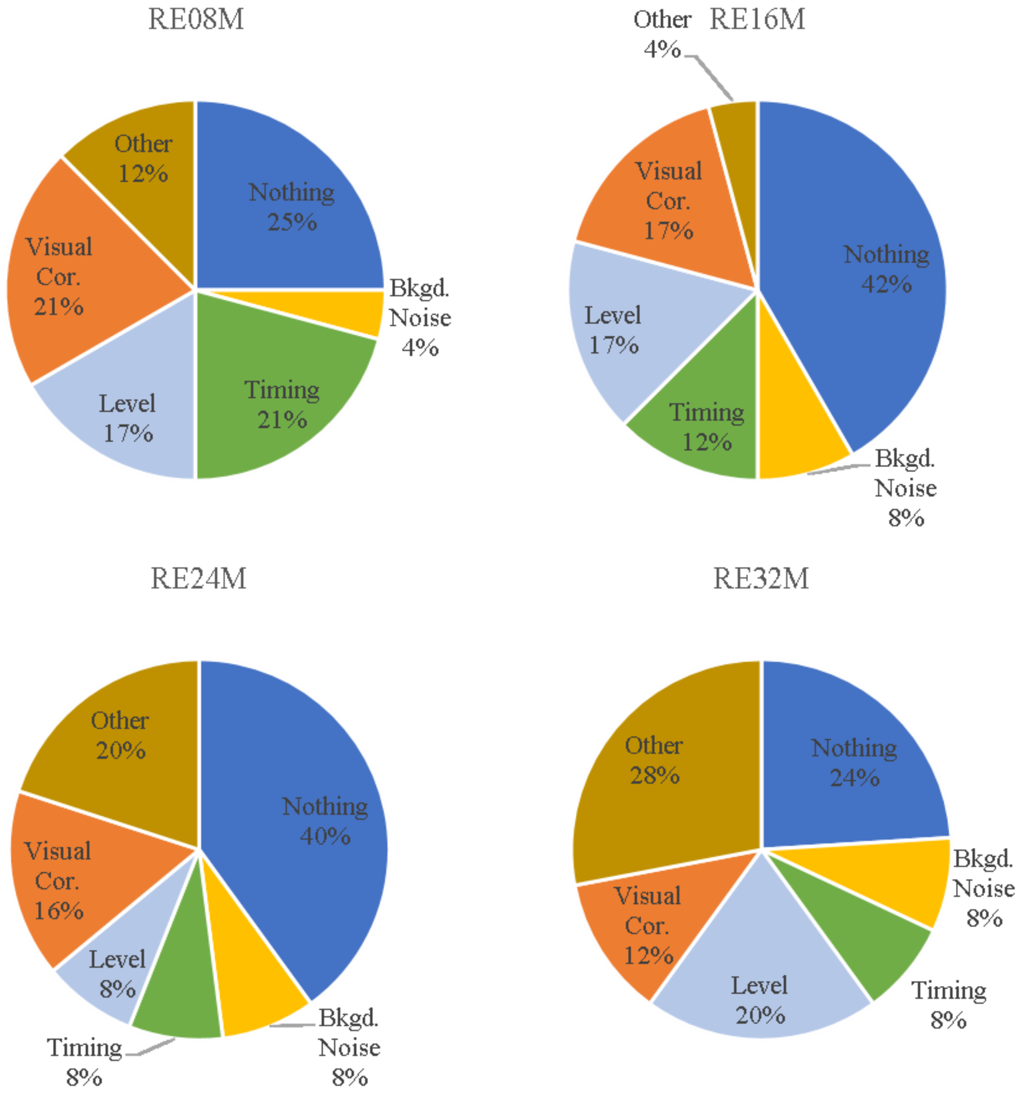


Figure 5.13. Unrealistic characteristics for simulated acoustic conditions RE08M, RE16M, RE24M, and RE32M.

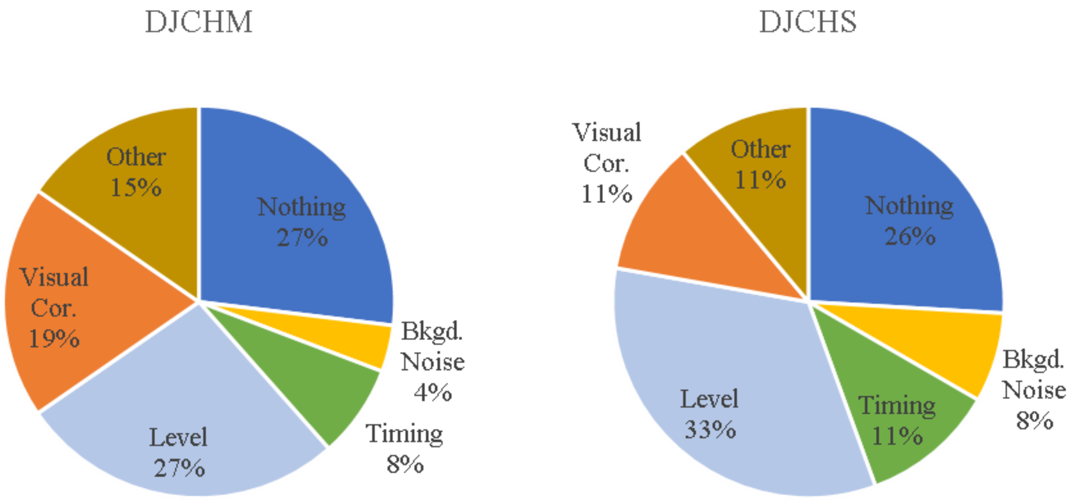


Figure 5.14. Unrealistic characteristics for DJCHM and DJCHS.

Chapter 6

Vocal Effort Study using RTCS

6.1 Introduction

This work was primarily motivated by the need to study vocal effort [23-36]. As indicated earlier, the RTCS was utilized to present virtual-acoustic environments to talkers. The study was conducted in a fashion similar to a recent vocal effort study performed at BYU in a reverberation chamber with varying amounts of added absorption [37].

Earlier studies have shown that talkers adjust their voices according to their acoustical environments [110,111]. Some researchers have probed the effects of basic room-acoustic properties [110,112-118], including voice support, reverberation time, and noise level, on simple vocal measures such as talker voice level [32], speech rate [119], and dose [26]. Since the effects of other important room-acoustic properties and more comprehensive effort-related vocal measures have not been reported, the present work has sought to help remedy the deficiency.

The methods for conducting the study using the RTCS are outlined in Sec. 6.2. Section 6.3 provides an initial look at the results of the vocal effort study. Section 6.4 compares similar data from this vocal effort study to the prior study. Section 6.5 discusses the results, and Sec. 6.6 provides some concluding remarks.

6.2 Methods

6.2.1 Conditions

The room conditions presented via the RTCS earphones were described in detail in Chs. 3 and 5. The amount of time allotted for each subject to complete the vocal effort experiment prevented the use of all prepared conditions. Those used in this study were (as listed in Table 5.1) the seven measured OBRIRs of the large BYU reverberation chamber with varying amounts of absorption. These were denoted by RE00M RE02M, RE04M, RE08M, RE16M, RE24M, and RE32M. Others included a measured OBRIR of the lecture hall C215 in the BYU Eyring Science Center, denoted by C215M; a measured OBRIR from the de Jong Concert Hall in the BYU Harris Fine Arts Center, denoted by DJCHM; and a simulated OBRIR of the de Jong Concert Hall, denoted by DJCHS. These conditions were presented to each subject in a random order, with randomization performed using the Microsoft Excel rand() function. A table of the randomizations is presented in Appendix F. The final condition or trial for each subject was the anechoic chamber itself. This condition was presented by disengaging the RTCS, but allowing the subject to continue to wear the microphones and earphones. The subjects were not warned beforehand that the final trial was unique or meant to be a control condition.

6.2.2 Speech Elicitation

The subjects completed three speech tasks while experiencing each room condition. The first was a reading of the phonetically-balanced Rainbow Passage, first paragraph [120]. This has been used in many speech studies, making data from this experiment comparable to other speech studies. The subject was instructed to read the passage using a conversational, clear tone. The subject then sustained the vowel /ɑ/ at a natural speaking pitch for five seconds, with three repetitions. The final task involved spontaneous speech. The subject was asked to describe a picture from a set of Diapix pictures for 45 seconds [121].

After these tasks, the subject rated his or her vocal effort and fatigue on a scale from 0 to 100, and predicted a level of vocal fatigue after speaking in the condition for twenty minutes on the same scale. The subject also rated the condition as described in Sec. 5.3. The signal to the earphones was then muted, and the subject was allowed some vocal and aural rest for about 30 seconds while the RTCS settings were updated for the next condition.

6.2.3 Recordings

The speech data gathered on the subjects came from recordings made using microphones and accelerometers on the subject. As indicated earlier, the subject wore a head-worn microphone positioned at the corner of his or her mouth. Its thin support arm was taped in place using medical tape. The subject also wore a Sonvox VoxLog collar that housed an accelerometer and microphone on his or her neck (see Fig. 4.4). The signals from each of these devices were routed to a PreSonus FireFace and recorded as *.wav files using Reaper with a sampling rate of 48 kHz and a depth of 32 bits.

6.2.4 Trimming

The recordings were saved using a file-naming protocol that retained information about the subject's gender and participant number, and the condition name and order in which it was presented to the subject. The full protocol is given in Appendix G. The files were trimmed using a MATLAB GUI developed by Mark Berardi to separate the speech tasks for analysis: (1) Rainbow Passage (RB), (2) Rainbow Passage sentences 2 and 3 (R2), (3) sustained vowel /ɑ/ (AH), and (4) spontaneous speech picture description (DE). The parenthetical letter codes were appended to the file names while retaining all the previous filename information.

6.2.5 Speech Analysis

The analysis on the trimmed recordings was performed using a MATLAB script that sent commands to the speech analysis program Praat. This program identified fundamental frequency, pitch strength, decibel level, alpha ratio, semitone standard deviation, activity factor, harmonics-to-noise ratio, shimmer, jitter, and the acoustic voice quality index (AVQI) [122,123] for each trimmed recording. A summary of the speech parameters computed for each speech task is given in Appendix H. A brief description of each of the speech parameters is given below.

- Fundamental frequency (F_0) is based on the period of the vocal folds vibrating. It was computed for the speech tasks RB and AH to provide a look at F_0 for a long-time average across many phonemes and at F_0 for a relatively short sustained vowel, respectively. Both the mean (mean) and the standard deviation (std) of F_0 across the speech tasks are reported. The unit of measurement for F_0 is Hz. F_0 is abbreviated F0 in the tables and figures that follow.
- Here, pitch strength is an objective measure that determines how salient the presence of pitch is. The mean pitch strength was determined via MATLAB implementation of Aud-SWIPE-P, based on the SWIPE' (sawtooth waveform inspired pitch estimator) script developed by Camacho [124]. Stronger pitch strength is associated with a clearer sense of tone, while speech with lower pitch strength is sometimes described as “breathy” or “airy.” The mean and standard deviation of pitch strength during the speech tasks RB and AH look at pitch strength for these two types of speech. Pitch strength is abbreviated Ps in the tables and figures that follow.
- Decibel level describes the intensity of the speech signal. The decibel level during the speech tasks RB, AH, and DE is reported by mean and standard deviation in units of dB. In addition, the mean decibel level normalized to the ANCH and C215 conditions for each subject is

reported. This normalizes the decibel level to make it more comparable across subjects. Decibel level is abbreviated dB in the tables and figures that follow.

- Activity factor (ActyFact) is the ratio of speech to silence during a speech task. It is reported for RB and DE. It is unitless and quantified between 0 and 1.
- Alpha Ratio (AlphaRto) is the ratio of the spectral energy in a speech signal above and below 1 kHz. It was computed for the speech task RB.
- The spectral slope from fundamental frequency (dBspcSpF) is reported for the speech task RB. This measure describes how rapidly the amplitudes of successive harmonic frequencies decrease as they get higher in frequency. It is commonly used as a measure of voice quality.
- Semitone standard deviation (STSD) is another measure of spectral deviations, but it is based on semitones, not fundamental frequency; it makes measures of males and females comparable. It is computed for the speech tasks RB and the extraction of two sentences from RB (R2).
- Syllable rate (syl_rate) is the division of a known number of syllables by the time it took the subject to pronounce those syllables (DurOTas). It is reported for the speech task R2, which had 29 syllables.
- Smoothed cepstral peak prominence (CCPS) analyzes the speech in the cepstral domain as a measure of dysphonia. A cepstrum of a speech signal is obtained by taking the inverse Fourier transform of the logarithm of the spectrum of the signal. Here CCPS is computed for a concatenation of the speech tasks R2 and AH. It is one of the components in computing AVQI.
- The acoustic voice quality index (AVQI) is a measure of dysphonia. It is computed from the concatenation of the speech tasks R2 and AH in order to incorporate the effects of running

speech and sustained vowel. According to Reynolds, a value greater than 3.5 is indicative of voice dysphonia [20,83].

- The sustained vowel also has a number of additional measures. Jitter is the deviation from true periodicity of a presumably periodic signal. This is the average absolute difference between consecutive periods of the sustained vowel, divided by the average period. Shimmer is the average absolute difference between the amplitudes of consecutive periods of the sustained vowel, divided by the average amplitude [84]. Harmonicity, or Harmonics-to-Noise Ratio (HNR), compares the energy of a speech signal that is periodic to the energy of the aperiodic, or noise part of a speech signal. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise, and the speech is dysphonic.

6.2.6 Statistical Analysis

The subject information, condition information, speech data, vocal effort ratings, and realism ratings were combined into a spreadsheet for statistical analysis. The independent variables of the statistical analysis are participant number, participant gender, trial number, and acoustic condition for each trial. The dependent variables for the statistical analysis are the speech parameters described in Sec. 6.2.5, and some self-reported vocal effort parameters. These are the subjects' self-reports of their levels of vocal effort (VE), vocal fatigue (VF), and predicted vocal fatigue if they had to keep talking in the acoustic condition for 20 minutes (VF20). Each of these parameters were scaled between 0 and 100.

The dependent variables were compared against the independent variables using mixed-design ANOVA tests. The vocal parameters significantly influenced by room-acoustics, gender, and trial number are presented in Sec. 6.3.

6.3 Results

The results of the mixed-design ANOVA tests are presented in Table 6.1. A p-value of 0.005 or smaller was used as the determining factor as to whether the vocal parameters were significantly influenced by the room acoustics, gender, and trial number.

Table 6.1. Statistical analysis results for speech measures in Sec. 6.2.5. Influence indicated for ANOVA results with $p \leq 0.005$.

Vocal Parameter	Influenced by Room Acoustics?	Influenced by Gender?	Influenced by Trial Number?	Influenced by interaction of room acoustics and trial number
F_0 mean (RB)	No	Yes	Yes	No
F_0 std (RB)	No	Yes	No	No
Ps mean (RB)	No	Yes	No	No
Ps std (RB)	No	Yes	No	No
dB mean (RB)	No	No	No	No
dB_mean_RB norm to ANCH	No	No	No	Yes
dB_mean_RB norm to C215	No	No	No	Yes
dB_std_RB	No	No	Yes	No
ActyFact_RB	No	No	No	Yes
AlphaRto_RB	No	No	No	Yes
dBspcSpF_RB	No	No	No	No
STSD_RB	No	Yes	No	No
DurOTas2_R2	No	No	No	No
syl_rate_R2	No	No	Yes	No
STSD_R2	No	Yes	No	No
CCPS_R2AH	Yes	No	No	No
AVQI_R2AH	Yes	No	No	No
Fo_mean_AH	No	Yes	No	No
Fo_std_AH	Yes	No	No	No
Ps_mean_AH	Yes	No	No	No
Ps_std_AH	No	Yes	No	No
dB_mean_AH	No	Yes	No	No
dB_std_AH	No	Yes	No	No
jitter_AH	No	Yes	No	No
shimmer_AH	Yes	Yes	No	No
HNR_AH	No	Yes	No	No
dB_mean_DE	Yes	No	No	No

dB_std_DE	No	Yes	Yes	No
ActyFact_DE	No	Yes	No	No
scaled VE	No	No	Yes	No
scaled VF	No	No	Yes	No
scaled VF20	No	No	Yes	No

6.3.1 Parameters Influenced by Gender

Several of the results were expected, but others were surprising. For instance, it is well known and expected that males and females have different fundamental frequencies, so to see the mean F_0 for both the RB and AH speech tasks as significantly influenced by gender was unsurprising. It was less expected to see that the standard deviation of F_0 for RB was also significantly influenced by gender, but the same was not true for AH. These parameters are shown in Fig. 6.1. The mean of each group is shown with a triangle, and the bars extending from the triangle are the standard error. The standard deviation results imply that females have more variation in their pitch during running speech than do males, but they have the same steadiness to the pitch as do males during sustained vowel. The STSD results indicate the same thing. Despite being on similar scales, females exhibited greater variation in semitones than did males during the running speech tasks RB and R2. These results are shown in Fig. 6.2.

Several more of the parameters pertaining to the sustained vowel task AH exhibited significant differences by gender. In addition to F_0, males had greater pitch strength standard deviation than did females, although the mean pitch strengths were not significantly different (Fig. 6.3). Males also had greater mean decibel level and greater decibel level standard deviation than did females. Males had greater jitter and shimmer than did females, and females had greater harmonics to noise ratio on average (Fig. 6.4). These all tend to indicate that during sustained vowel, the males were not as steady as females, which could be an indication of trending towards dysphonia.

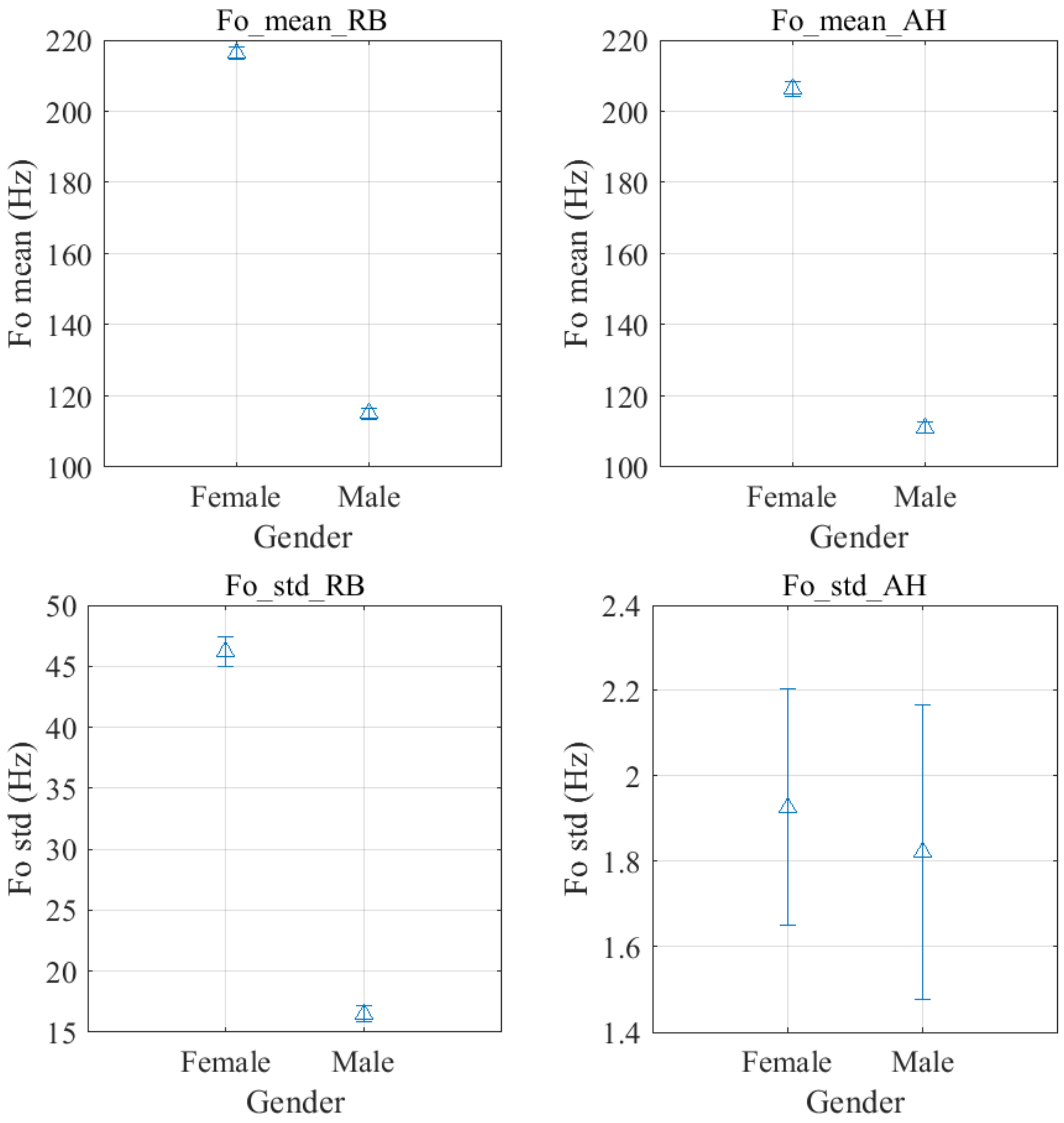


Figure 6.1. Fundamental frequency mean and standard deviation for the speech tasks AH, separated by gender. Females had higher fundamental frequencies, as expected. Durign running speech (RB), females also had greater deviations in their fundamental frequency than did males, although during sustained speech (AH), the deviations are similar for the genders.

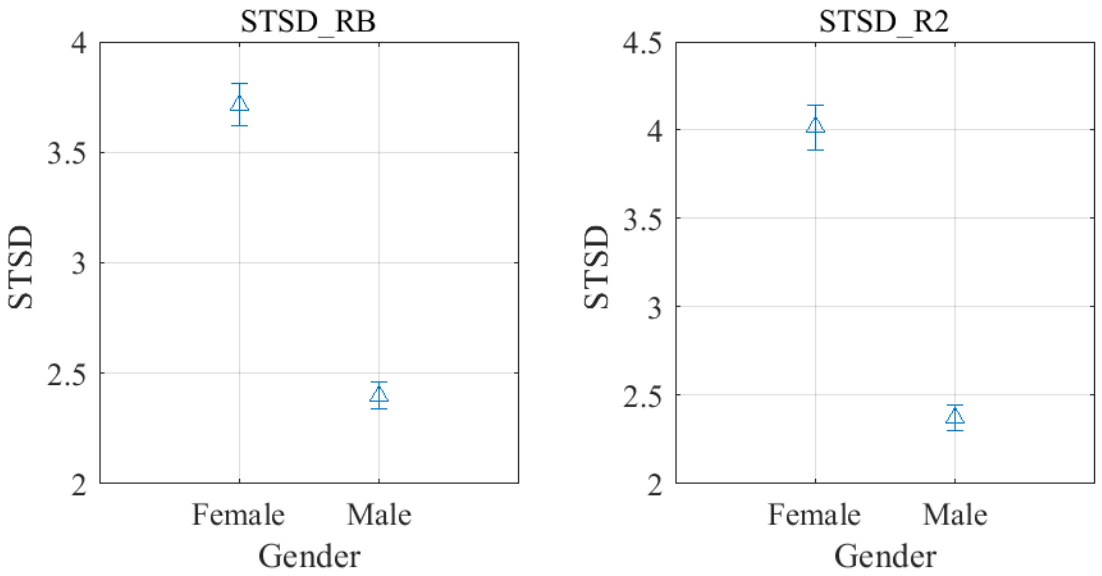


Figure 6.2. STSD by gender. The running speech tasks RB and R2 both indicate that females had greater STSD than males. This is interesting because STSD is meant to remove the differences in fundamental frequency betweem males and females to make the variations comparable on the same scale.

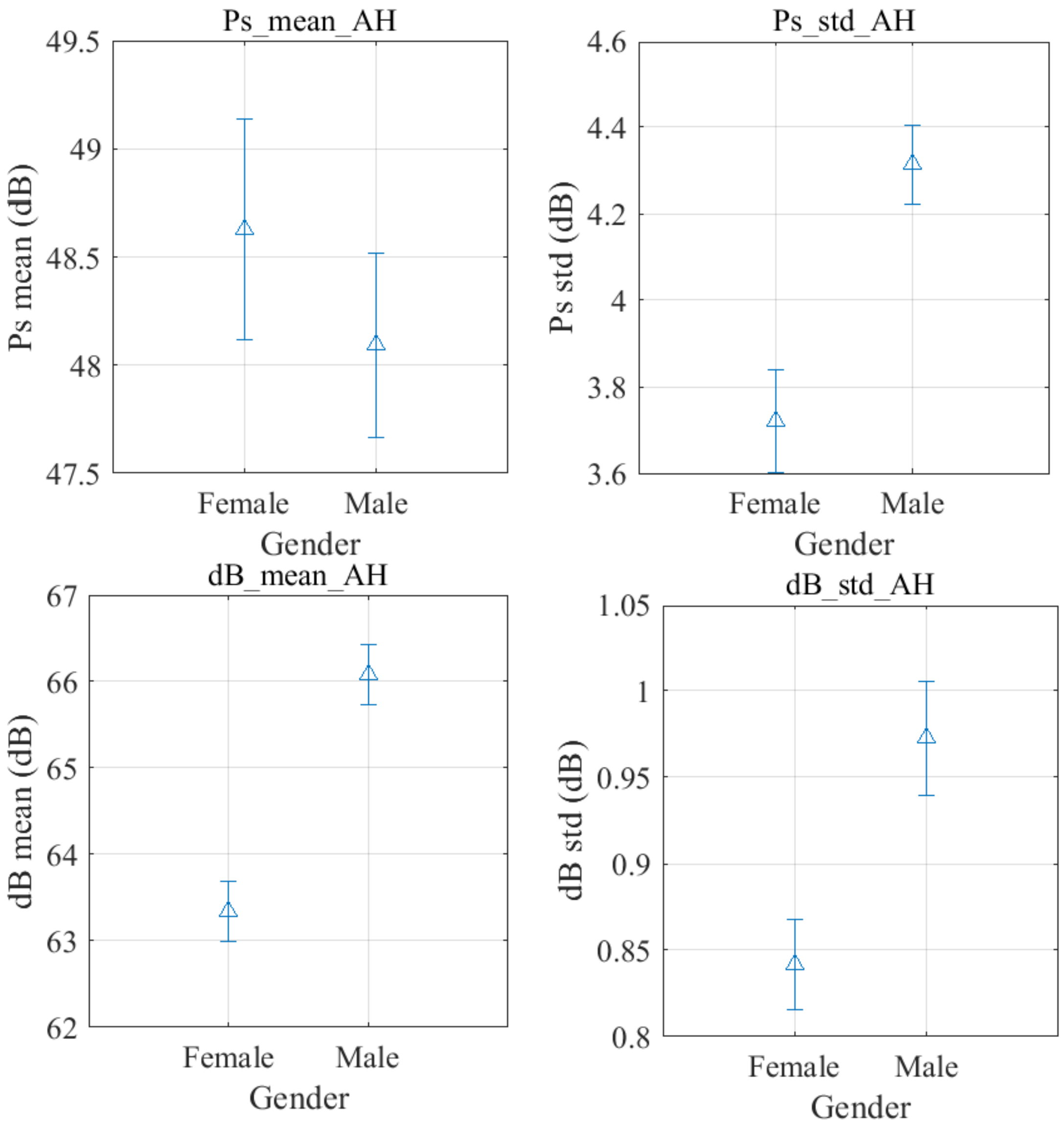


Figure 6.3. Pitch Strength and decibel level for the speech task AH, separated by Gender. Females had greater mean pitch strength and lower standard deviations in pitch strength than did males. Males had louder mean decibel levels and greater standard deviations in decibel levels.

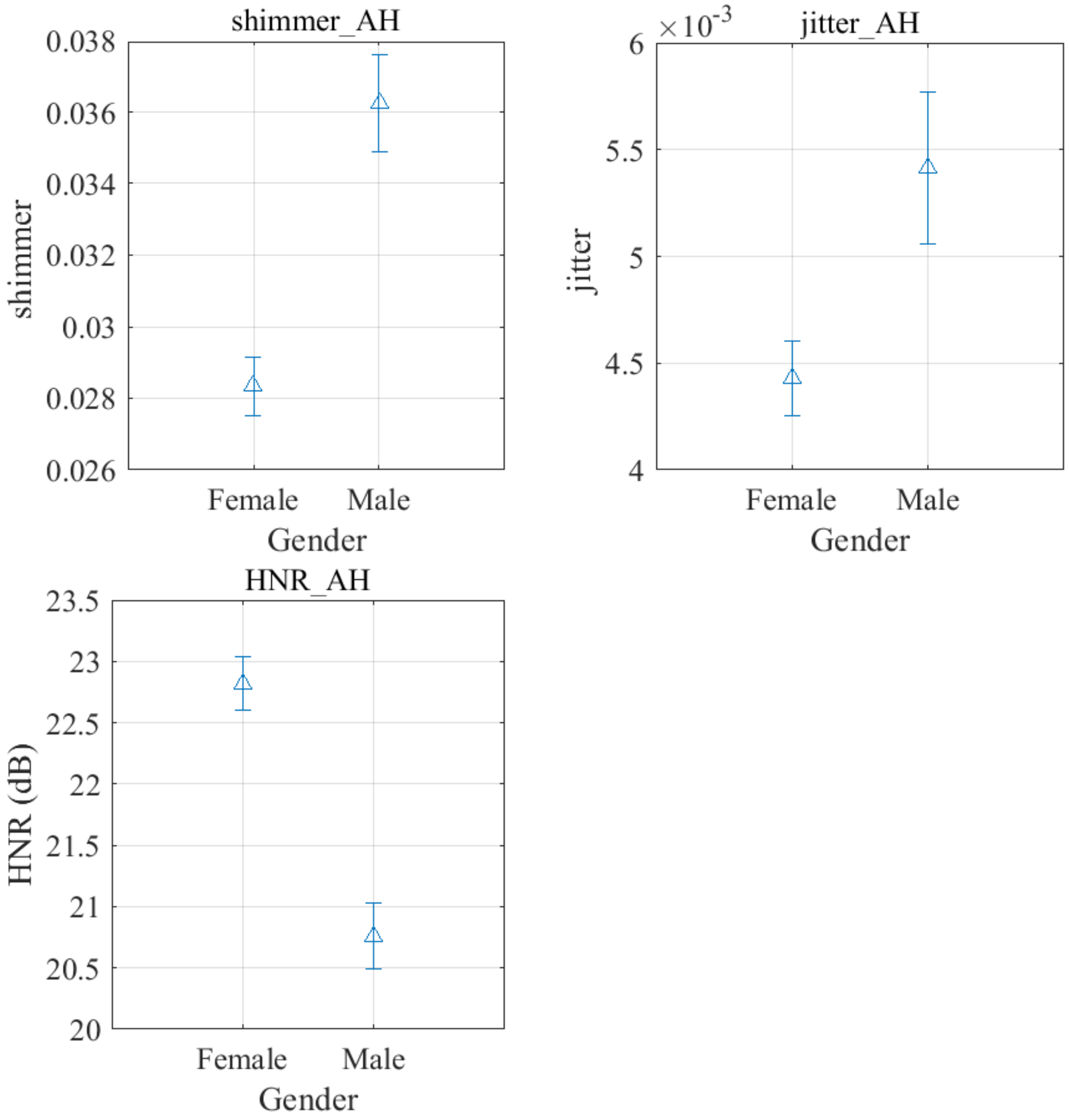


Figure 6.4. Shimmer, jitter, and HNR separated by Gender. Males had greater shimmer and jitter and lower HNR than did females, indicating a less steady sustained speech.

6.3.2 Parameters Influenced by Acoustic Condition

The parameters influenced by room acoustics came from all the speech tasks, showing once again that multiple types of speech are needed to show a complete picture of vocal effort. The room conditions are here represented by their room gains, defined in Sec. 3.6 as the difference in energy level in a room OBRIR compared to that of an anechoic OBRIR. AVQI was significantly influenced by room condition, as was CCPS, shown in Figure 6.5. While no linear trends across room gain are apparent for these parameters, the significant difference in AVQI is evident for DJCHM and DJCHS. Despite having similar values for room gain, the AVQI mean is different for these two conditions. However, the mean value of AVQI for any of the room conditions does not exceed 3.5, which is indicative of dysphonic speech. CCPS showed a significant difference in the mean value for ANCH and RE24M, but again, no linear trends across room gain are evident.

The sustained vowel also showed room-based differences in fundamental frequency standard deviation and mean pitch strength. They are shown in Fig. 6.6. Mean pitch strength tends to decrease with mid-value room gains, indicating less tonality and more noise in the signal. Fundamental frequency standard deviation does not show as clear a trend, as several of the rooms had extremely wide standard errors and may be influenced by outliers. In addition, the free-response speech task showed differences in mean decibel level. As room gain increased, the mean decibel level decreased. This is shown in Fig. 6.7 .

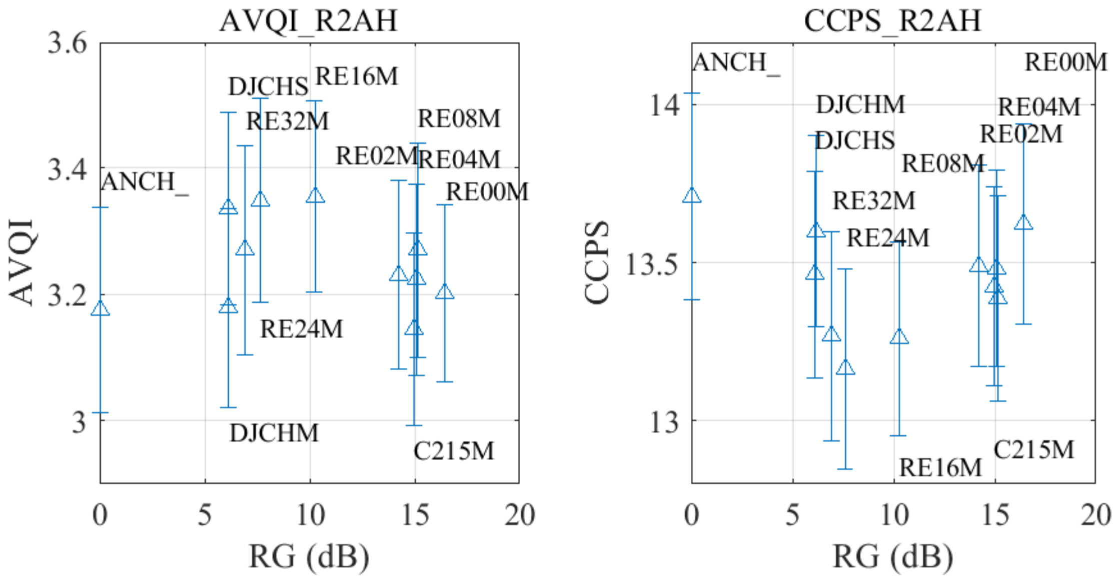


Figure 6.5. AVQI and CCPS against room gain.

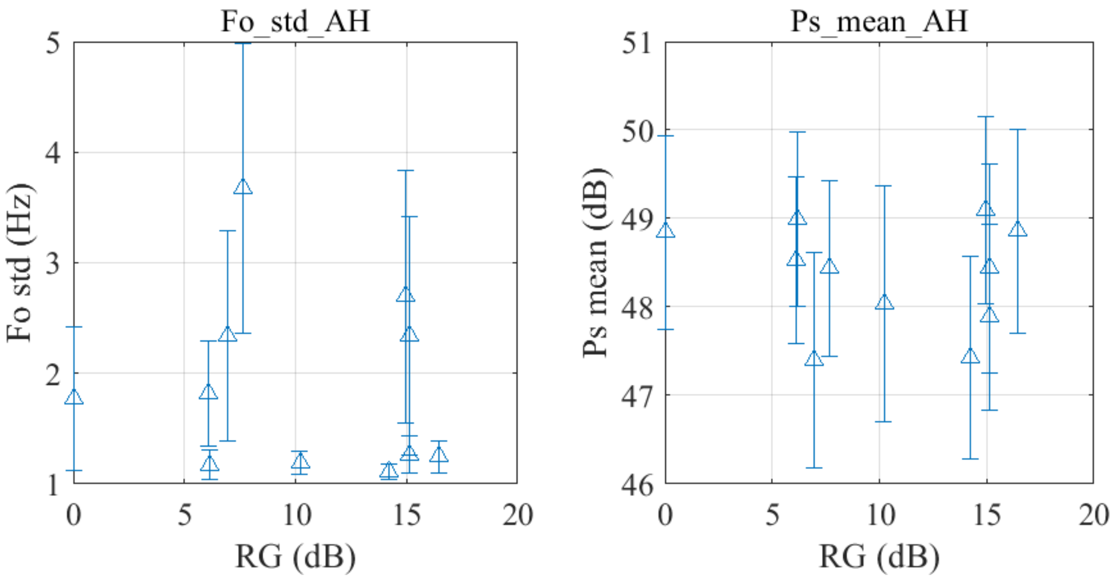


Figure 6.6. Fundamental frequency standard deviation and mean pitch strength for the speech task AH plotted against room condition.

Shimmer was the only vocal effort parameter to be significantly influenced by both room condition and gender. It is shown in Fig. 6.8. Males tended to have greater shimmer than did

females. As room gain increased, males also exhibited greater variation in shimmer than did females. Along with the results of jitter and HNR, this could be an indication that males did not have as steady a sustained vowel as did females. It is uncertain what trends across room condition may be drawn.

6.3.3 Parameters Influenced by Trial Number

Surprisingly, several parameters were slightly influenced by trial number, despite the random presentation of the room conditions. A slight increase in fundamental frequency was observed: about 2 Hz on average over the course of the trials, though no gender differences were reported as significant over trial number. The standard deviation of decibel value also showed a slight increase of about 0.3 dB for the speech task RB or 0.4 dB for the speech task DE. These small values are not large enough to indicate that participants were fatigued by the end of the study. A longer study might reveal a greater certainty in the trend and greater differences across

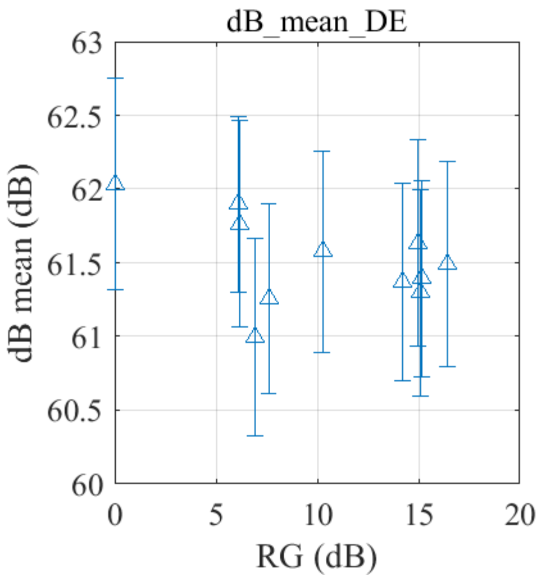


Figure 6.8. Mean decibel level for the speech task DE plotted against room condition. As room gain increases, the mean decibel level decreases.

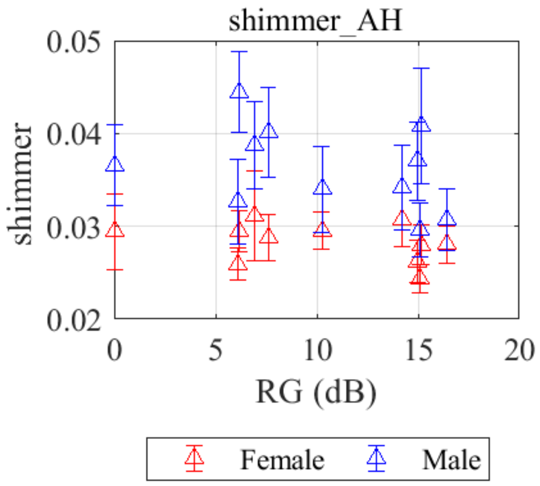


Figure 6.8. Shimmer for the speech task AH plotted against room condition and separated by gender.

trial number. Syllable rate for the two sentences of the rainbow passage increased about 0.3 syllables per second, showing that participants generally became faster at reading the passage over the course of the trials, likely due to the familiarity effect. However, this small value while

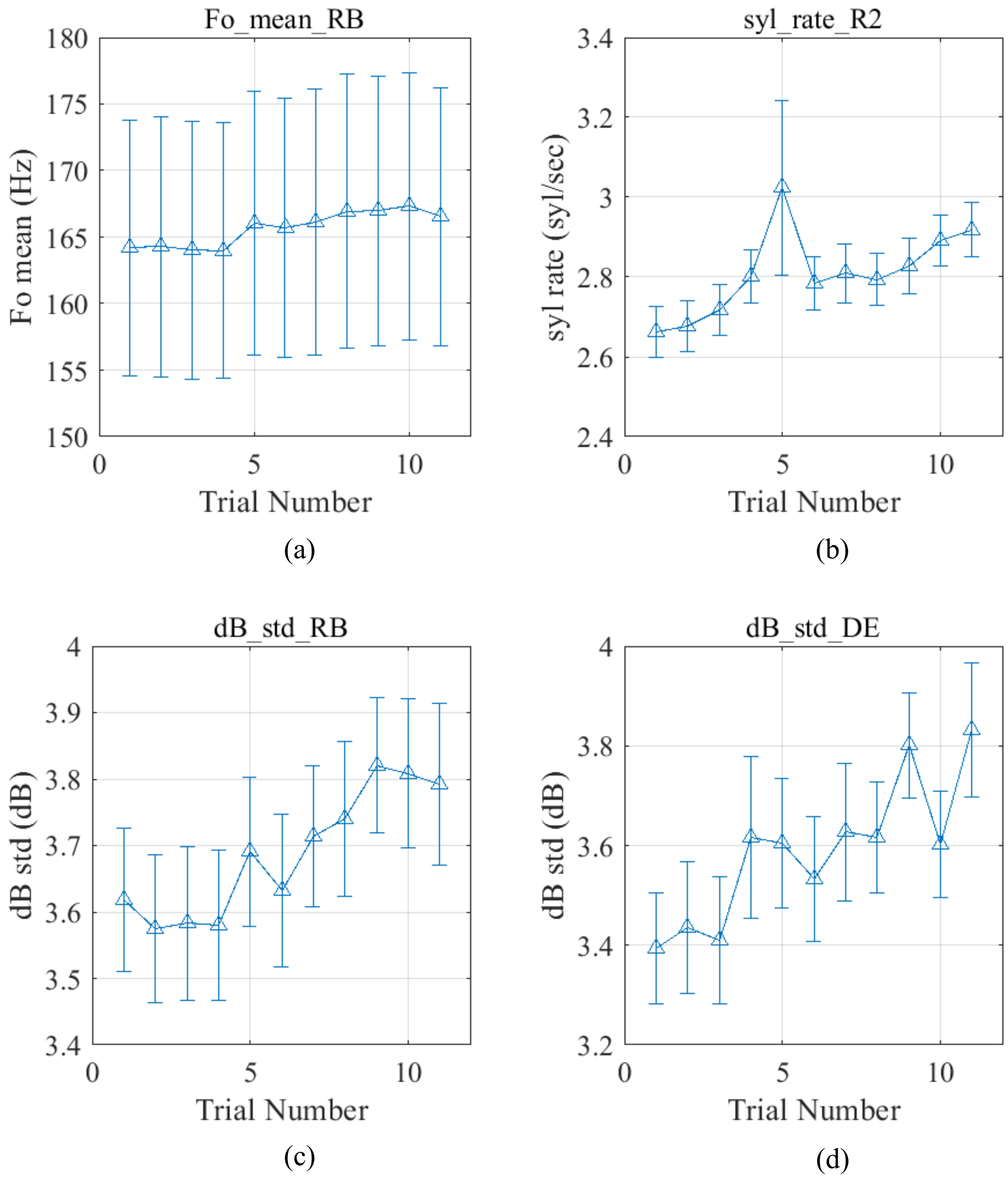
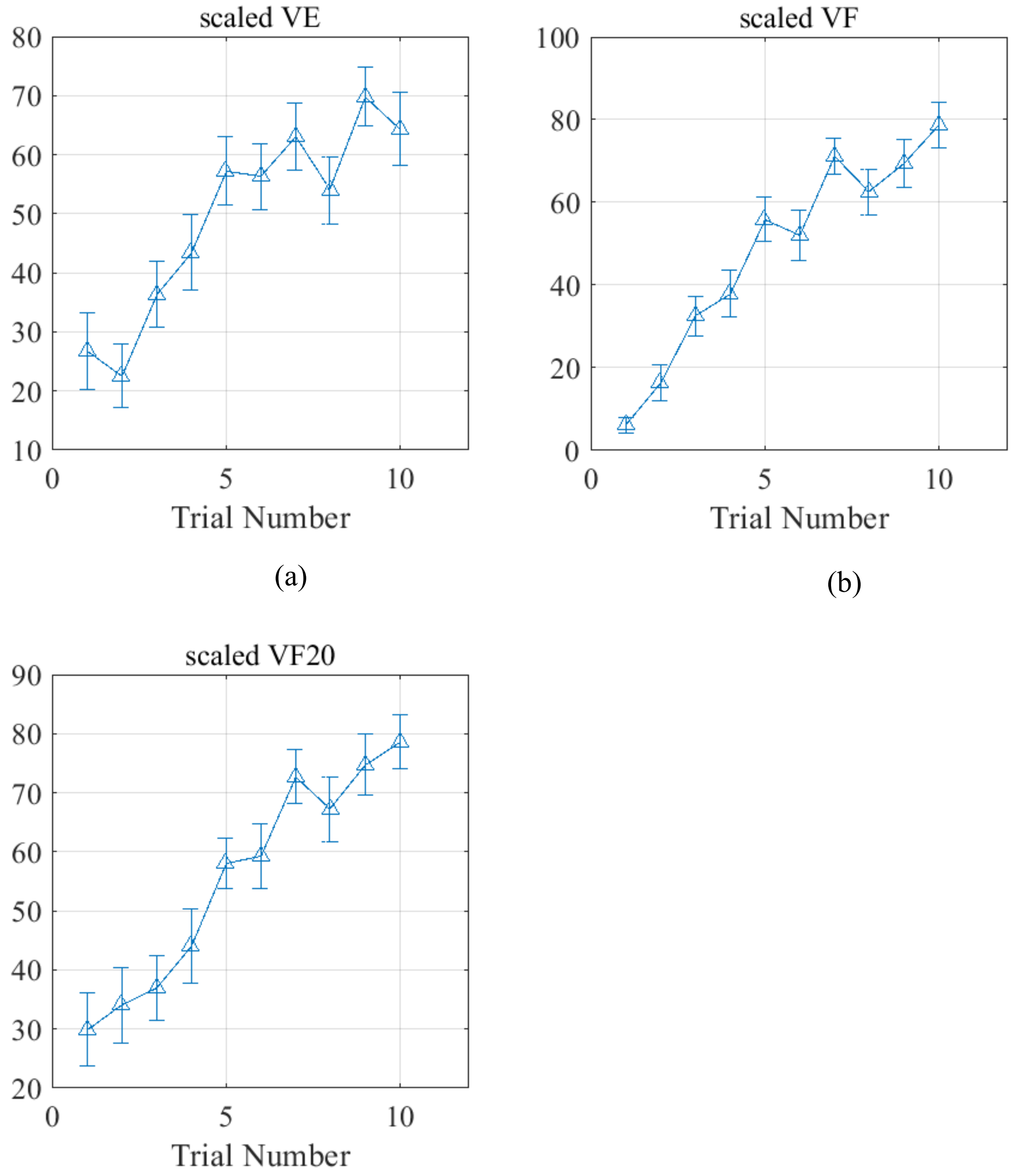


Figure 6.9. Vocal effort parameters significantly different by trial number. Subplot (a) shows the mean fundamental frequency for the speech task RB. Subplot (b) shows the syllable rate for the speech task R2. Subplot (c) shows the standard deviation for the decibel level for the speech task RB. Subplot (d) shows the standard deviation for the speech task DE.

significant is not particularly insightful. Fig. 6.9 shows these results. In addition, it appears that the self-reported vocal effort metrics all tended to increase with trial number. However, these self-reported metrics show a lack of correlation with the calculated vocal effort parameters. They are shown in Fig. 6.10.

The normalized decibel values showed significant differences due to the interaction of room acoustics and trial number due to the nature of the normalization to a specific acoustical



(a)

(b)

(c)

Figure 6.10. Self-reported vocal effort parameters significantly different by trial number. Subplot (a) shows the self reported vocal effort. Subplot (b) shows the self-reported vocal fatigue. Subplot (c) shows the self-reported 20 minute vocal fatigue.

condition, and due to the control condition (ANCH) always being in the 11th trial-number position. While this is true, it is not particularly insightful.

6.4 Comparison to Prior Vocal Effort Research

The design and implementation of the vocal effort study was strongly influenced by that of Rollins et al. [37] wherein the acoustical conditions were presented within a reverberation chamber with varying amounts of added absorption. The results for the similar conditions presented via the RTCS are compared to the results of that study.

Several trends in vocal effort metrics showed good agreement in vocal effort trends between the two studies. The parameters shown here are those found to be significantly different by acoustical condition in the RTCS vocal effort study. Mean and standard deviation of pitch strength exhibited no significant change across the reverberation chamber conditions (See Fig. 6.11). Mean decibel level increased with increasing room gain, while standard deviation of decibel level decreased slightly (see Fig. 6.12).

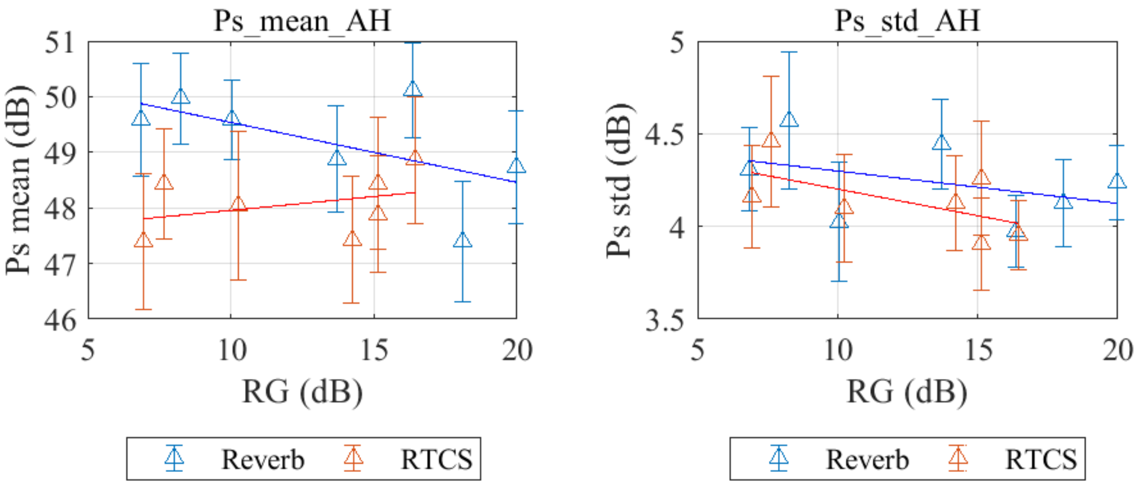


Figure 6.11. Pitch strength mean and standard deviation for the speech task AH, comparing RTCS and Rollins vocal effort studies.

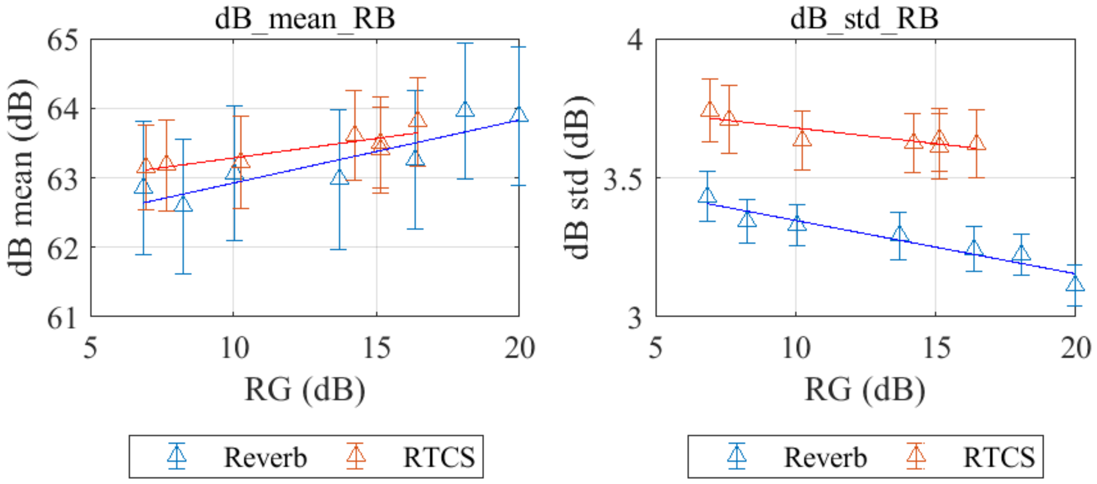


Figure 6.13. Decibel level mean and standard deviation for the speech task RB, comparing RTCS and Rollins vocal effort studies.

Harmonics to noise ratio and jitter are very similar for the two studies (see Fig. 6.13). Shimmer shows some differences by condition (see Fig. 6.14), but this parameter was also influenced by gender in the RTCS study. A look at the data from each of the genders plotted against room gain may yield some insights. Syllable rate decreased slightly with increasing room

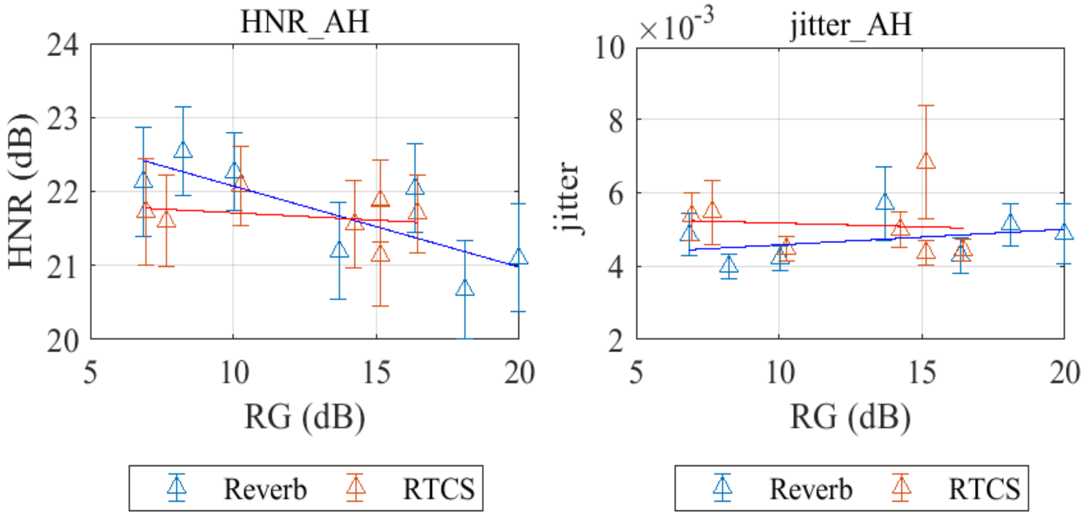


Figure 6.12. Harmonics-to-noise ratio and jitter for the speech task AH, comparing RTCS and Rollins vocal effort studies.

gain (see Fig. 6.14). The RTCS data has a clear outlier, but otherwise follows the trend of the Rollins data.

Rollins et al. found that AVQI increased toward dysphonia with increasing T20. They also indicate an increase towards dysphonia with increasing room gain, as seen in Fig. 6.15. The AVQI results from the RTCS study appear not to increase with room gain, but stay consistent. Indeed, the values for AVQI in the RTCS study all appear to remain quite close to each other, indicating no statistical difference in the values. It is quite possible that an additional acoustic condition in the RTCS study was statistically different in its value for AVQI while these AVQI values are not significantly different from each other. Further statistical tests on the subset of RTCS vocal effort data containing only the reverberation-chamber conditions have yet to be carried out.

The differences between results of the Rollins study and the present study could be due to a number of dissimilarities between the studies. For one, the participants in the Rollins study had both visual and externally induced acoustical cues that the acoustical condition was changing while the RTCS study participants had none of these. In the Rollins study, participants saw the

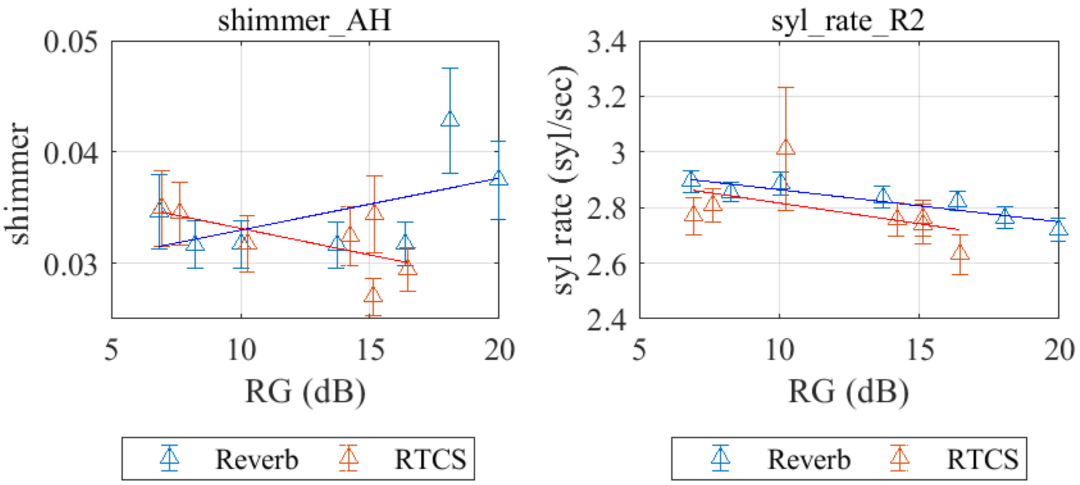


Figure 6.14. Shimmer for the speech task AH and syllable rate for the speech task R2, comparing RTCS and Rollins vocal effort studies.

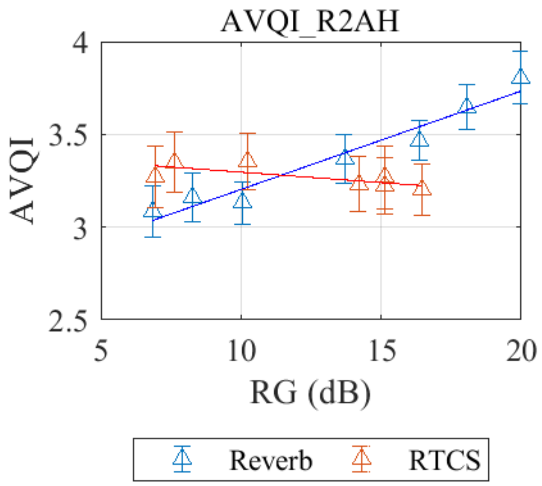


Figure 6.15. AVQI compared for the RTCS and Rollins vocal effort studies.

absorptive wedges being added to and removed from the room and talked with the interviewer in the altered environment prior to beginning the speaking tasks. In the RTCS study, there was no visual indication that the acoustical condition had changed, and the only acoustical cue that a new condition was being presented was the auralization of the participant's own voice. The interviewer's voice was always received anechoically and not included in the simulated acoustical environment. This may have led to the participant responding to interviewer cues more than to psychoacoustic cues, resulting in the difference in AVQI between the two studies. A first look at the results plotted against traditional room characterization parameters suggested that the vocal effort changes seen in relationship to the room parameters may have been due to the sound each talker heard from the interviewer, perhaps as much or more than from his or her own voice.

6.5 Discussion

The vocal effort parameters significantly different by trial number are small enough that they do not pose concern for the validity of the results for the study. Slight changes in mean fundamental frequency and standard deviation in decibel level are not enough to say that participants were fatigued by the end of the study. However, the self-reported vocal effort data do indicate that participants felt more fatigued over the course of their time in the study. There were no significant differences in the self-reported vocal effort over acoustic conditions.

The vocal effort parameters significantly different by gender yield some interesting results. Females had higher mean fundamental frequencies than males as expected. They also had greater fundamental frequency standard deviations, greater mean pitch strength, and greater pitch strength standard deviations for the speech task RB. Females also had higher semitone standard deviations than males. Males exhibited greater mean decibel level and standard deviations in decibel level for the speech tasks AH and DE.

The vocal effort parameters significantly different by room condition showed that voices tended towards dysphonia with mid-value room gains. In other words, the subjects exhibited more vocal effort for conditions that were not as extreme in their room gains, as seen in the parameters of mean decibel level, mean pitch strength, fundamental frequency standard deviation, shimmer, CCPS, and AVQI.

6.6 Conclusions

The vocal effort study showed the utility of the RTCS in being able to quickly change acoustic conditions without changing physical location. Participants were able to experience 11 acoustic conditions in approximately 45 minutes. An initial look at the vocal effort data revealed significant differences in female and male behavior over the course of the study, especially in parameters. Trends across room condition were less clear, but significant differences between several of the room conditions were evident. A few of the vocal effort parameters were significantly different across trial number, including the self-reported vocal effort and vocal fatigue. A comparison to a similar, prior study showed similar results in the vocal effort data, but further investigation into the vocal effort data and its trends across room conditions is needed.

Chapter 7

Conclusions

This thesis has described the development of a real-time convolution system (RTCS) intended to create virtual-acoustical conditions for live talkers. It included high-resolution directivity measurements of human speech, several key hardware and software system components, various measured and simulated oral-binaural room impulse responses (OBRIRs), and equalization filters necessary to ensure realistic performance. The utility of the RTCS was demonstrated through a brief study on vocal effort that investigated talker responses to many room conditions through several speech tasks.

A significant contribution of the work involved its high-resolution speech directivity measurements, which represent the highest-resolution results for live human speech acquired to date. They will inform speech scientists, architectural acousticians, audio engineers, hearing-aid engineers, telecommunications engineers, automotive engineers, and other specialists in ways that will help improve their efforts. The measurements were averaged over small groups of males and females, but a larger sample would produce more general averages. Speech directivities taken at closer radii would enable use of spherical near-field acoustical holography to further explore speech radiation in the near field and at many other radii.

Another noteworthy contribution of the work included its improved modeling and implementation of a RTCS. The development of unique inversion filters and subsequent objective and subjective validation measures were especially important for RTCS optimization.

The objective measures assessed the RTCS validity in both the time and frequency domains. When used in combination with subjective evaluations, they allowed the realism of the presented acoustical conditions to be ascertained and improved. These measures demonstrated that the inclusion of compensation filters improved the RTCS performance.

A further contribution included the use of the RTCS for a vocal effort study. While the system could be used for more scenarios than those discussed, the conditions used and reported for the study represent a step forward in the field of speech acoustics involving the use of virtual-acoustic reality. The study was unique in that it utilized both measured and simulated OBRIRs in the RTCS and did not rely on artificial reverberation. Instead, it employed measurements and simulations of actual environments from the acoustical perspective of a talker-listener. More investigation is needed to determine if the study results align with expectations for vocal effort parameters. Nevertheless, the methods and results will be useful to those seeking to create similar tests involving virtual environments.

The research effort was successful in reaching stated objectives. The RTCS was developed and validated for the presentation of specific acoustic environments for talkers in vocal effort studies. The system could also be useful for other virtual acoustic studies, in addition to those presented here. It could be improved through the addition of channels, such as one dedicated to convolution of an interviewer's voice within the virtual acoustic environment. This would further increase realism and perhaps elicit vocal effort responses more congruent with those that would be produced in actual acoustical environments. Individual HRTFs could also be used with architectural-acoustic simulations for the specific test subjects. Visual components representing the environments and head-tracking capabilities would also improve realism as part of a larger virtual-reality system. The author recommends these and other improvements to enhance future work in this area.

Bibliography

[1] D. Cabrera, M. Yadav, and W. L. Martens, "A system for simulating room acoustical environments for one's own voice," Applied Acoustics **73**, 409-414 (2012).

[2] C. Porschmann, "Influences of bone conduction and Air Conduction on the sound of one's own voice," Acta Acustica United With Acustica **86**, 1038-1045 (2000).

[3] F. Otondo and J. H. Rindel, "The influence of the directivity of musical instruemnts in a room," Acta Acustica united with Acustica **90**, 1178-1184 (2004).

[4] F. Hohl, "Kugelmikrofonarray zur Abstrahlungsvermessung von Musikinstrumenten" ("Spherical microphone array for radiation survey of musical instruments")," Masters Thesis, University of Music and Performing Arts, , Graz, Austria (2009).

[5] F. Hohl and F. Zotter, "Similarity of musical instrument radiation-patterns in pitch and partial," Fortschritte der Akusti, DAGA,, Berlin, (2010).

[6] J. L. Carrou, Q. Leclere, and F. Gautier, "Some characteristics of the concert harp's acoustic radiation," J. Acoust. Soc. Am. **127**, 3202-3211 (2010).

[7] J. Bodon, "Development, Evaluation, and Validation of a High Resolution Directivity Measurement System for Played Musical Instruments," Masters of Science Thesis, Brigham Young University, Provo (2016).

[8] M. Pollow, "Measuring directivities of musical instruments for auralization," Fortschritte der Akustik **35**, 1471-1473 (2009).

[9] M. Pollow, G. Behler, and B. Masiero, "Measuring directivities of natural sound sources with a spherical microphone array," Ambisonics Symposium 2009, **1**. (2009).

[10] M. Pollow, G. K. Behler, and F. Schultz, "Musical instrument recording for building a directivity database," Fortschritte der Akustik, **36**. Deutsche Jahrestagung für Akustik, Berlin, Germany, (2010).

[11] M. Vorlander, *Auralization Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. (Springer, Verlag Berlin Heidelberg, 2008).

[12] AFMG, "Enhanced Acoustic Simulator for Engineers," [available online at <http://ease.afmg.eu/index.php/ears_module.html> (Last viewed 2016)].

[13] Odeon, "ODEON room acoustics software," [available online at <www.odeon.dk> (Last viewed 2017)].

[14] CATT-Acoustic, "CATT-Acoustic™ v9.1b," [available online at <www.catt.se> (Last viewed 2018)].

[15] H. Moller, "Fundamentals of Binaural Technology," Applied Acoustics **36** (1992).

[16] D. Schroder, D. Rausch, F. Wefers et al., "Virtual reality system at RWTH Aachen University," Proceedings on the International Symposium on Room Acoustics, 1-9, (2010).

[17] M. Kleiner, B.-I. Dalenback, and P. Svensson, "Auralization - an overview," J. Audio Engineering Soc. **41** (11), 861-875 (1993).

[18] D. S. Brungart, B. D. Simpson, R. L. McKinley et al., "The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources," Proceedings of ICAD 04 - Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, (2004).

[19] S. Weinzierl and M. Vorländer, "Room acoustical parameters as predictors of room acoustical impression: what do we know and what would we like to know?," Acoustics Australia **43** (1), 41-48 (2015).

[20] D. Cabrera, W. L. Martens, D. Lee et al., "Binaural measurement and simulation of the room acoustical response from a person's mouth to their ears," Acoustics Australia **37** (3), 98-103 (2009).

[21] N. J. Eyring, "Development and Validation of an Automated Directivity Acquisition System used in the Acquisition, Processing, and Presentation of the Acoustic Far-Field Directivity of Musical Instruments in an Anechoic Space," Masters of Science, Brigham Young University, Provo, UT (2013).

[22] BYU-ARG, "Acoustics Facilities at BYU," [available online at <http://acoustics.byu.edu/content/acoustics-facilities-byu> (Last viewed 2018)].

[23] V. L. Ahlander, R. Rydell, and A. Lofqvist, "Speaker's comfort in teaching environments: voice problems in Swedish teaching staff," J. Voice **25** (4), 460-440 (2011).

[24] A. Astolfi, A. Vallan, A. Carullo et al., "Influence of classroom acoustics on the vocal behavior of teachers," Proceedings of Meetings on Acoustics **19**, 1-9 (2013).

[25] B. B. Boren and A. Roginska, "Sound radiation of trained vocalizers," Proceedings of Meetings on Acoustics **19** (2013).

[26] P. Bottalico and A. Astolfi, "Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms," J. Acoust. Soc. Am. **131** (4), 2817-2827 (2012).

[27] J. Brunskog and A. C. Gade, "Increase in voice level and speaker comfort in lecture rooms," J. Acoust. Soc. Am. **125**, 2072-2082 (2009).

[28] N. Durup, B. Shield, S. Dance et al., "Vocal strain in UK teachers: An investigation into the acoustic causes and cures," J. Acoust. Soc. Am. **133** (5), 3553-3553 (2013).

[29] M. George and M. Youssef, "Acoustical quality assessment of the classroom environment," arXiv preprint arXiv: 1201. 2902 (2012).

[30] M. Kob, A. Kamprolf, C. Neuschaefer-Rube et al., "Experimental investigations of the influence of room acoustics on the teacher's voice," Acoustics in Science and Technology **29** (1), 86-94 (2008).

[31] L. Nijs and M. Rychtarikova, "Calculating the optimum reverberation time and absorption coefficient for good speech intelligibility in classroom design using U50," Acta Acustica United With Acustica **97**, 93-102 (2011).

[32] D. Pelegrin-Garcia and J. Brunskog, "Speakers' comfort and voice level variation in classrooms: Laboratory research," J. Acoust. Soc. Am. **132** (1), 249-260 (2012).

[33] B. Rasmussen, D. Hoffmeyer, and J. Brunskog, "Reverberation time in classrooms - comparison of regulations and classification criteria in the Nordic countries," Joint Baltic-Nordic Acoustics Meeting 2012, 1-6 (2012).

[34] D. Rostolland, "Acoustic features of shouted voice," Acta Acustica United With Acustica **50** (1982).

[35] A. Russell, J. Oates, and K. M. Greenwood, "Prevalence of voice problems in teachers," J. Voice **12** (4), 467-479 (1998).

[36] I. R. Titze, J. Lemke, and D. Montequin, "Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report," J. Voice **11** (3), 254-259 (1997).

[37] M. Rollins, "The influence of room acoustics on the voice," Bachelor of Science Capstone Project Report, Brigham Young University, Provo, UT (2016).

[38] H. K. Dunn and D. W. Farnsworth, "Exploration of pressure field around the human head during speech," J. Acoust. Soc. Am. **10** (1) (1939).

[39] G. Studebaker, "Directivity of the human vocal source in the horizontal plane," Ear Hear **6** (6), 315-319 (1985).

[40] G. Studebaker, "Directivity of the human vocal source," J. Acoust. Soc. Am. **73**, S105-S105 (1983).

[41] W. T. Chu and A. C. C. Warnock, "Detailed Directivity of Sound Fields Around Human Talkers," NRC Publications Archive (2002).

[42] A. H. Marshall and J. Meyer, "The directivity and auditory impressions of singers," Acta Acustica United With Acustica **58** (1985).

[43] B. Katz and C. d'Alessandro, "Measurement of 3-D phoneme specific radiation patterns in speech and singing," Scientific Report (2007).

[44] B. Katz, C. d'Alessandro, and F. Prezat, "Human voice phoneme directivity pattern measurements," J. Acoust. Soc. Am. **120**, 3359 (2006).

[45] B. B. Monson, E. J. Hunter, and B. H. Story, "Horizontal directivity of low- and high-frequency energy in speech and singing," J. Acoust. Soc. Am. **132** (1), 433-441 (2012).

[46] M. Kob, "Directivity measurement of a singer," Collected Papers from the Joint Meeting "Berlin 99" (1999).

[47] F. Bazzoli, P. Bilzi, and A. Farina, "Influence of artificial mouth's directivity in determining speech transmission index," Audio Engineering Society Convention 119, 6571, (2005).

[48] F. Bazzoli and A. Farina, "Directivity balloons of real and artificial mouth simulators for measurement of the Speech Transmission Index " Audio Engineering Society Convention 115, 5953, (2003).

[49] F. Bazzoli, A. Farina, and M. Viktorovitch, "Balloons of directivity of real and artificial mouth used in determining speech transmission index," Audio Engineering Society Convention 118, 6492, (2005).

[50] T. Halkosaari, "Radiation Directivity of Human and Artificial Speech," PhD Dissertation, Helsinki University of Technology, Helsinki, Finland (2004).

[51] T. Halkosaari and M. Vaalgamaa, "Directivity of Human and Artificial Speech," Joint Baltic-Nordic Acoustics Meeting, 8-10, (2004).

[52] T. Halkosaari, M. Vaalgamaa, and M. Karjalainen, "Directivity of Artificial and Human Speech," J. Audio Engineering Soc. **53** (7/8), 620-631 (2005).

[53] T. Snaidero, F. Jacobsen, and J. Buchholz, "Measuring HRTFs of Bruel and Kjaer Type 4128-C G.R.A.S. KEMAR Type 45 BM and Head Acoustics HMS II.3 Head and Torso Simulators", (Tehnical University of Denmark, Department of Electrical Engineering, 2011).

[54] AES 56-2008 AES standard on acoustics - Sound source modeling - Loudspeaker polar radiation measurements, Audio Engineering Society Incorporated, New York, (2008).

[55] E. M. Lai, G. A. Carrijo, R. Bennett et al., "An English language speech database at the University of Western Australia," Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, 101-104, (1990).

[56] T. Leishman, S. Rollins, and H. Smith, "An experimental evaluation of regular polyhedron loudspeakers as omnidirectional sources of sound," J. Acoust. Soc. Am. **120**, 1411-1422 (2006).

[57] C. M. Pincock, "High-Resolution Speech Directivity Balloons," Bachelor of Science, Brigham Young University, Provo, UT (2017).

[58] J. D. Maynard, E. G. Williams, and Y. Lee, "Nearfield acoustic holography I. Theory of generalized holography and the development of NAH," J. Acoust. Soc. Am. **78**, 1395-1413 (1985).

[59] F. T. Ulaby and A. E. Yagle, *Engineering Signals and Systems*. (2013).

[60] GRAS, "Sound & Vibration A/S, Head and torso simulators," [available online at <http://www.gras.dk/products/head-torso-simulators-kemar.html> (Last viewed 2016)].

[61] GRAS, " Sound & Vibration A/S, GRAS 45BC KEMAR Head & Torso with Mouth Simulator, Non-configured," [available online at <http://www.gras.dk/products/head-torso-simulators-kemar/product/749-45bc> (Last viewed 2018)].

[62] M. Burkhard and R. Sachs, "Anthropometric manikin for acoustic research," J. Acoust. Soc. Am. **58** (1), 214-222 (1975).

[63] AFMG, "EASERA Universal Measuring Platform," [available online at <http://easera.afmg.eu/> (Last viewed 2018)].

[64] D. Nutter, "Sound absorption and Sound Power Measurements in Reverberation Chambers Using Energy Density Methods," Masters of Science, Brigham Young University, Provo, UT (2006).

[65] L. L. Beranek and H. P. S. Jr., "The Design and Construction of Anechioc Sound Chambers," J. Acoust. Soc. Am. **18** (1), 140-150 (1946).

[66] BYU-Arts, "Venues," [available online at <http://arts.byu.edu/venues/> (Last viewed 2017)].

[67] ISO 3382-2:2008 Measurement of room acoustic parameters Part 2: Reverberation time in ordinary rooms, International Organization for Standardization, Geneva, Switzerland, (2008).

[68] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," J. Acoust. Soc. Am. **66**, 497-500 (1979).

[69] A. Farina, "Convolution of anechoic music with binaural impulse responses," Proceedings of PARMA-CM Users Meeting, (1993).

[70] S. Favrot and J. M. Buchholz, "LoRA: A loudspeaker-based room auralization system," Acta Acustica United With Acustica **96**, 364-375 (2010).

[71] C. Muller-Tomefelde, "Low Latency Convolution for Real Time Applications," Audio Engineering Society Conference: 16th Internation Conference: Spatial Sound Reproduction., (1999).

[72] M. Noisternig and B. Katz, "Framework for Real-Time Auralization in Architectural Acoustics," Acta Acustica United With Acustica **94**, 1000-1015 (2008).

[73] S. Pelzer, M. Pollow, and M. Vorlander, "Auralization of a virtual orchestra using directivities of measured symphonic instruments," Proceedings of the Acoustics 2012 Nantes Conference, Nantes, France, (2012).

[74] D. Cabrera, B. Hartmann, D. Lee et al., "Characterising the variation in oral-binaural room impulse responses for horizontal rotations of a head and torso simulator," Proceedings on the International Symposium on Room Acoustics, 1-10 (2010).

[75] C. Knufinke, "SIR2 Reverb," [available online at <https://www.siraudiotools.com/sir2.php> (Last viewed 2018)].

[76] M. Yadav, D. Cabrera, L. Miranda et al., "Variations in acoustical parameters in oral-binaural room impulse response of a real and a computer-modeled room," Proceedings of Acoustics - Fremantle, 1-5 (2012).

[77] M. Yadav, D. Cabrera, and W. L. Martens, "Auditory room size perceived from a room acoustic simulation with autophonic stimuli," Acoustics Australia **39** (3), 101-105 (2011).

[78] M. Yadav, D. Lee, D. Cabrera et al., "The regulation of voice levels in various room acoustic conditions," Proceedings of Acoustics 2013, Victor Harbor, Australia, (2013).

[79] D. Cabrera, W. L. Martens, M. Yadav et al., "Evaluation of stage acoustics preference for a singer using oral-binaural room impulse responses," Proceedings of Meetings on Acoustics **19**, 1-9 (2013).

[80] D. Cabrera, D. Lee, R. Collins et al., "Variation in oral-binaural room impulse responses for horizontal rotations of a head and torso simulator," Building Acoustics **18** (1,2), 227-252 (2011).

[81] H. Sato, M. Morimoto, and K. Fukunaga, "Effects of reverberation sounds on conversing difficulty in living rooms," Tech. Rep. Architectural acoustics Acoustical Society of Japan, 1-8 (2008).

[82] G. V. Bekesy, "The structure of the middle ear and the hearing of one's own voice by bone conduction," J. Acoust. Soc. Am. **21** (3), 217-232 (1949).

[83] C. Porschmann, "Eigenwahrnehmung der Stimme in virtuellen auditiven Umgebungen," Fortschritte der Akustik **24**, 550-551 (1998).

[84] J. Blauert, L. Hilmar, J. Sahrhage et al., "An interactive virtual-environment generator for psychoacoustic research. I: Architecture and implementation," Acta Acustica united with Acustica **86** (1), 94-102 (2000).

[85] T. Djelani, C. Porschmann, J. Sahrhage et al., "An interactive virtual-environment generator for psychoacoustic research II: Collection of head-related impulse responses and evaluation of auditory localization.," Acta Acustica united with Acustica **86** (6), 1046-1053 (2000).

[86] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," Presence: Teleoperators and virtual environments **7** (3), 225-240 (1998).

[87] D. P. Garcia, M. Rychtanikova, C. Glorieux et al., "Interactive auralization of self-generated oral sounds in virtual acoustic environments for research in human echolocation," Proceedings of Forum Acusticum 2014, (2014).

[88] D. Pelegrin-Garcia, O. Fuentes-Mendizabal, J. Brunskog et al., "Equal autophonic level curves under different room acoustics conditions," J. Acoust. Soc. Am. **130** (1), 228-238 (2011).

[89] D. Pelegrin-Garcia, "Local variations of speaker-oriented acoustic parameters in typical classrooms: a simulation study," Proceedings of Euronoise 2015, 703-708, (2015).

[90] C. Cheng and G. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," J. Audio Engineering Soc. **49** (4), 231-249 (2001).

[91] M. Vorlander, "Virtual acoustics: opportunities and limits of spatial sound reproduction for audiology," German Society for Medical Physics **39**, 1-4 (2008).

[92] T. Lentz, "Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments," J. Acoust. Soc. Am. **54** (4), 293-294 (2006).

[93] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. (Prentice Hall, Upper Saddle River, New Jersey, 1999), 2 ed.

[94] Z. Scharer and A. Lindau, "Evaluation of equalization methods for binaural signals," Audio Engineering Society Convention 126, (2009).

[95] J. N. Mourjopoulos, "Digital equalization of room acoustics," J. Audio Engineering Soc. **42** (11), 884-900 (1994).

[96] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics, Speech, and Signal Processing **29** (2), 254-272 (1981).

[97] J. N. Mourjopoulos, P. M. Clarkson, and J. K. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," Proceedings of hte 1982 IEEE International Conference on ASSP, Paris, (1982).

[98] J. N. Mourjopoulos, P. M. Clarkson, and J. K. Hammond, "Spectral phase and transient equalization for audio systems," J. Audio Engineering Soc. **33** (3), 127-132 (1985).

[99] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," J. Audio Engineering Soc. **47** (7), 583-595 (1999).

[100] R. Bucklein, "The audibility of frequency response irregularities," J. Audio Engineering Soc. **29** (3), 126-131 (1981).

[101] F. E. Toole and B. M. Sayers, "Lateralization judgements and the nature of binaural acoustic images," J. Acoust. Soc. Am. **37**, 319-324 (1965).

[102] F. Brinkmann and S. Weinzierl, presented at the Audio Engineering Society 142nd Convention, Berlin, Germany, 2017 (unpublished).

[103] O. Kirkeby, P. A. Nelson, H. Hamada et al., "Fast deconvolution of multichannel systems using regularization," IEEE Transactions on Acoustics, Speech, and Signal Processing **6** (2), 189-194 (1998).

[104] S. G. Norcross, M. Bouchard, and G. A. soulodre, "Inverse filtering design using a minimal phase target function from regularization," 121st AES Convention, Convention Paper 6929, San Francisco, USA, (2006).

[105] B. C. Moore, *Hearing*. (Academic Press, 1995).

[106] F. Brinkmann, "Individual headphone compensation for binaural synthesis," Masters, Technical University Berlin, (2011).

[107] Z. Scharer, "Kompensation von Frequenzgangen in Kontext der Binauraltechnik.," Magisterabeit, Technische Universitat, Berlin, Germany (2008).

[108] M. Slaney, "Auditory toolbox. Version 2", (Interval Research Corporation, 1998), Vol. Technical report 10.

[109] J. H. Johnson, C. W. Turner, J. J. Zqislocki et al., "Just noticeable differences for intensity and their relation to loudness," J. Acoust. Soc. Am. **93** (2), 983-991 (1993).

[110] P. Bottalico, S. Graetzer, and E. J. Hunter, "Effects of voice style, noise level, and acoustic feedback on objective and subjective voice evaluations," J. Acoust. Soc. Am. **138** (6), EL498-EL503 (2015).

[111] E. J. Hunter, "A comparison of a child's fundamental frequencies in structured elicited vocalizations versus unstructured natual vocalizations: A case study," Int. J. Pediatr. Otorhi. **73** (4), 561-571 (2009).

[112] L. C. C. Cutiva, P. Bottalico, and E. J. Hunter, "Vocal fry and vowel height in different room acoustics," Folia Phoniatr Logop (in press).

[113] P. Bottalico, S. Graetzer, and E. J. Hunter, "Effect of training and level of external auditory feedback on the singing voice: Pitch inaccuracy," J. Voice **31** (1), 122e129-122e116 (2017).

[114] P. Bottalico, S. Graetzer, and E. J. Hunter, "Effects of speech style, room acoustics, and vocal fatigue on vocal effort.," J. Acoust. Soc. Am. **139** (5), 2870-2879 (2016).

[115] P. Bottalico, S. Graetzer, and E. J. Hunter, "Effects of voice style, noise level and acoustic feedback on objejctive and subjective voice evaluations," J. Acoust. Soc. Am. **138** (6), EL498-503 (2015).

[116] P. Bottalico, I. Ipsaro-Passione, S. Graetzer et al., "Evaluation of the starting point of the Lombard Effect," Acta Acustica united with Acustica **103** (1), 169-172 (2017).

[117] S. Graetzer, P. Bottalico, and E. J. Hunter, "Speech produced in noise: Relationship between listening difficulty and acoustic and durational parameters," J. Acoust. Soc. Am. (in press).

[118] P. Bottalico, A. Astolfi, and E. J. Hunter, "Teachers' accumulation of voicing and silence periods of continuouse speech in classrooms with different reverberation times," J. Acoust. Soc. Am. **141** (1), EL26-EL31 (2017).

[119] J. W. Black, "the effect of room characteristics upon vocal intensity and rate," J. Acoust. Soc. Am. **22** (174) (1950).

[120] G. Fairbanks, *Voice and Articulation Drillbook*. (Harper and Row, New York, 1960), 2 ed.

[121] M. Kim, W. S. Horton, and A. R. Bradlow, "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance," Laboratory phonology **2** (1), 125-156 (2011).

[122] Y. Maryn, P. Corthols, P. V. Cauwenberge et al., "Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels," J. Voice, 540-555 (2010).

[123] V. Reynolds, A. Buckland, J. Baley et al., "Objective assessment of pedaitric voice disorders with the acoustic voice quality index," J. Voice **26** (5), 627e621-627e627 (2012).

[124] A. Camacho, "SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music.," Ph.D. Dissertation, University of Florida, Gainesville, FL (2007).

[125] P. Boersma and D. Weenik, "Praat Manual: Voice," [available online at <http://www.fon.hum.uva.nl/praat/manual/Voice.html> (Last viewed 2018)].

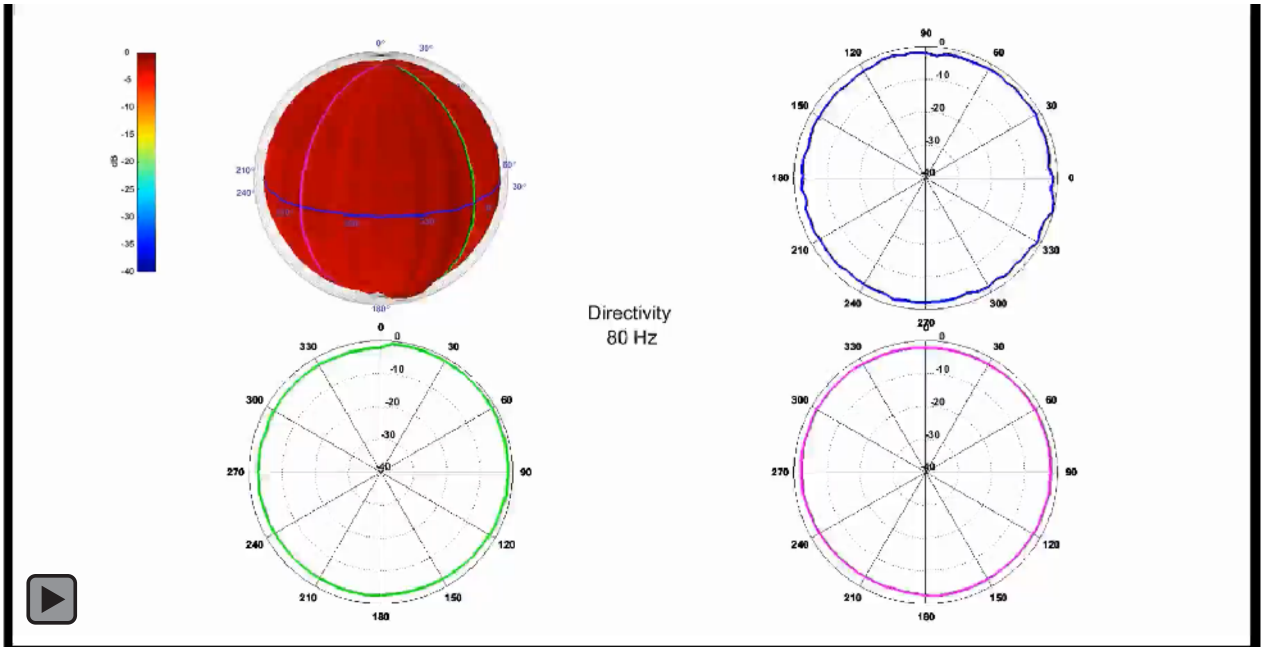
Appendix A

Directivity Animations

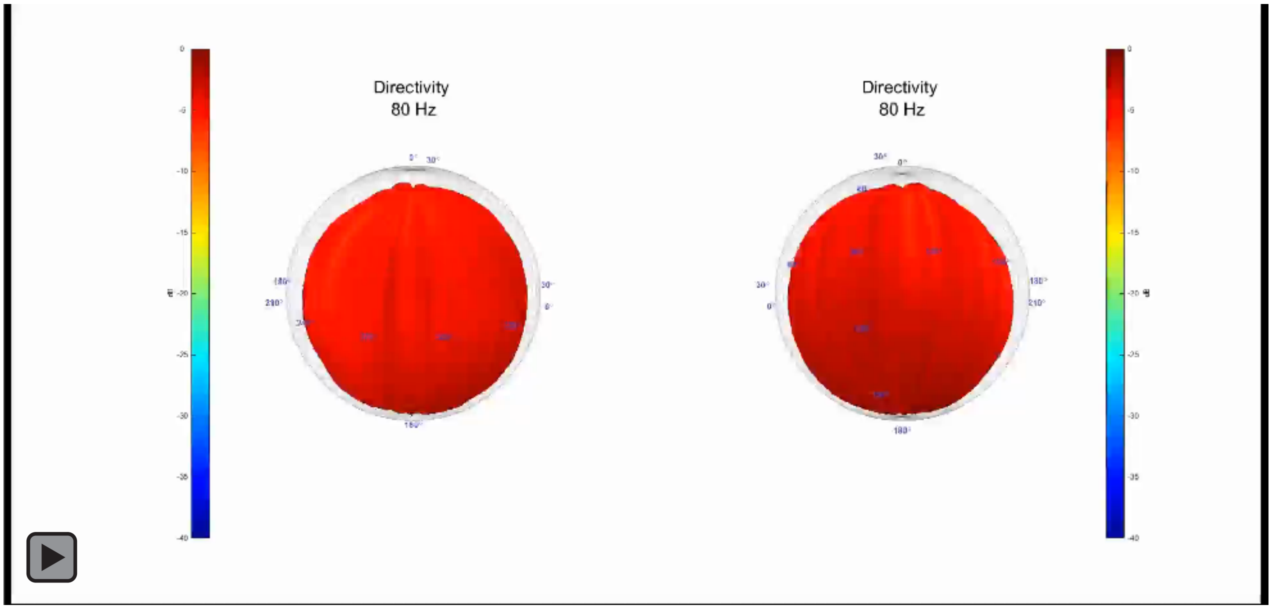
The animations included here are also hosted online at

<https://drive.google.com/drive/folders/0B2xHEZdUcroZeWRjWWRqYkFsNXM>.

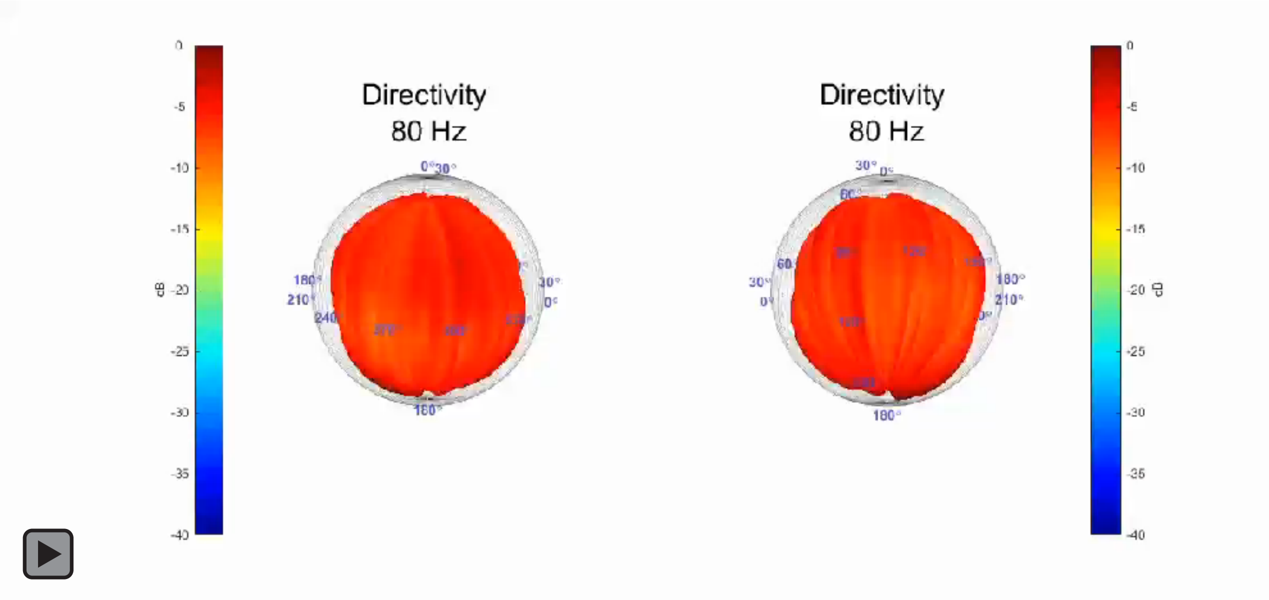
A.1 KEMAR



A.2 Female


0
-5
-10
-15
-20
-25
-30
-35
-40
dB
Directivity
80 Hz
0°
30°
180°
210°
30°
0°
180°
Directivity
80 Hz
30°
0°
30°
0°
180°
210°
180°
0
-5
-10
-15
-20
-25
-30
-35
-40
dB

A.3 Male


0
-5
-10
-15
-20
-25
-30
-35
-40
dB
Directivity
80 Hz
0°
30°
180°
210°
240°
180°
Directivity
80 Hz
30°
0°
60°
180°
210°
180°
0
-5
-10
-15
-20
-25
-30
-35
-40
dB

Appendix B

Simulation Parameters for the de Jong Concert Hall

Filter	missingfirstpulse_de Jong_Male_Stage_stagelistener.bir
Folder	Z:\Jenny Lund\2015SpSu EASE\de Jong EASE Original\Auralisation\
Level (dB)	0
Delay (ms)	50
Length (ms)	897
Window (ms)	1024
Samples	43078
Channels	2
Sampling Rate (Hz)	48000
Frame Length	8192
Frame Number	6

Listener	stagelistener
Position	X = 0, Y = 2, Z = 4 ft.
Orientation	Hor = 0 °, Ver = 0 °
HRTF Type	KEMAR Dummy
Phase Comp	AURA Method

Project	de Jong		
Town	BYU Campus		
Volume	519682 cu.ft.		
Surface	86889 sq.ft.		
Humidity	30 %		
Air Temp	23 °C		
RT Formula	Eyring		
No.	Band (Hz)	RTime (s)	Absorp
1	100	0.24	0.716
2	125	0.24	0.716
3	160	0.24	0.704
4	200	0.25	0.693

5	250	0.26	0.682
6	315	0.26	0.686
7	400	0.25	0.691
8	500	0.25	0.696
9	630	0.25	0.699
10	800	0.24	0.703
11	1000	0.24	0.707
12	1250	0.24	0.704
13	1600	0.24	0.701
14	2000	0.25	0.698
15	2500	0.24	0.698
16	3150	0.24	0.699
17	4000	0.23	0.700
18	5000	0.23	0.701
19	6300	0.22	0.702
20	8000	0.20	0.704
21	10000	0.18	0.704

Loudspeaker	stage speaker		
Active	True		
Position	X = 0, Y = 2 , Z = 4.99 ft.		
Aiming	Hor = 0 ° , Ver =-90 °, Rot = 0 °		
Delay	0 msec		
Speaker	Male Speech		
No.	Band (Hz)	SPL at 1m (dB)	Directivity (dB)
1	100	62	3
2	125	62	3
3	160	62	4
4	200	62	4
5	250	62	5
6	315	62	5
7	400	62	5
8	500	62	6
9	630	62	8
10	800	62	10
11	1000	62	11
12	1250	62	11
13	1600	62	12
14	2000	62	12
15	2500	62	13
16	3150	62	14
17	4000	62	14
18	5000	62	14
19	6300	62	15
20	8000	62	15

21	10000	62	15

Reverb Tail	None

Number of Particles	16792000 (Very High Resolution, Slow)
Length (ms)	560 (Long)
Default Scattering (%)	20
Scattering Method	Standard
Threads	8 (8 Threads)
Max. Diameter after 1s (m)	0.40
Cut off Order	10
Density Factor	10
Tail Resolution	5

Appendix C

OBRIR Modification

C.1 runmetomodifyOBRIRs_newandimproved.m

```
clear; close all; clc;
set(0,'defaultfigurewindowstyle','docked');
%% Load OBRIRs
% Ear IR
[roomname,roomIR,fs] = loadIR('etx', 'room');
room_t=0:1/fs:(length(roomIR)-1)/fs;
ears=[room_t',roomIR];

%% Mouth IR
[mouthname,mouthIR,fs] = loadIR('txt', 'mouth');
mouth_t=0:1/fs:(length(mouthIR)-1)/fs;
cheek=mouthIR(:,1);
mouth=[mouth_t',cheek];
original=fftinfo(roomIR,48e3,'Original');
%% remove mouth simulator from responses

[time,newleft,newright,FFTpic,TFpic,IRpic,IRshortpic]=...
    removemouthsimv5(ears,fs,mouth,fs);

%% Plot original and no mouth sim (Step 1).

step1=fftinfo([newleft,newright],48e3,'Step 1: Mouth Simulator Divided Out');
Step1=QuickCompare2(original,step1);

%% remove noise floor at end of OBRIR
% fitendtime=input('What time does completely-flat noise floor stop being flat? ');
% fitendindex=find(diff(sign(time-fitendtime)),1);
[nntime,newleft,newright,FITpic,WINDpic,IRpic2]=...
    removenoisefloorv3(time,...
    step1.data(:,1),step1.data(:,2));
```

```
%% Plot Step 1 and Step 2
step2=fftinfo([newleft,newright],48e3,'Step 2: No Noise Floor');
Step2=QuickCompare2(step1,step2);
%% truncate by system latency

% design Tukey Window
L=length(newleft);
r=1;
n=6e-3;
N=1e-3*fs*2;

testdatal=circshift(newleft,-n*fs);
testdatar=circshift(newright,-n*fs);

%
% w1 = gausswin(N);
w1 = tukeywin(N,r);
% w2 = chebwin(N);
% w3 = blackman(N);
% w4 = blackmanharris(N);
% w5 = gausswin(N);
% w6 = taylorwin(N);
% w7 = flattopwin(N);
% w8 = barthannwin(N);
% w9 = bohmanwin(N);
% w10 = nuttallwin(N);
% w11 = parzenwin(N);
% w12 = bartlett(N);
%
% windows=[w1,w2,w3,w4,w5,w6,w7,w8,w9,w10,w11,w12]';
% windownames='Tukey','Chebychev','Blackman','Blackman-HArris','Gaussian','Taylor','Flat Top','Bart-Hann','Bohman','Nuttall','Parzen','Bartlett';
ww=ones(1,L);
ww(1:N/2)=w1(1:N/2);
ww(end-N/2:end)=w1(end-N/2:end);
%
% % apply Tukey Window
newleft=testdatal'.*ww;
newleft=newleft(1:end-6e-3*fs)';
newright=testdatar'.*ww;
newright=newright(1:end-6e-3*fs)';

step3=fftinfo([newleft,newright],48e3,'Step 3: Truncated by 6 ms');
Step3=QuickCompare2(step2,step3);
```

```
%step3b=fftinfo([newleft,newright],48e3,'Truncated by 6 ms and Tukey Window Applied');
%Step3ab=QuickCompare2(step3a,step3b);

%% Plot Step 2 and Step 3


%% bandpass filter
lowfreq=60;
highfreq=21000;
d = fdesign.bandpass('N,F3dB1,F3dB2',10,lowfreq,highfreq,fs);
Hd = design(d,'butter');
newleft=filter(Hd,step3.data(:,1));
newright=filter(Hd,step3.data(:,2));
% kronecker delta function, bandpass filter applied
tmp=zeros(size(newleft));  tmp(1)=1;
df=fs/length(tmp);
Freq=-fs/2:df:fs/2-df;
tmp2=filter(Hd,tmp);
Tmp2=fft(tmp2);

%% Plot Step 3 and Step 4
step4=fftinfo([newleft,newright],48e3,'Step 4: Bandpass Filtered');
finaltime=step4.t';
Step4=QuickCompare2(step3,step4);

%% plot fully-modified OBRIR
IRpic3=figure;
subplot(2,1,1)
plot(step4.t,step4.data(:,1))
xlabel('Time (sec)')
ylabel('Amplitude (Pa)')
title('Left')
subplot(2,1,2)
plot(step4.t,step4.data(:,2))
xlabel('Time (sec)')
ylabel('Amplitude (Pa)')
title('Right')
suptitle('Modified IR: Filtered, Windowed, and Truncated')

%% Compare Original and Modified OBRIRs

AllSteps=QuickCompare2(original, step4);
%% write to file for SIR2.
finalleft=newleft;
```

```
finalright=newright;

% The txt file has units of Pascals. The wav file has scaled units (-1 to 1).
[FILENAME, PATHNAME, FILTERINDEX] = uiputfile('*.wav', 'Save the Modified OBRIR');
%FILENAME=input('What to call this room? ','s');
%PATHNAME='Q:\Final Validation Measurements\Room OBRIRs (for SIR2)\';
FILENAME=FILENAME(1:end-4);
fid=fopen([PATHNAME,FILENAME,'.txt'],'w');
fprintf(fid, [ 'Time' ' ' 'Left' ' ' 'Right' '\n']);
fprintf(fid, '%1.12f \t %2.14f \t %2.14f \n', [finaltime finalleft finalright]');
fclose(fid);
scalewav=input('Use a previously determined normalization factor? (enter the number for yes, 0 for no) ');
if scalewav==0
    scalewav=[(1.01*max(abs(finalleft))),(1.01*max(abs(finalright)))];
end
disp(max(scalewav))
audiowrite([PATHNAME,FILENAME,'.wav'],...
    [finalleft/max(scalewav),...
    finalright/max(scalewav)],fs,'BitsPerSample',32);
%%
disp('Load the recently saved IRs for a quick check')
%[wavname,wavIR,fsw]=loadIR('wav','room');
%[txtname,txtIR,fst]=loadIR('txt','room');
%[FILENAME, PATHNAME, ~] = uigetfile('*.wav', strtxt);
        [wavIR,fsw]=audioread([PATHNAME,FILENAME,'.wav']);
        wavname=FILENAME;
WAVIR=fftinfo(wavIR,fsw,wavname);
%TXTIR=fftinfo(txtIR,fst,txtname);
QuickCompare2(WAVIR,WAVIR);
truncateagain=input('Would you like to shorten the IR further? (1 yes, 0 no) ');

if truncateagain
   newendtime=input('At what time to stop IR? (where the wav file has a noise floor) (sec) ');
   newendindex=find(diff(sign(finaltime-newendtime)));
    newtime=finaltime(1:newendindex);
    newLIR=finalleft(1:newendindex);
    newRIR=finalright(1:newendindex);

%[FILENAME, PATHNAME, FILTERINDEX] = uiputfile('*.wav', 'Save the Modified OBRIR');
finaltime=newtime;
finalleft=newLIR;
```

```
finalright=newRIR;
fid=fopen([PATHNAME,FILENAME,'.txt'],'w');
fprintf(fid, [ 'Time' ' ' 'Left' ' ' 'Right' '\n']);
fprintf(fid, '%1.12f \t %2.14f \t %2.14f \n', [finaltime finalleft finalright]');
fclose(fid);

audiowrite([PATHNAME,FILENAME,'.wav'],...
    [finalleft/max(scalewav),...
    finalright/max(scalewav)],fs,'BitsPerSample',32);
end
%% save plots and organize files
% parentFolder=PATHNAME;
% folderName=FILENAME(1:end-4);
% mkdir(parentFolder,folderName);
% path=[parentFolder,folderName,'\'];
%
% saveas(FFTpic,[path,'FFT of IR'],'tiff');
% saveas(TFpic,[path,'Transfer Functions'],'tiff');
% saveas(IRpic,[path,'Impulse Response Mouth Sim Removed'],'tiff');
% saveas(IRshortpic,[path,'Impulse Response Mouth Sim Removed first 50 ms'],'tiff');
% saveas(FITpic,[path,'Curve Fit to IR'],'tiff');
% saveas(WINDpic,[path,'Noise Floor Window'],'tiff');
% saveas(IRpic2,[path,'Impulse Response Noise Floor Removed'],'tiff');
% saveas(IRpic3,[path,'Fully Modified Impulse Response'],'tiff');
% %savefig(Step1,[path,strrep(step1.name,':',' ')],'compact');
% saveas(Step1,[path,strrep(step1.name,':',' ')],'tiff');
% %savefig(Step2,[path,strrep(step2.name,':',' ')],'compact');
% saveas(Step2,[path,strrep(step2.name,':',' ')],'tiff');
% %savefig(Step3,[path,strrep(step3.name,':',' ')],'compact');
% saveas(Step3,[path,strrep(step3.name,':',' ')],'tiff');
% %savefig(Step4,[path,strrep(step4.name,':',' ')],'compact');
% saveas(Step4,[path,strrep(step4.name,':',' ')],'tiff');
% %savefig(AllSteps,[path,'All Steps'],'compact');
% saveas(AllSteps,[path,'All Steps'],'tiff');
% clear FFTpic TFpic IRpic IRshortpic FITpic WINDpic IRpic2 IRpic3;
% clear Step1 Step2 Step3 Step4 AllSteps Step3ab;
% % save workspace variables
% save(FILENAME(1:end-4))
% %movefile([FILENAME(1:end-4),'.mat'],[path,FILENAME(1:end-4),'.mat'])
% copyfile([PATHNAME,FILENAME(1:end-4),'.txt'],[path,FILENAME(1:end-4),'.txt'])
% copyfile([PATHNAME,FILENAME(1:end-4),'.wav'],[path,FILENAME(1:end-4),'.wav'])
% % copyfile([PATHNAME,roomname], [path,'(original)_',roomname])
```

```
display ('Finished!')
```

C.2 loadIR.m

```

function [FILENAME,IR,fs] = loadIR(file_type, IR_type)

if strcmp(file_type,'wav')

    if strcmp(IR_type,'room')
        strtxt='Open Room OBRIR (*.wav)';
    elseif strcmp(IR_type,'RTCS')
        strtxt='Open Unfiltered RTCS OBRIR (*.wav)';
    elseif strcmp(IR_type,'RTCSf')
        strtxt='Open Filtered RTCS OBRIR (*.wav)';
    elseif strcmp(IR_type,'filter')
        strtxt='Open Filter Wavform (*.wav)';
    end
        [FILENAME, PATHNAME, ~] = uigetfile('*.wav', strtxt);
        [IR,fs]=audioread([PATHNAME,FILENAME]);

elseif strcmp(file_type,'etx')
    if strcmp(IR_type,'room')
        strtxt='Open Room OBRIR (*.etx)';
    elseif strcmp(IR_type,'RTCS')
        strtxt='Open RTCS OBRIR (*.etx)';
    elseif strcmp(IR_type,'RTCSf')
        strtxt='Open Filtered RTCS OBRIR (*.etx)';
    elseif strcmp(IR_type,'mouth')
        strtxt='Open anechoic OBRIR (*.etx)';
    end
    [FILENAME, PATHNAME, ~] = uigetfile('*.etx', strtxt);
    % num_ch=input('How many channels in the etx file? ');
    num_ch=3;
    fID=fopen([PATHNAME,FILENAME]);

    if num_ch == 2
        data = textscan(fID,'%f %f %f %f','HeaderLines',22);
    elseif num_ch == 3
        data = textscan(fID,'%f %f %f %f %f','HeaderLines',22);
    elseif num_ch == 4
        data = textscan(fID,'%f %f %f %f %f %f','HeaderLines',22);
    end
    fclose(fID);
    left=data{:,2};       % Left KEMAR ear microphone% time data
```

```
        right=data{:,3};      % Right KEMAR ear microphone
        if strcmp(IR_type,'mouth')
            mouth=data{:,4};
        end
        fs=48e3; % dt=1/fs;      % sampling rate and time per sample
        if strcmp(IR_type,'mouth')
            IR=[left,right,mouth];
        else
            IR=[left,right];
        end
    elseif strcmp(file_type,'txt')
        if strcmp(IR_type,'room')
            strtxt='Open Room OBRIR (*.txt)';
        elseif strcmp(IR_type,'RTCS')
            strtxt='Open RTCS OBRIR  (*.txt)';
        elseif strcmp(IR_type,'RTCSf')
            strtxt='Open Filtered RTCS OBRIR (*.txt)';
        elseif strcmp(IR_type,'mouth')
            strtxt='Open OBRIR containing cheek mic info (*.txt)';
        end
        [FILENAME, PATHNAME, ~] = uigetfile('*.txt', strtxt);
        fID=fopen([PATHNAME,FILENAME]);
        data = textscan(fID,'%f %f %f %f','HeaderLines',1);
        fclose(fID);
        t=data{:,1};
        left=data{:,2};      % Left KEMAR ear microphone% time data
        right=data{:,3};      % Right KEMAR ear microphone
        fs=48e3;
        IR=[left,right];
        if strcmp(IR_type,'mouth')
            IR=left;
        end
        FILENAME=strrep(FILENAME,'_',' ');
    end






    end
```

C.3 removemouthsimv5

```
function [time,trunc_Left_IR,trunc_Right_IR,p1,p2,p5,p6]=removemouthsimv3(earIR,earfs,mouthIR,mouthfs)
% IR is a three-column array containing time data, left and right ear
% microphone data for an OBRIR.
% IRfs is the sampling rate for that impulse response.
% mouthmic is an impulse response (two-column array) for a microphone close
% to the mouth.
% mouthfs is the sampling rate for that impulse response.
%% get data from input
t=earIR(:,1);
left=earIR(:,2);
right=earIR(:,3);
mouth=mouthIR(:,2);

if earfs~=mouthfs
sprintf('Error: Sampling Rates do not Match');
end
fs=earfs;

l1=length(earIR);
l2=length(mouthIR);
N=l1+l2;
N=2*l1;
% pad array with zeros such that each IR is the same length
newmouth=padarray(mouth,N-l2,0,'post');
newleft=padarray(left,N-l1,0,'post');
newright=padarray(right,N-l1,0,'post');

% Identify direct sound in Mouth
twindow=tukeywin(5e-3*fs,0.1);
twindow_ext=padarray(twindow,length(newmouth)-length(twindow),0,'post');
mouth_direct=newmouth.*twindow_ext;

%% perform Fourier Transform

Left=fft(newleft);
Right=fft(newright);
% smooth mouth spectrum in 6th octave bands

    data_current=mouth_direct;

    AVG          = fft(data_current);
    [AVG, is_even] = both2single_sided_spectrum(AVG);
    AVG_SM       = zeros(size(AVG));
    numChannel=1;
    for idx_c = 1:numChannel
        AVG_SM(:, idx_c) = fract_oct_smooth(AVG(:, idx_c), 'welti', fs, 24);
    end
    AVG_SM = single2both_sided_spectrum(AVG_SM, is_even);
    average_sm = ifft(AVG_SM, 'symmetric');
    %disp('Applying min-phase after smoothing averaged transfer function')
    average_sm = phase_manipulation(average_sm,fs, 'min_phase',2);

    data_current = average_sm;

Mouth=fft(data_current);


df=earfs/length(Left);
Freq=-fs/2:df:fs/2-df;
p1=figure;
subplot(3,1,1)
plot(Freq,20*log10(abs(fftshift(Left/20e-6))))
xlabel('Frequency (Hz)')
ylabel('dB')
title('Left')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;
subplot(3,1,2)
plot(Freq,20*log10(abs(fftshift(Right/20e-6))))
xlabel('Frequency (Hz)')
ylabel('Relative dB')
title('Right')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;
subplot(3,1,3)
plot(Freq,20*log10(abs(fftshift(Mouth/20e-6))))
xlabel('Frequency (Hz)')
ylabel('Relative dB')
title('Mouth')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;


figure
subplot(2,1,1)
```

```
plot(Freq,20*log10(abs(fftshift(Left/20e-6))),Freq,20*log10(abs(fftshift(Mouth/20e-6))))
xlabel('Frequency (Hz)')
ylabel('Relative dB')
legend('left ear','mouth')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;
subplot(2,1,2)
plot(Freq,20*log10(abs(fftshift(Right/20e-6))),Freq,20*log10(abs(fftshift(Mouth/20e-6))))
xlabel('Frequency (Hz)')
ylabel('Relative dB')
legend('right ear','mouth')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;


%% Apply 12 dB / oct roll off to Ear IRs

% plot(Freq,20*log10(abs(fftshift(Left/20e-6))),Freq,20*log10(abs(fftshift(Mouth/20e-6))),...
%     Freq,15-20*log10(abs(fftshift(Try1*1e-12))))
% xlabel('Frequency (Hz)')
% ylabel('Relative dB')
% legend('left','mouth','12dB/oct roll off')
% xlim([1 fs/2])
% set(gca,'xscale','log')
% grid on;

% fllim=133;
% n1=find(diff(sign(Freq-(-fllim))));
% n2=find(diff(sign(Freq-fllim)));
% test_dataL=fftshift(Left);
% test_dataR=fftshift(Right);
% try1=Freq.^2/Freq(n2)^2; %Try1=fftshift(fft(try1));
% test_dataL(n1:n2)=test_dataL(n1:n2).*try1(n1:n2)';
% test_dataR(n1:n2)=test_dataR(n1:n2).*try1(n1:n2)';
cont = 0;
while cont ==0
    order=input('What order filter to use? (try 3 to start)');
    fllim=input('Which Frequency to use as Corner Frequency for High Pass Filter? (near 150 Hz)');
% fllim=148;
d = fdesign.highpass('N,F3dB',order,fllim,fs);
```

```
%M = designmethods(d,'full','SystemObject',true)
Hd = design(d,'butter');
test_dataL=filter(Hd,newleft);
test_dataR=filter(Hd,newright);


%{
dfm=earfs/length(mouth);
Freqm=-fs/2:dfm:fs/2-df;

figure
plot(Freqm,20*log10(abs(fftshift(fft(mouth))/20e-6)),...
    Freq,20*log10(abs(fftshift(fft(newmouth)/20e-6))),...
    Freq,20*log10(abs(fftshift(fft(mouth_direct))/20e-6)),...
    Freq,20*log10(abs(fftshift(Mouth)/20e-6)),...
    Freq,(107-80)+20*log10(Freq.^2),'linewidth',2);
xlabel('Frequency (Hz)')
ylabel('Relative dB')
legend('Original','Zero-Padded','Direct Sound Only','Direct Sound Smoothed 6th oct')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;
title('Cheek Mic IR')

%%
figure
plot(Freq,20*log10(abs(fftshift(Mouth)/20e-6)),...
    Freq,(107-80)+20*log10(Freq.^2));
xlabel('Frequency (Hz)')
ylabel('Relative dB')
legend('Modified Cheek Mic IR','12 dB / oct','location','southeast')
xlim([20 500])
set(gca,'xscale','log')
set(gca,'xtick',[20,40,60,80,100,200,400,600])

grid on;
title('Cheek Mic IR')
%}
%%

figure
subplot(2,1,1)
plot(Freq,20*log10(abs(fftshift(Left/20e-6))),Freq,20*log10(abs(fftshift(Mouth/20e-6))),...
Freq,20*log10(abs(fftshift(fft(test_dataL/20e-6)))));
xlabel('Frequency (Hz)')
```

```
ylabel('Relative dB')
legend('left','mouth','left * high pass')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;
subplot(2,1,2)
plot(Freq,20*log10(abs(fftshift(Right/20e-6))),Freq,20*log10(abs(fftshift(Mouth/20e-6))),...
Freq,20*log10(abs(fftshift(fft(test_dataR/20e-6)))));
xlabel('Frequency (Hz)')
ylabel('Relative dB')
legend('right','mouth','right * high pass')
xlim([1 fs/2])
set(gca,'xscale','log')
grid on;


newLeft=fft(test_dataL);
newRight=fft(test_dataR);
%% Compute Transfer Functions from Auto- and Cross- Spectra
% No blocking/averaging necessary as IRs were used instead of continous data.

Left_TF=(conj(Mouth).*newLeft)./(conj(Mouth).*Mouth);
Right_TF=(conj(Mouth).*newRight)./(conj(Mouth).*Mouth);

%    AVG          = Left_TF;
%    [AVG, is_even] = both2single_sided_spectrum(AVG);
%    AVG_SM       = zeros(size(AVG));
%    numChannel=1;
%    for idx_c = 1:numChannel
%        AVG_SM(:, idx_c) = fract_oct_smooth(AVG(:, idx_c), 'welti', fs, 48);
%    end
%    AVG_SM = single2both_sided_spectrum(AVG_SM, is_even);
%    average_sm = ifft(AVG_SM, 'symmetric');
%    %disp('Applying min-phase after smoothing averaged transfer function')
%    % data.average_sm = phase_manipulation(data.average_sm, s.fs, 'min_phase', s.Nfft_double);
%
%    data_current = average_sm;
% Left_TF=fft(data_current);
%
%
%    AVG          = Right_TF;
%    [AVG, is_even] = both2single_sided_spectrum(AVG);
%    AVG_SM       = zeros(size(AVG));
%    numChannel=1;
%    for idx_c = 1:numChannel
%        AVG_SM(:, idx_c) = fract_oct_smooth(AVG(:, idx_c), 'welti', fs, 48);
%    end
%    AVG_SM = single2both_sided_spectrum(AVG_SM, is_even);
%    average_sm = ifft(AVG_SM, 'symmetric');
%    %disp('Applying min-phase after smoothing averaged transfer function')
%    % data.average_sm = phase_manipulation(data.average_sm, s.fs, 'min_phase', s.Nfft_double);
%
%    data_current = average_sm;
%
%    Right_TF=fft(data_current);
p2=figure;
subplot(2,1,1)
plot(Freq,(20*log10(abs(fftshift(Left_TF/20e-6)))))
xlim([1 fs/2])
set(gca,'xscale','log')
xlabel('Frequency (Hz)')
ylabel('Relative dB')
title('Mouth to Left')
grid on;
subplot(2,1,2)
plot(Freq,(20*log10(abs(fftshift(Right_TF/20e-6)))))
xlim([1 fs/2])
set(gca,'xscale','log')
xlabel('Frequency (Hz)')
ylabel('Relative dB')
title('Mouth to Right')
grid on;
suptitle('Transfer Function Magnitude')

cont = input('Happy with this result? (1 for yes, 0 to modify) ');
end
%% Inverse Fourier Transform
Left_IR=ifft(Left_TF,'symmetric');
Right_IR=ifft(Right_TF,'symmetric');
trunc_Left_IR=Left_IR(1:l1);
trunc_Right_IR=Right_IR(1:l1);

p5=figure;
subplot(2,1,1)
plot(t,10*log10(left.^2/(20e-6)^2),t,10*log10(trunc_Left_IR.^2/(20e-6)^2));
xlim([min(t) max(t)])
legend('Original','Modified')
xlabel('Time (sec)')
ylabel('dB ref. 20\muPa')
```

```
title('Left')
subplot(2,1,2)
plot(t,10*log10(right.^2/(20e-6)^2),t,10*log10(trunc_Right_IR.^2/(20e-6)^2));
legend('Original','Modified')
xlim([min(t) max(t)])
xlabel('Time (sec)')
ylabel('dB ref. 20\muPa')
xlabel('Time (sec)')
title('Right')
suptitle('Impulse Responses Mouth to Ear Mic')

p6=figure;
subplot(2,1,1)
plot(t,10*log10(left.^2/(20e-6)^2),t,10*log10(trunc_Left_IR.^2/(20e-6)^2));
xlim([min(t) 0.05])
legend('Original','Modified')
xlabel('Time (sec)')
ylabel('dB ref. 20\muPa')
title('Left')
subplot(2,1,2)
plot(t,10*log10(right.^2/(20e-6)^2),t,10*log10(trunc_Right_IR.^2/(20e-6)^2));
xlim([min(t) 0.05])
legend('Original','Modified')
xlabel('Time (sec)')
ylabel('dB ref. 20\muPa')
xlabel('Time (sec)')
title('Right')
suptitle('Impulse Responses Mouth to Ear Mic: First 50 ms')

time=t;
end
```

C.4 QuickCompare2.m

```
function [f]=QuickCompare2(s1,s2)
f=figure;
for ch=1:2
    subplot(3,2,ch)
    plot(s1.freq2s,s1.fftlog(:,ch),s2.freq2s,s2.fftlog(:,ch));
        set(gca,'xscale','log')
        xlim([50 11e3])
```

```
     xlabel('Frequency (Hz)')
     ylabel('Amplitude (dB)')
      grid on;
      legend(s1.name,s2.name,'location','south')
      title(['Channel: ',num2str(ch)])
    subplot(3,2,ch+2)
plot(s1.freq2s,fftshift(s1.fftphase(:,ch)),s2.freq2s,fftshift(s2.fftphase(:,ch)))
      set(gca,'xscale','log')
      xlim([0 11e3])
      ylim([-360 360])
      xlabel('Frequency (Hz)')
      ylabel('Unwrapped Phase (Deg)')
      grid on;
      legend(s1.name,s2.name,'location','south')
    subplot(3,2,ch+4)
    plot(s1.t,s1.log(:,ch),s2.t,s2.log(:,ch))
    xlabel('Time (sec)')
    ylabel('Amplitude (dB)')
    grid on;
    legend(s1.name,s2.name)
end

end
```

C.5 removenoisefloorv3

```
function [newtime,newLIR,newRIR,p1,p2,p3]=removenoisefloorv2(time,left,right)

%% fit a curve to data profile (One Channel Only)

cont=0;
mpd=input('Choose MinPeakDistance: (suggest 500 to start) ');
[pks,locs] = findpeaks(20*log10(abs(left/20e-6)),'MinPeakDistance',mpd);
pkstime=time(locs);
pksinterp=interp1(pkstime,pks,time);
while cont==0

% Design a fit to gradually remove noise floor
figure;
plot(time,20*log10(abs(right)/20e-6),time,pksinterp,'r','linewidth',2)
xlabel('Time (sec)')
ylabel('Amplitude (dB)')
legend('Full Data','Peaks')
```

```
grid on;

endtime=input('What time to use to find slope? (sec) ');
endindex=find(diff(sign(pkstime-endtime)),1);
noisefloor=pks(endindex);
startindex=find(diff(sign(pks-(noisefloor+20))),1);
starttime=pkstime(startindex);
fitendtime=input('What time does completely-flat noise floor stop being flat? ');
fitendindex=find(diff(sign(pkstime-fitendtime)),1);
f1=polyfit(pkstime(1:fitendindex),pks(1:fitendindex),9);
finalfit1 = polyval(f1,time);
%ffit1=f1.a*exp(f1.b*newtime);
%finalfit1=f1.a*exp(f1.b*time);


f2=fit(pkstime(startindex:endindex),pks(startindex:endindex),'poly1');
ffit2=f2.p1*pkstime+f2.p2;
finalfit2=f2.p1*time+f2.p2;
goal_slope=(pks(endindex)-pks(startindex))/(pkstime(endindex)-pkstime(startindex));
goal_slope=f2.p1;


p1=figure;
plot(time,20*log10(abs(left)/20e-6),...
    pkstime,ffit2,'cyan',time,pksinterp,'r',time,finalfit1,'g','linewidth',2)
legend('Full Data',...
    'Noise Floor Removal Curve','Sparsed Data','Polynomial Fit to Data')
line([starttime starttime],[-100 100],'color','k')
line([endtime endtime],[-100 100],'color','k')
text(starttime,90,'Region Used for Fitting Curve for Noise Removal')
hold on;
xlabel('Time (sec)')
ylabel('Amplitude (dB)')
title('Data and Fitted Line for Noise Removal')
xlim([0 fitendtime])
ylim([-100 100])
cont=input('Are you happy with this fit for noise removal? (1 for yes, 0 for no) ');

end
%% compute window
%endtime=input('Estimate the time the noise floor begins: (When to start the decay window)');
customwindowdb=20*log10(ones(1,length(left)));%/20e-6);
starti=find(diff(sign(time-endtime)),1);
% smoothfactor=20*log10(ones(1,1)/20e-6)-finalfit2(starti); % smoothing factor
% smoothfactor=0;

% endi=find(diff(sign(time-4.5)),1);
endi=length(time);
N=round(0.05/(time(2)-time(1)));
for n=starti:endi-N-1

    startindex=n-N;
    %starttime=time(startindex);
    %startindex=find(diff(sign(newtime-starttime)),1);
    endindex=n+N;
    %endtime=time(endindex);
    %endindex=find(diff(sign(newtime-endtime)),1);

    current_slope(n-starti+1)=(finalfit1(endindex)-finalfit1(startindex))/(time(endindex)-time(startindex));
    current_value(n-starti+1)=mean(pksinterp(startindex:endindex));
    goal_value(n-starti+1)=finalfit2(n);

end
slope_diff=1; arb_factor=0;
%while slope_diff>0.5
% customwindowdb(starti:end)=time(1:end-starti+1).*f2.p1+20*log10(1/20e-6);
% customwindowdb(starti:n)=time(1:n-starti+1)'.*(arb_factor+(goal_slope-current_slope));%+20*log10(1/20e-6);
customwindowdb(starti:n)=(goal_value-current_value);
customwindow=10.^(customwindowdb/20);%*20e-6;



% finalendtime=input('At what time to stop IR? (at least 100 dB down from start) ');
% newendindex=find(diff(sign(time-finalendtime)),1);
% if finalendtime > time(end)
%     newendindex=length(time);
% end
newendindex=n-starti;
%{
n=20*log10(1/20e-6)-40;
newendindex=find(diff(sign(customwindowdb-n)),1);
while isempty(newendindex)
    n=n-1;
    newendindex=find(diff(sign(customwindowdb-n)),1);
    if n<=0
       newendindex=length(customwindow);
    end
end
%}
newLIR=left.*customwindow';
newRIR=right.*customwindow';
newIR=[newLIR,newRIR];

% newIR=newIR(1:newendindex,:);
% newtime=time(1:newendindex);
% newLIR=newLIR(1:newendindex);
% newRIR=newRIR(1:newendindex);

lmax=max(10*log10(newLIR.^2/20e-6));
rmax=max(10*log10(newRIR.^2/20e-6));
mmax=(min(lmax,rmax));
newenddb=mmax-120;
newenddbindex=find(diff(sign(ffit2-newenddb)),1);
newenddbtime=pkstime(newenddbindex);

if isempty(newenddbtime)
    newendindex=length(time);
else
newendindex=find(diff(sign(time-newenddbtime)),1);
end


newtime=time(1:length(newLIR));
%end

[pks2,locs] = findpeaks(20*log10(abs(newLIR/20e-6)),'MinPeakDistance',mpd);
pks2time=time(locs);
pks2interp=interp1(pks2time,pks2,newtime);
f3=fit(pks2time,pks2,'poly1');
ffit3=f3.p1*pkstime+f3.p2;
finalfit3=f3.p1*newtime+f3.p2;
f3_slope=f3.p1;


slope_diff=abs(f3.p1-f2.p1)
%arb_factor=arb_factor+1;
% display(f3.p1);
% display(goal_slope);
% display(f2.p1);
%end
display(arb_factor);
p2=figure;
plot(time,20*log10(abs(left/20e-6)),time,customwindowdb)
xlabel('Time (sec)')
```

```
xlim([0 fitendtime])
ylabel('Amplitude (dB)')
title('Window Function')

p3=figure;
subplot(2,1,1)
plot(time,10*log10(left.^2/(20e-6)^2),newtime,10*log10(newLIR.^2/(20e-6)^2),pkstime,ffit2,'cyan',newtime,finalfit3)
xlim([min(newtime) max(newtime)])
xlabel('Time (sec)')
ylabel('Amplitude (dB)')
legend('Original','Modified','original linear fit','new linear fit')
grid on;
title('Left')
subplot(2,1,2)
plot(time,10*log10(right.^2/(20e-6)^2),newtime,10*log10(newRIR.^2/(20e-6)^2),pkstime,ffit2,'cyan',newtime,finalfit3)
xlim([min(newtime) max(newtime)])
xlabel('Time (sec)')
ylabel('Amplitude (dB)')
legend('Original','Modified','original linear fit','new linear fit')
title('Right')
grid on;
suptitle('Modified IR: Noise Floor Removed')

end
```

C.6 Fftinfo.m

```
function [ vector_info ] = fftinfo( time_domain_vector ,fs,filename)
%UNTITLED Summary of this function goes here
%   Detailed explanation goes here
vector_info=struct;
vector_info.name=filename;
vector_info.data=time_domain_vector;
vector_info.log=10*log10(time_domain_vector.^2);
vector_info.t=0:1/fs:(length(time_domain_vector)-1)/fs;
df=fs/length(time_domain_vector); dt=1/fs;
vector_info.freq2s=-fs/2:df:(fs/2-df);
vector_info.fft=dt*fftshift(fft(time_domain_vector,[],1),1);
vector_info.fftlog=10*log10(abs(vector_info.fft/20e-6).^2);
vector_info.fftphase=180/pi*unwrap(angle(vector_info.fft));
```

```
end
```

Appendix D

Inversion Filter Computation and Simulation Results

As in Section 4.2.4.3, a visualization of the compensation filter computation and subsequent simulation of RTCS performance with inclusion of the filter for each of the rooms being considered are presented here.

D.1 Reverberation Chamber 00 Wedges

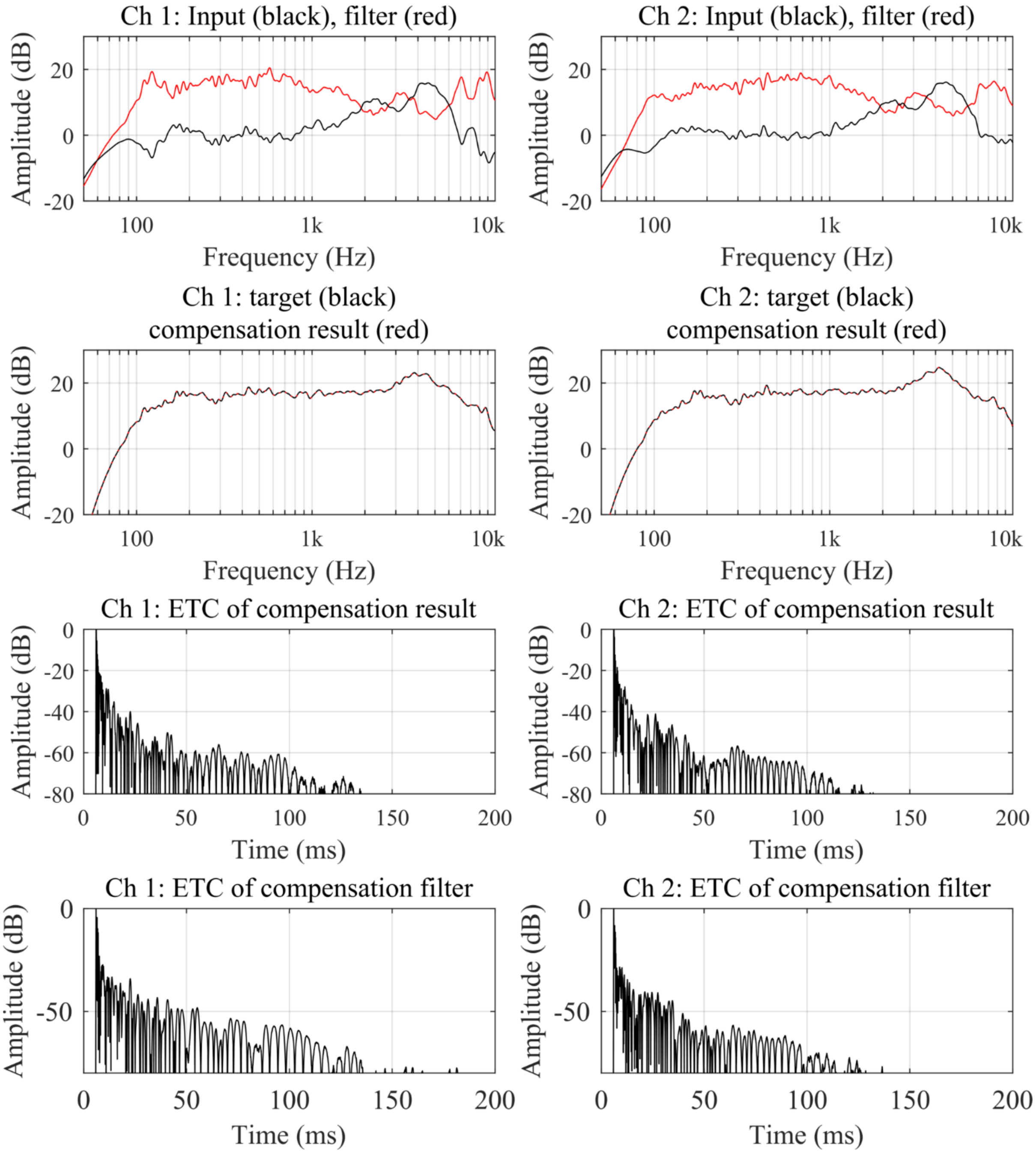
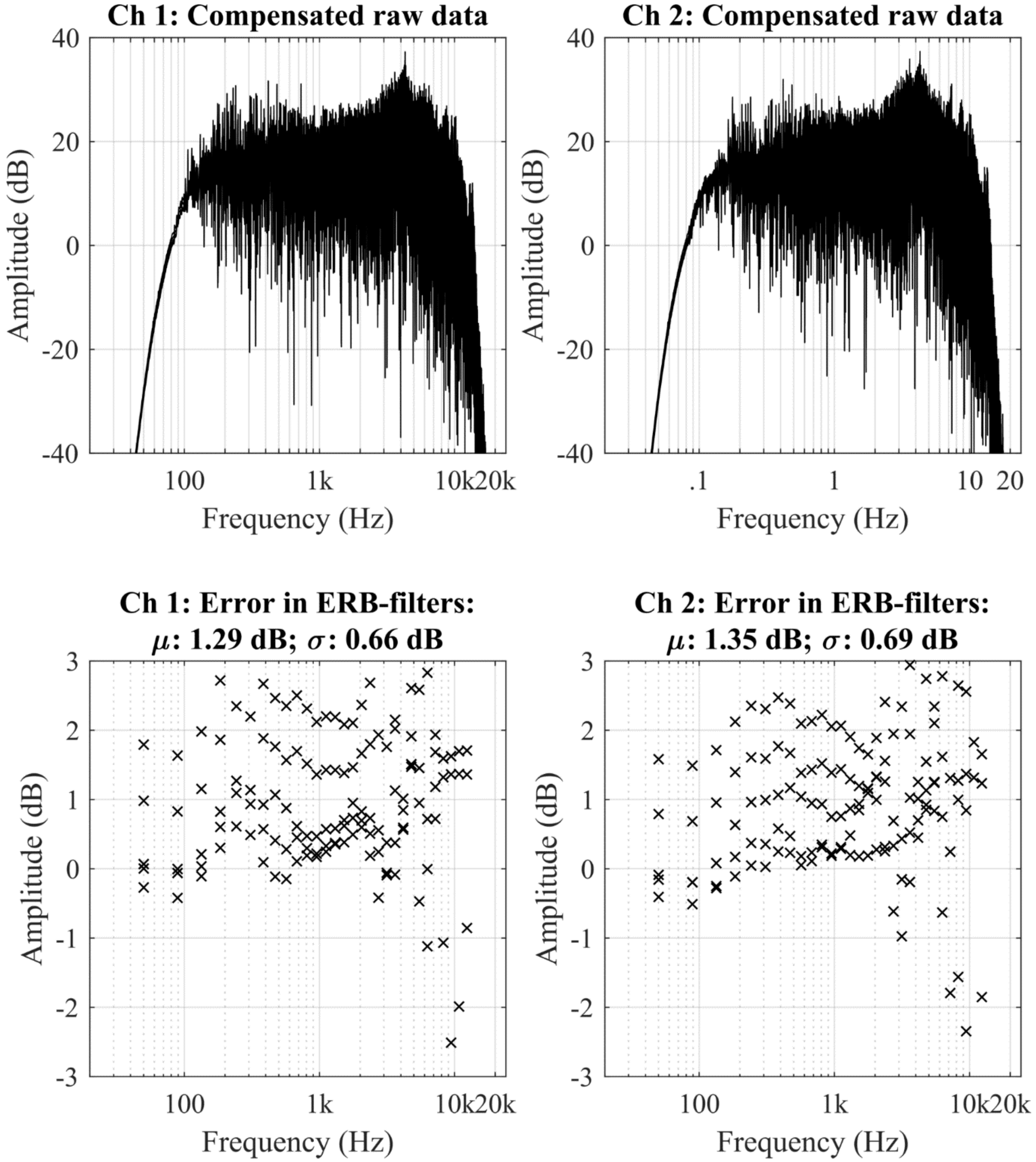


Figure D.1. Compensation filter for Reverberation Chamber 00 wedges.



Ch 1: Compensated raw data

Ch 2: Compensated raw data

Ch 1: Error in ERB-filters: μ: 1.29 dB; σ: 0.66 dB

Ch 2: Error in ERB-filters: μ: 1.35 dB; σ: 0.69 dB

Figure D.2. Simulated Errors for Compensation Filter inclusion, Reverberation Chamber 00 wedges.

D.2 Reverberation Chamber 04 Wedges

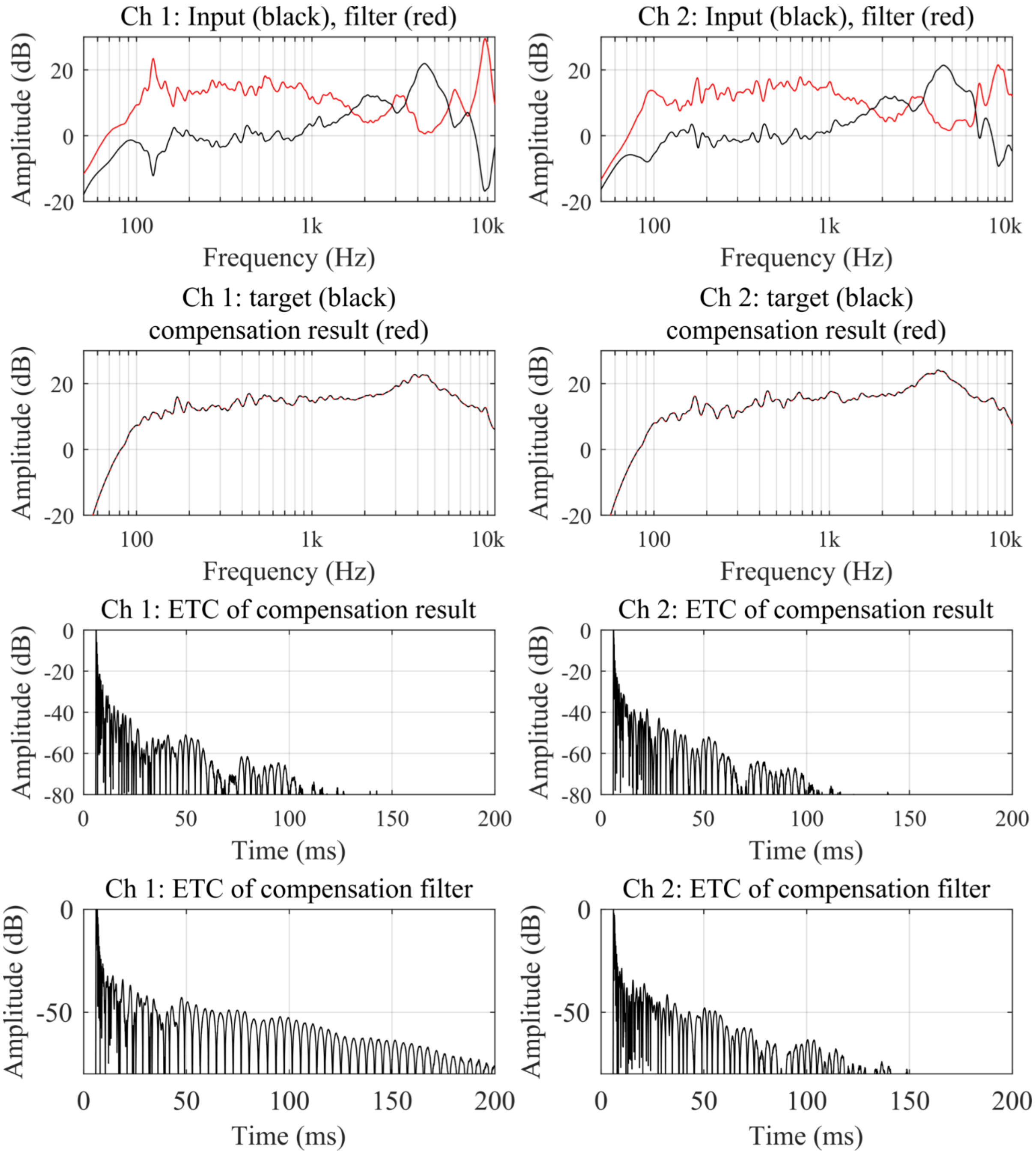


Figure D.3. Compensation filter for Reverberation Chamber 04 wedges.

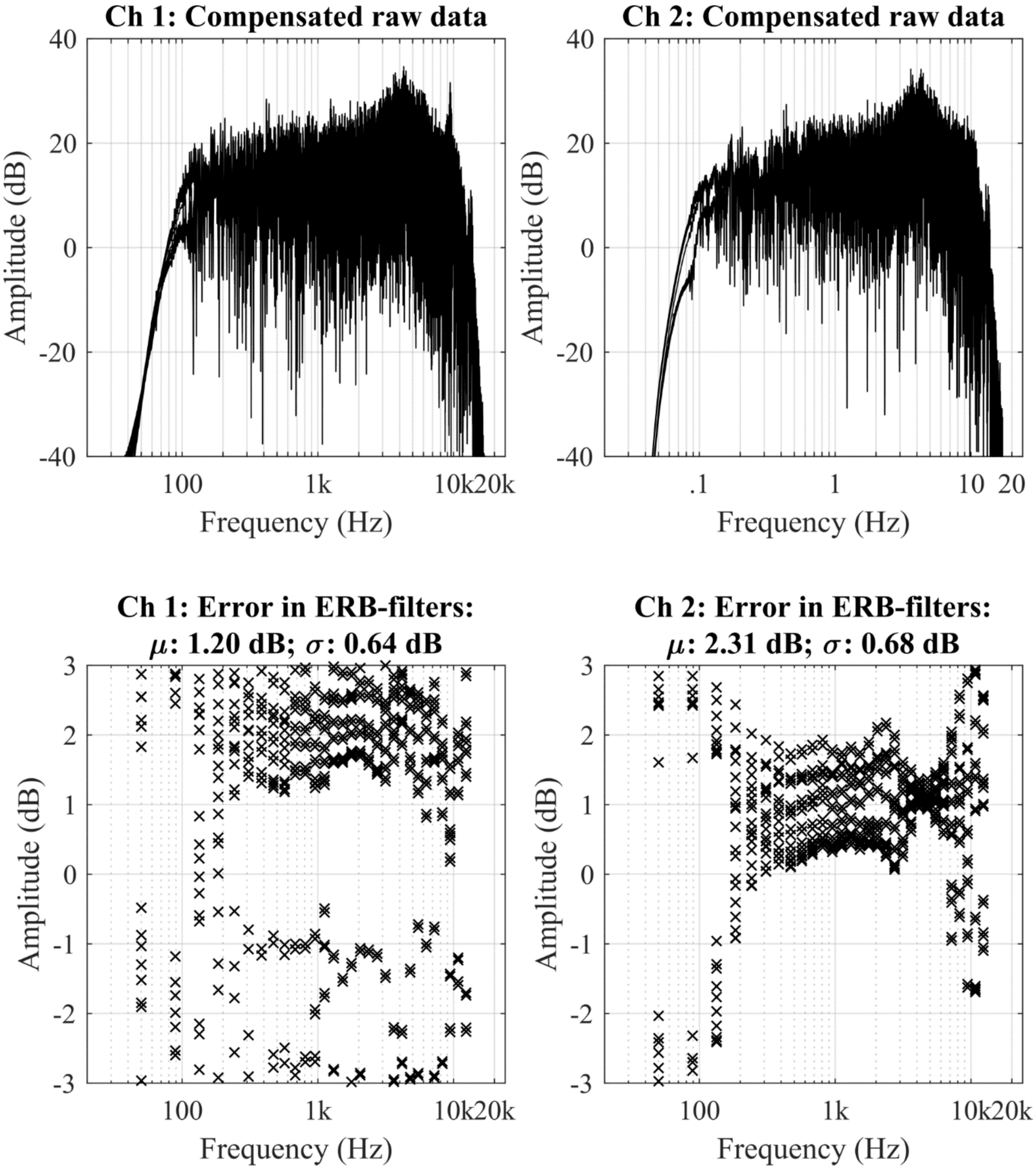


Figure D.4. Simulated Errors for Compensation Filter inclusion, Reverberation Chamber 04 wedges.

D.3 Reverberation Chamber 08 Wedges

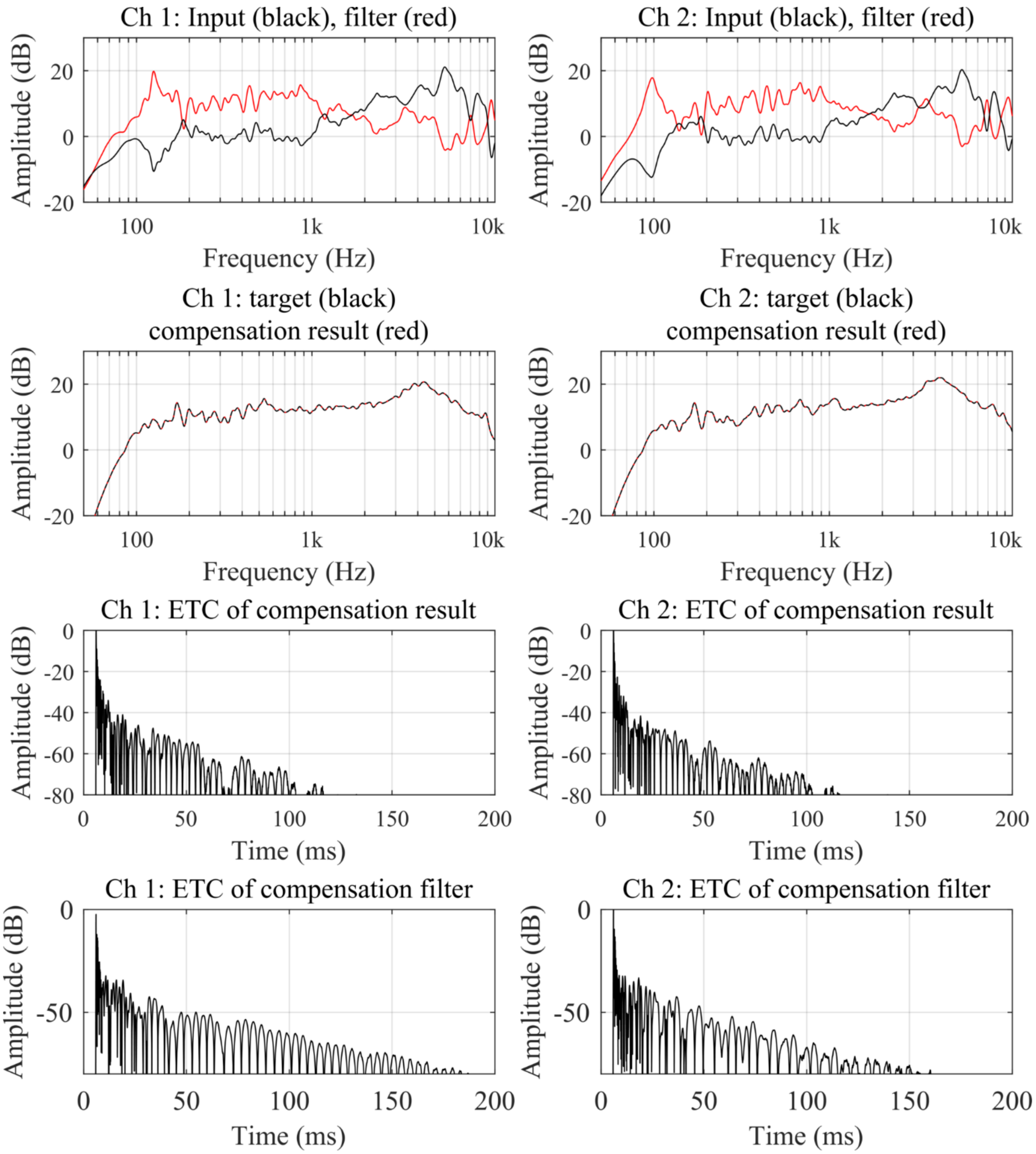


Figure D.5. Compensation filter for Reverberation Chamber 08 wedges.

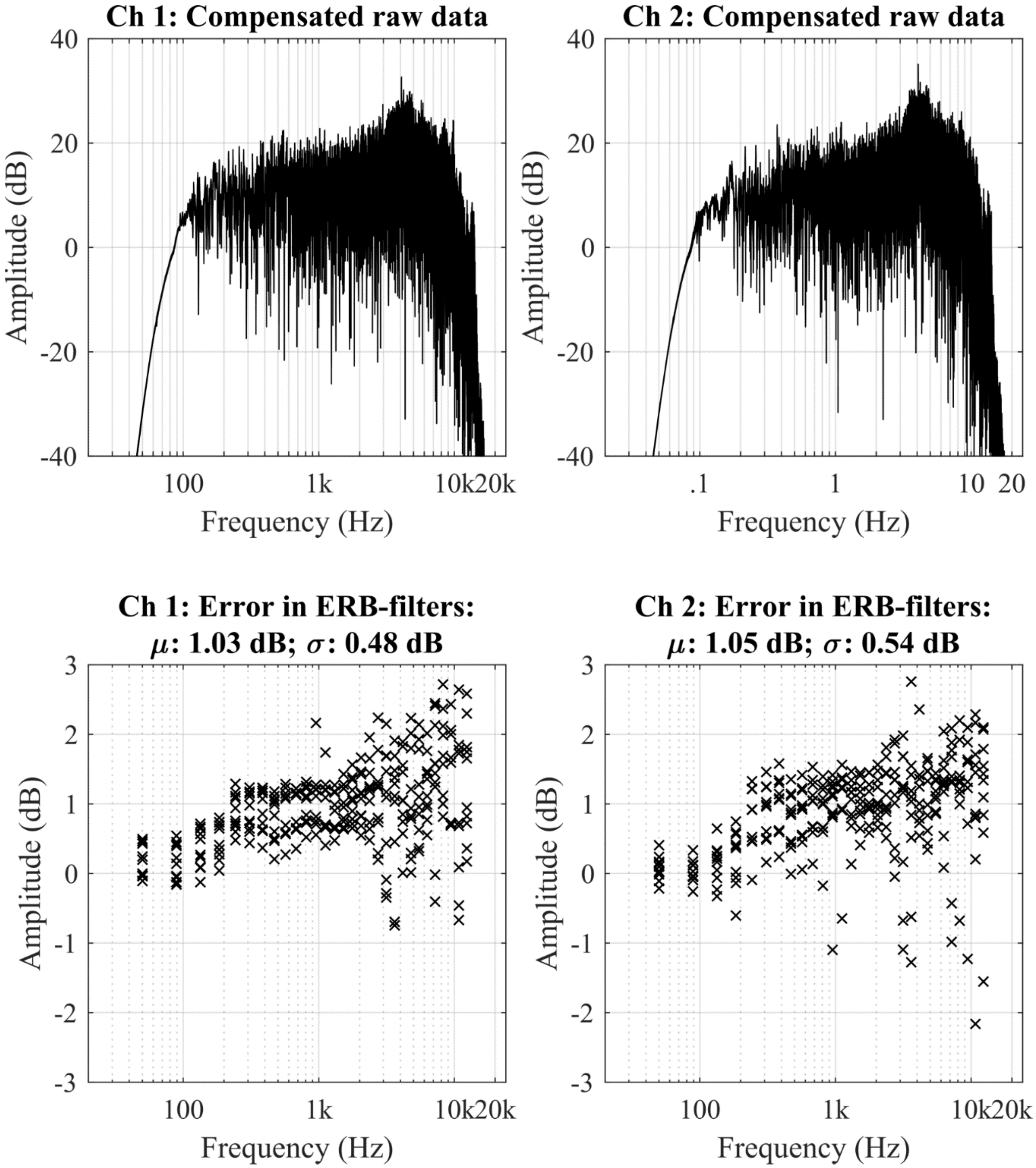


Figure D.6. Simulated Errors for Compensation Filter inclusion, Reverberation Chamber 08 wedges.

D.4 Reverberation Chamber 16 Wedges

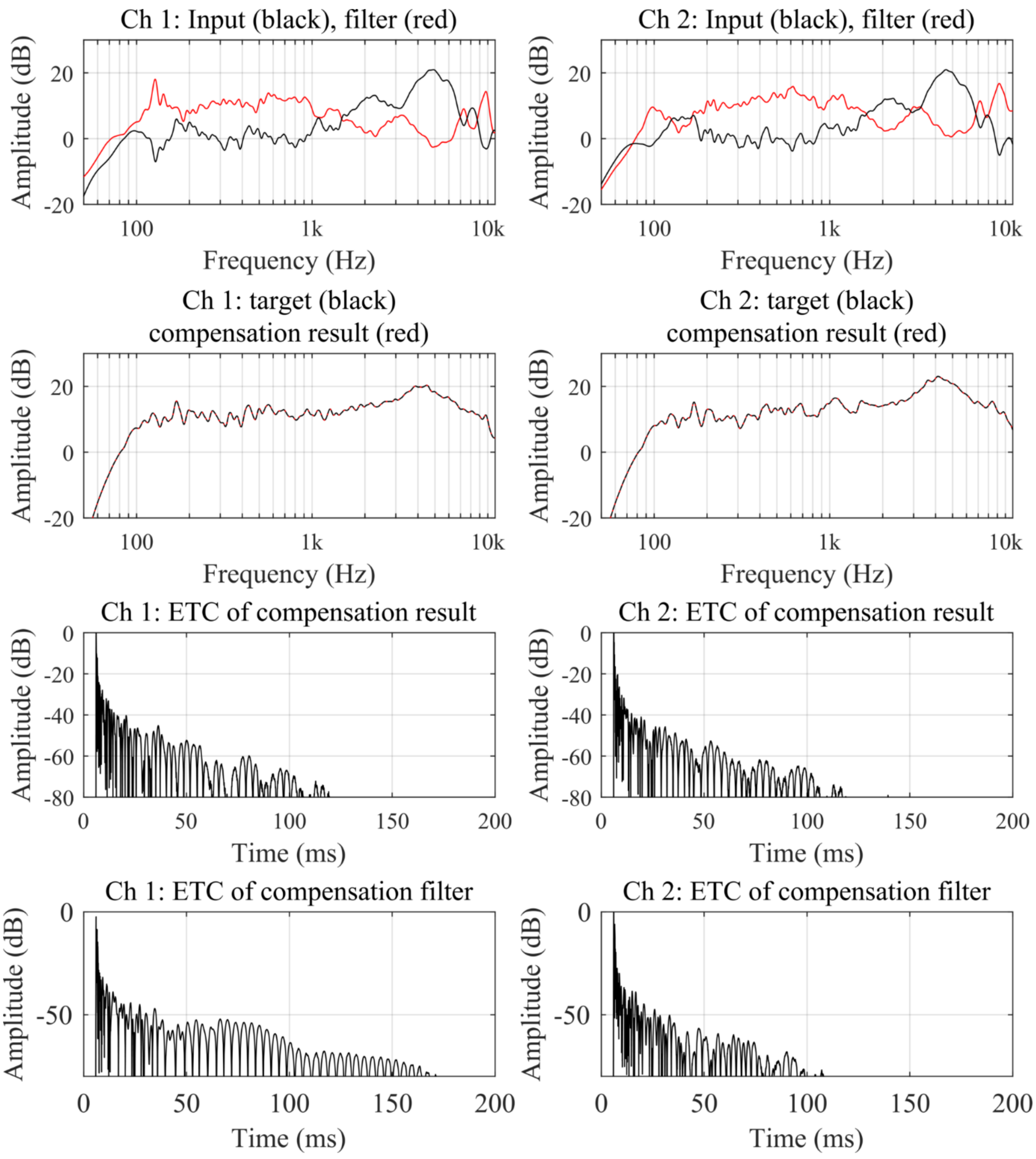


Figure D.7. Compensation filter for Reverberation Chamber 16 wedges.

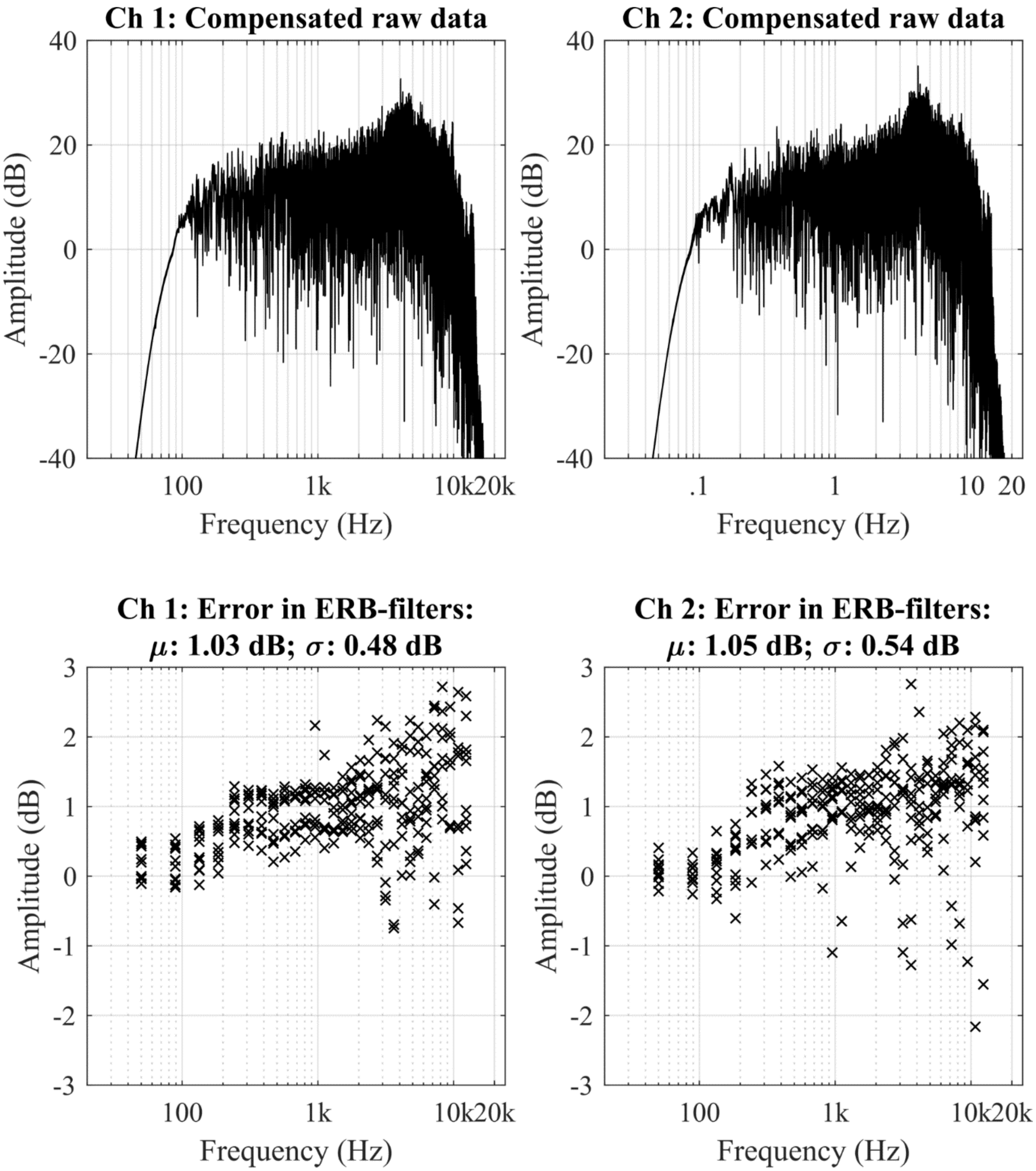


Figure D.8. Simulated Errors for Compensation Filter inclusion, Reverberation Chamber 16 wedges.

D.5 Reverberation Chamber 24 Wedges

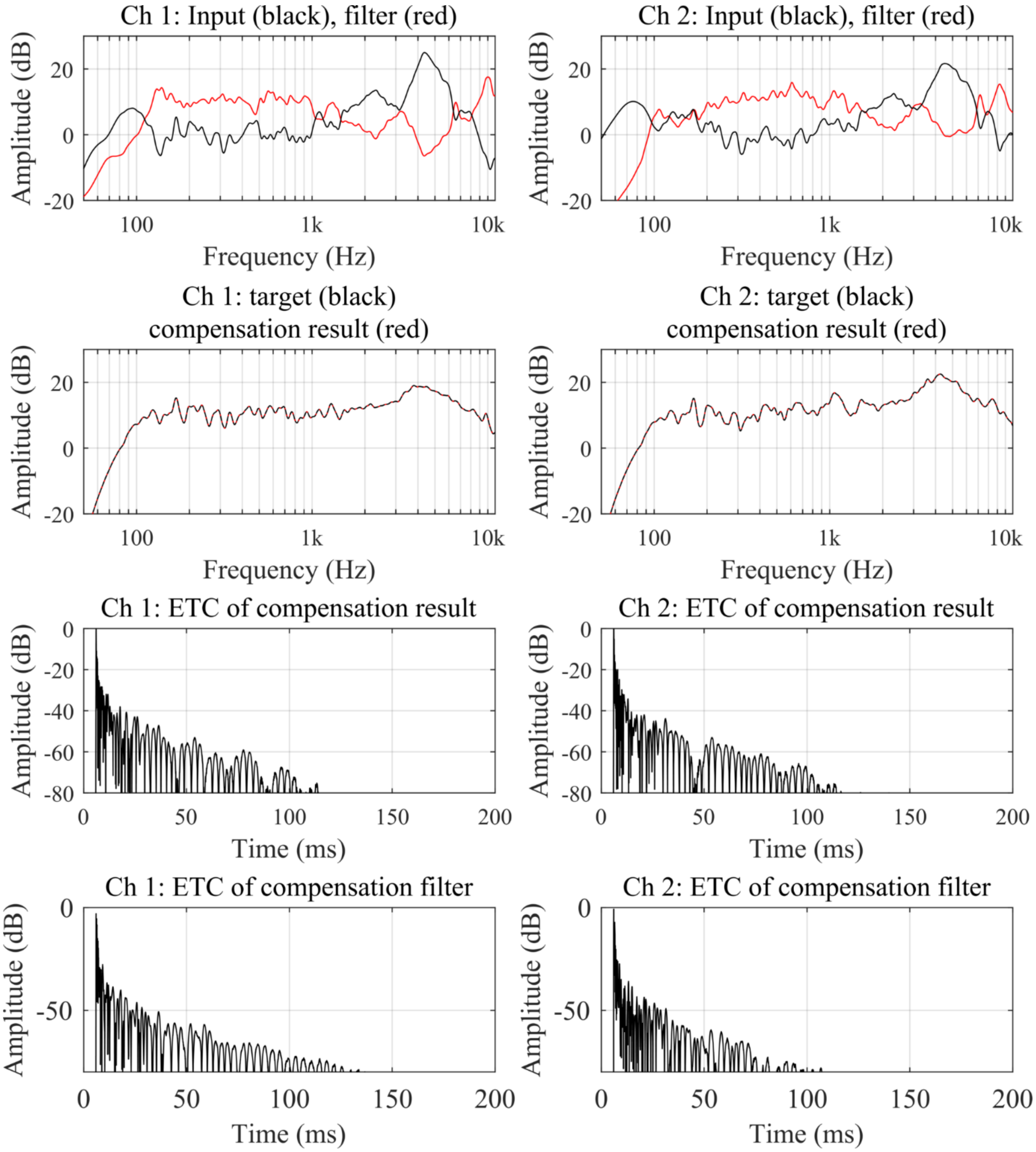


Figure D.9. Compensation filter for Reverberation Chamber 24 wedges.

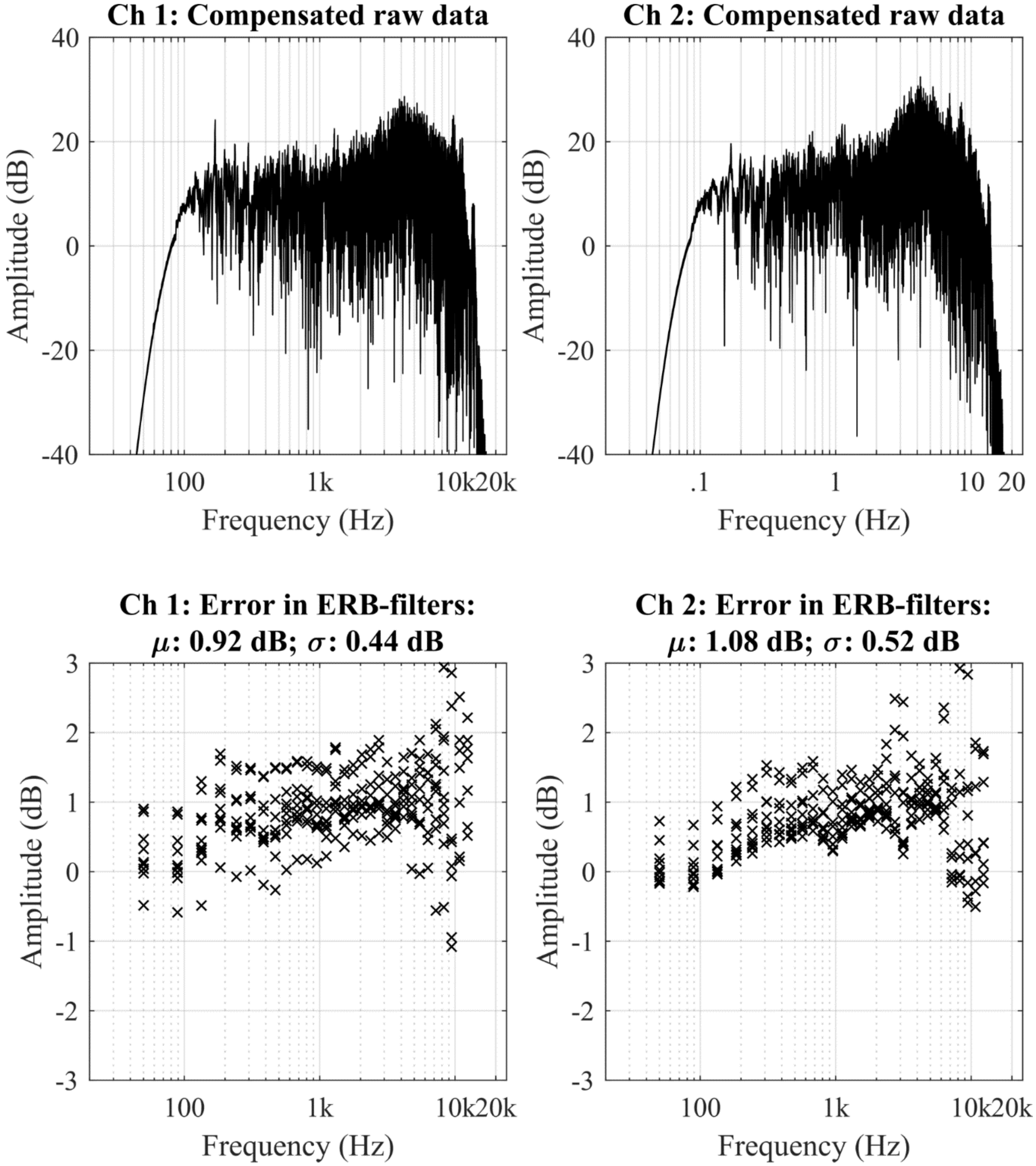


Figure D.10. Simulated Errors for Compensation Filter inclusion, Reverberation Chamber 24 wedges.

D.6 Reverberation Chamber 32 Wedges

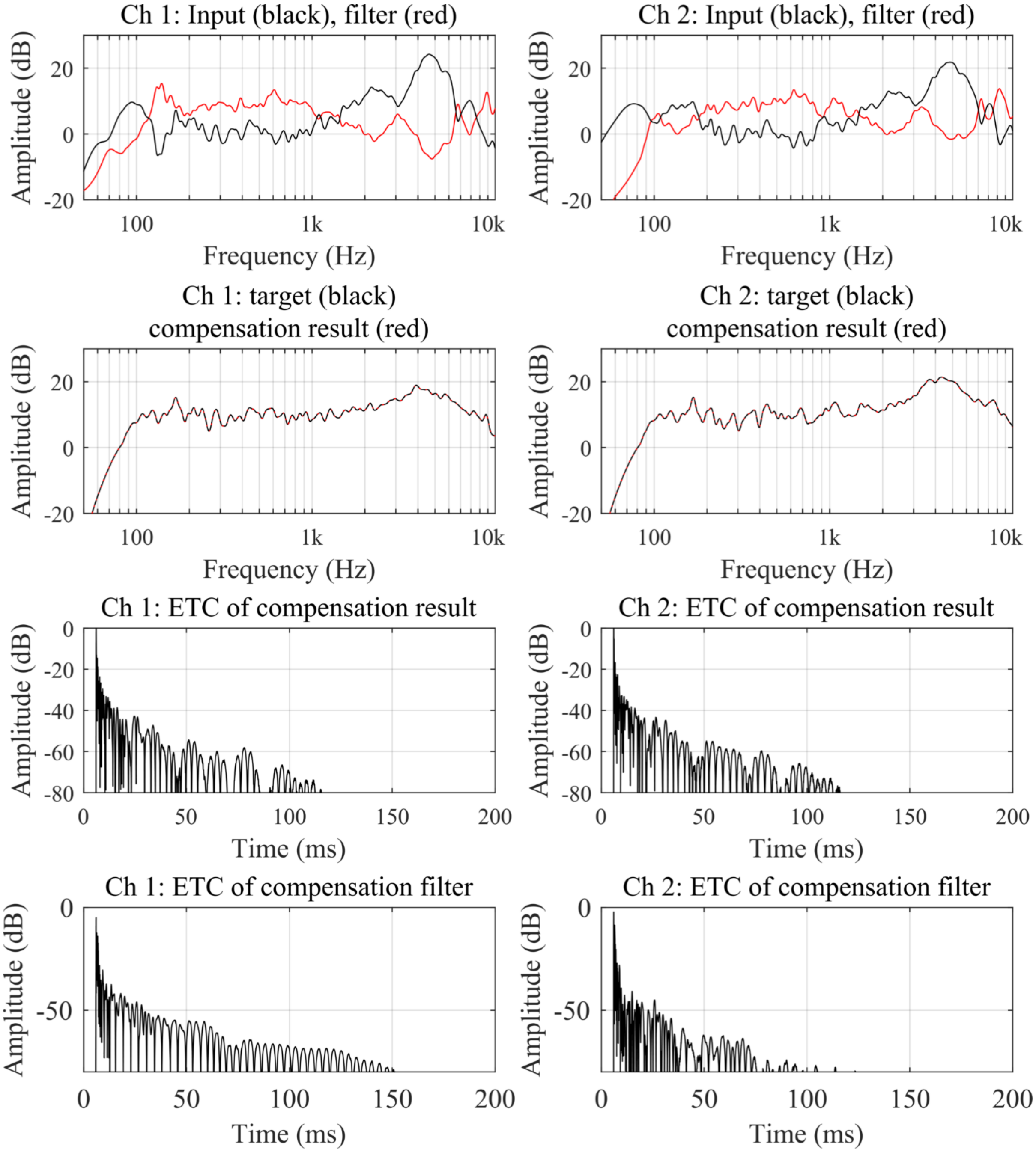


Figure D.11. Compensation filter for Reverberation Chamber 32 wedges.

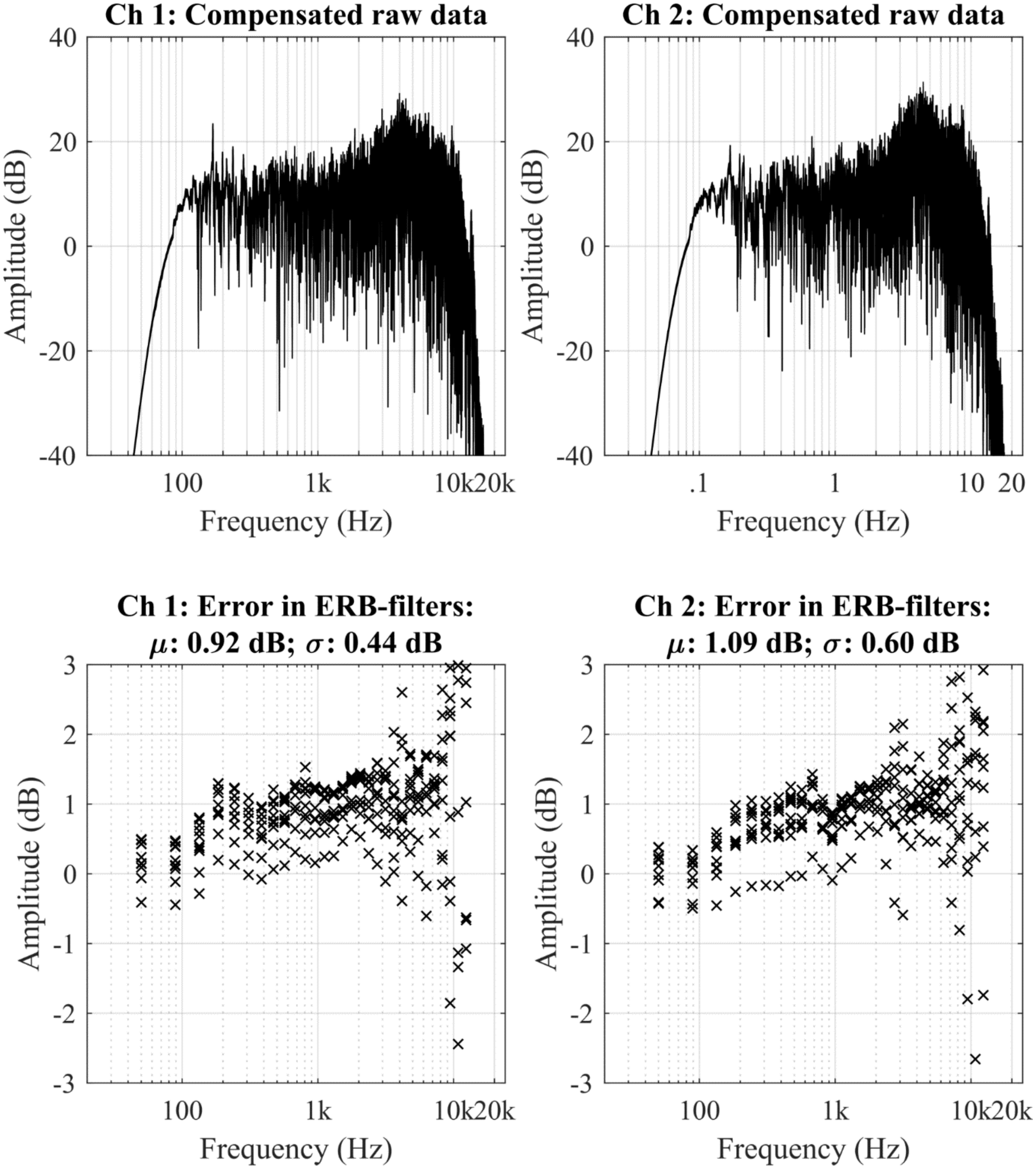


Figure D.12. Simulated Errors for Compensation Filter inclusion, Reverberation Chamber 32 wedges.

D.7 Classroom C215

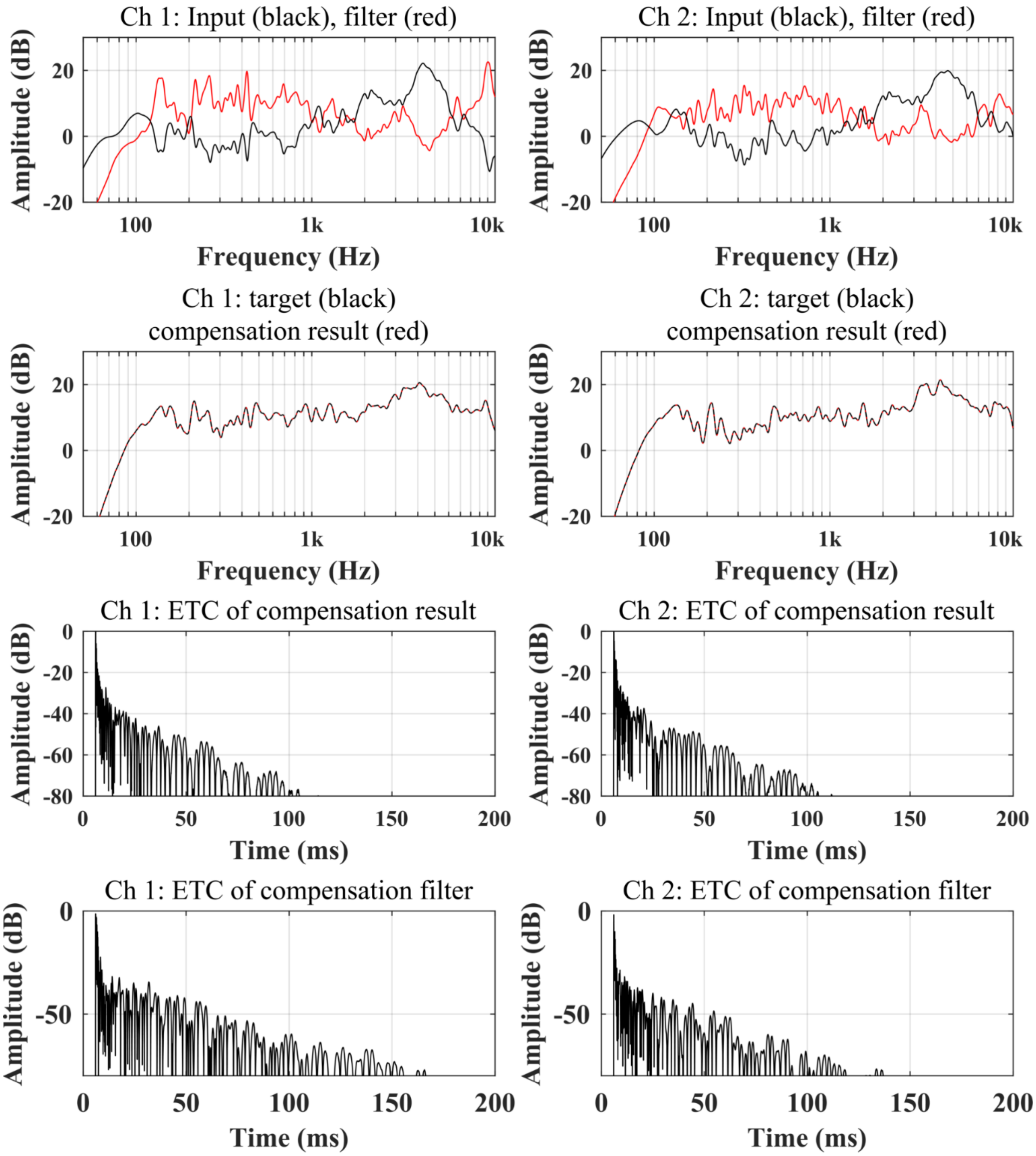


Figure D.13. Compensation filter for ESC C215.

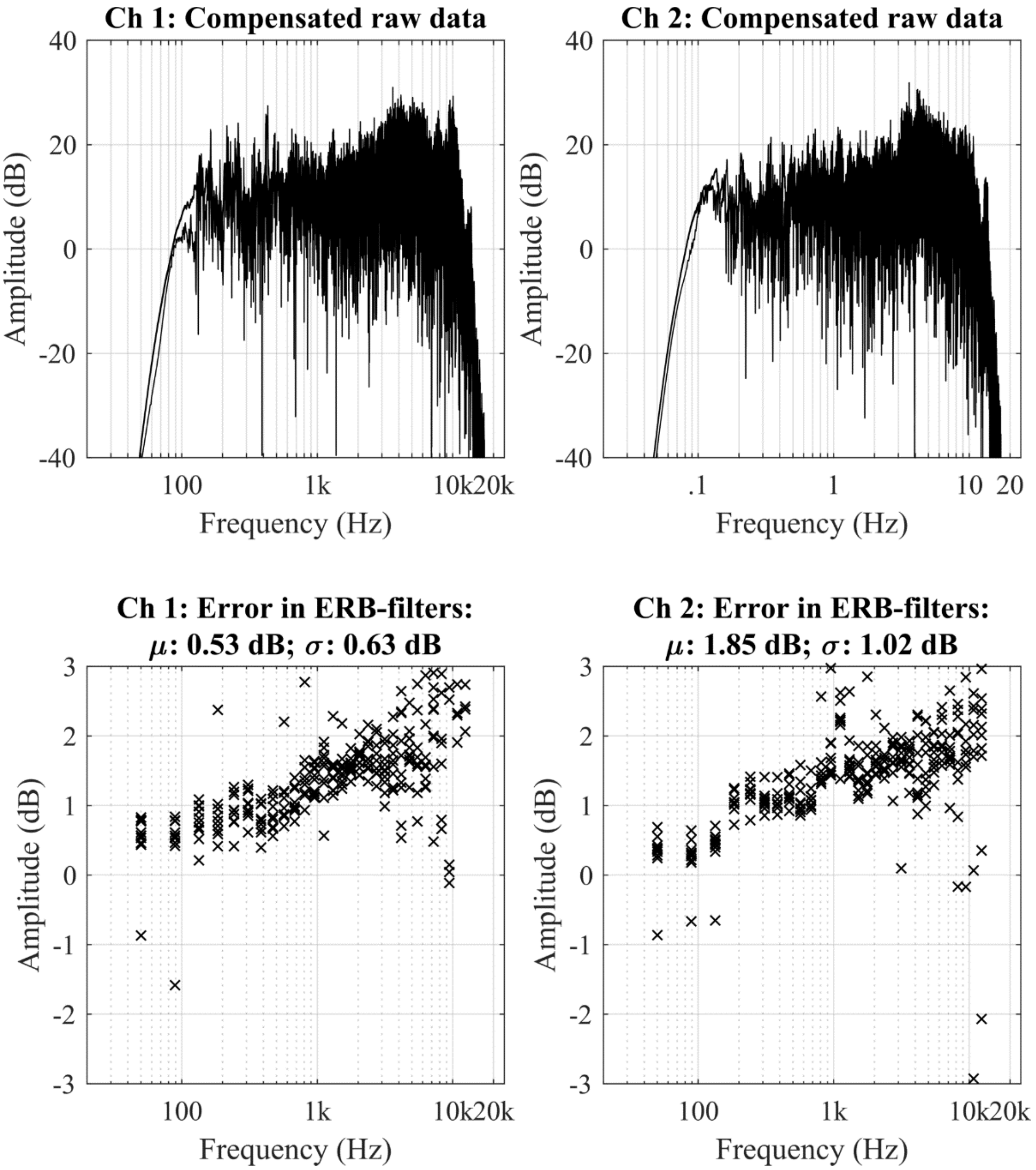


Figure D.14. Simulated Errors for Compensation Filter inclusion, ESC C215.

D.8 Classroom C261

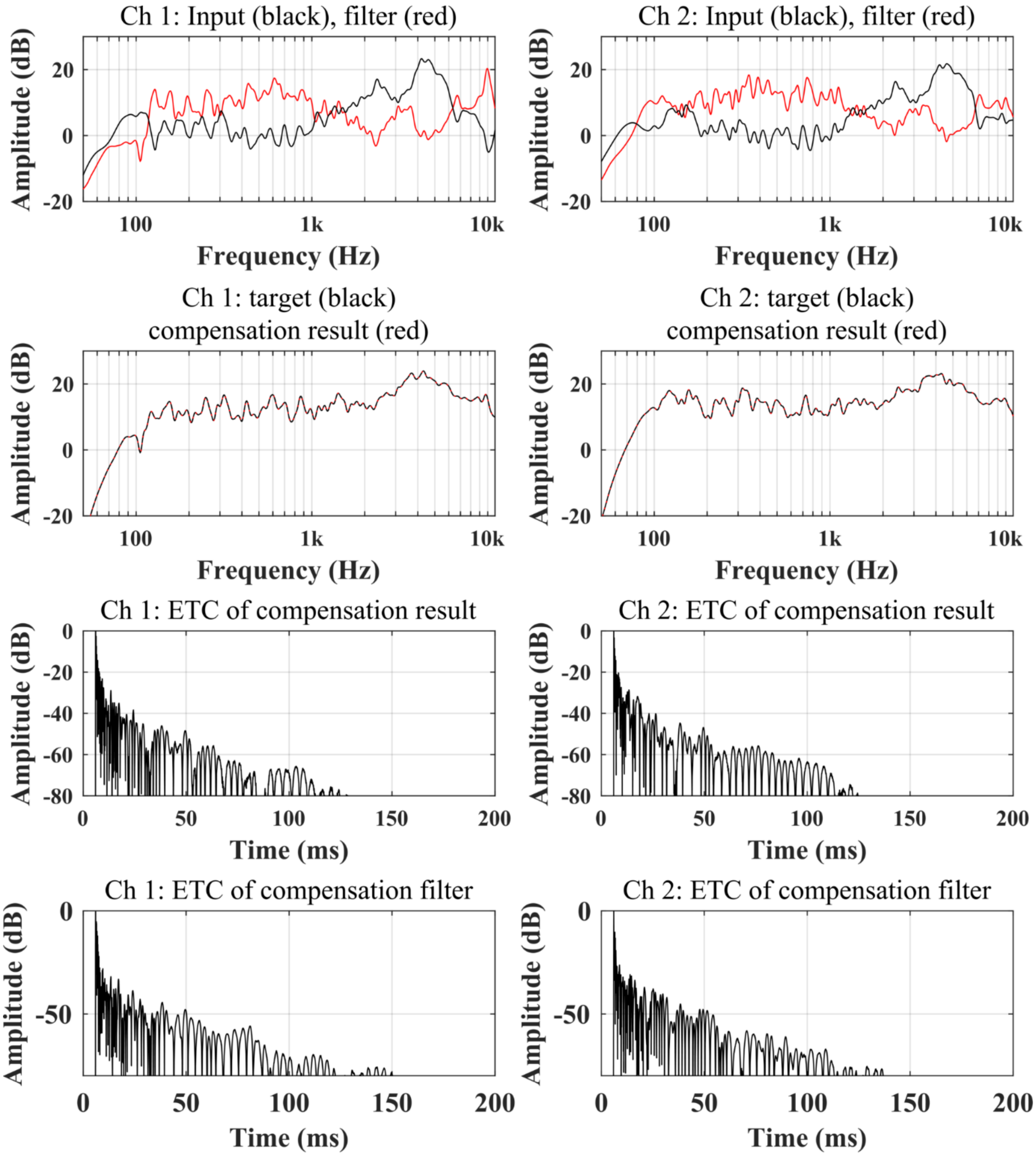


Figure D.15. Compensation filter for ESC C261.

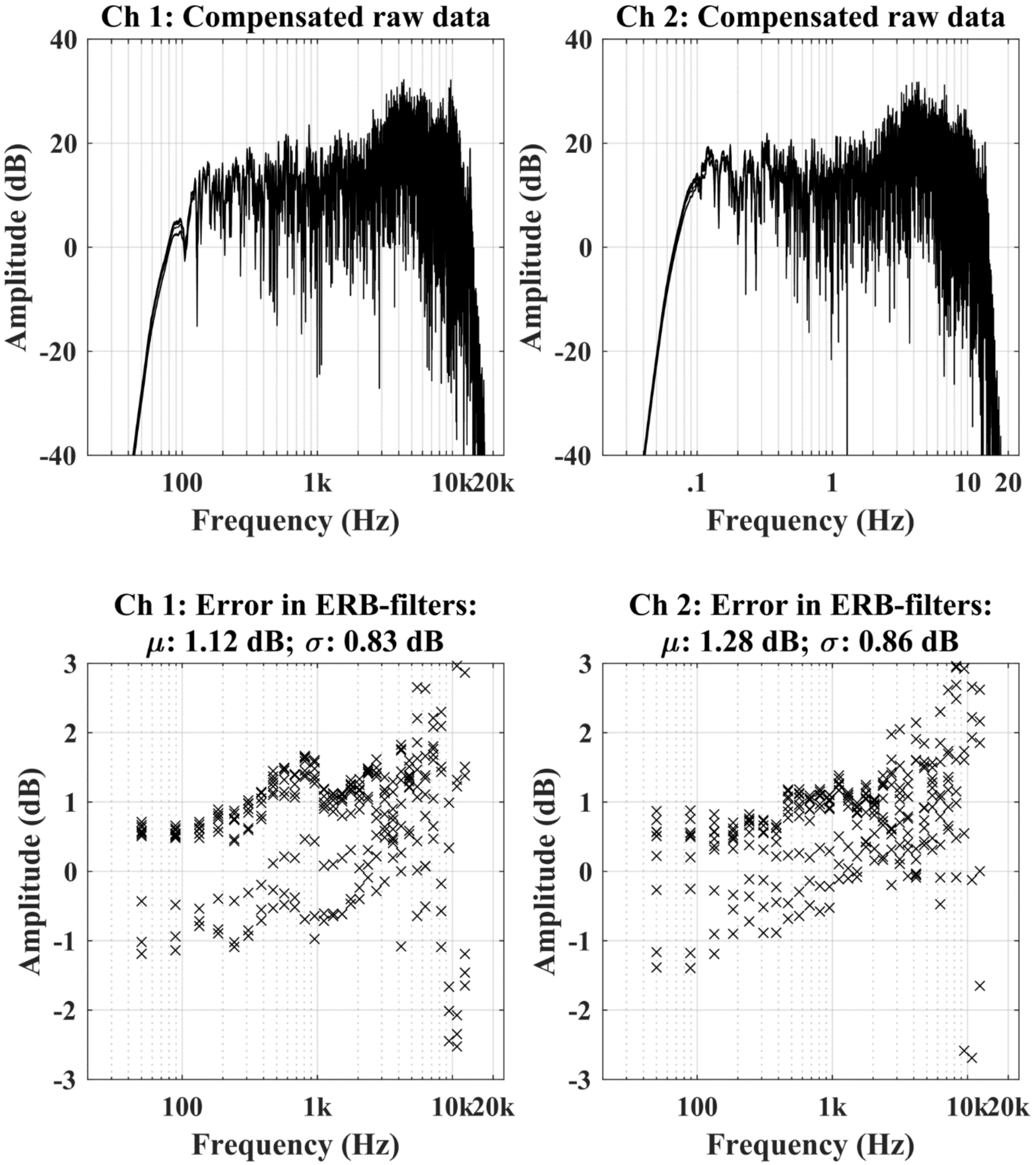


Figure D.16. Simulated Errors for Compensation Filter inclusion, ESC C261.

D.9 Simulated ESC C261

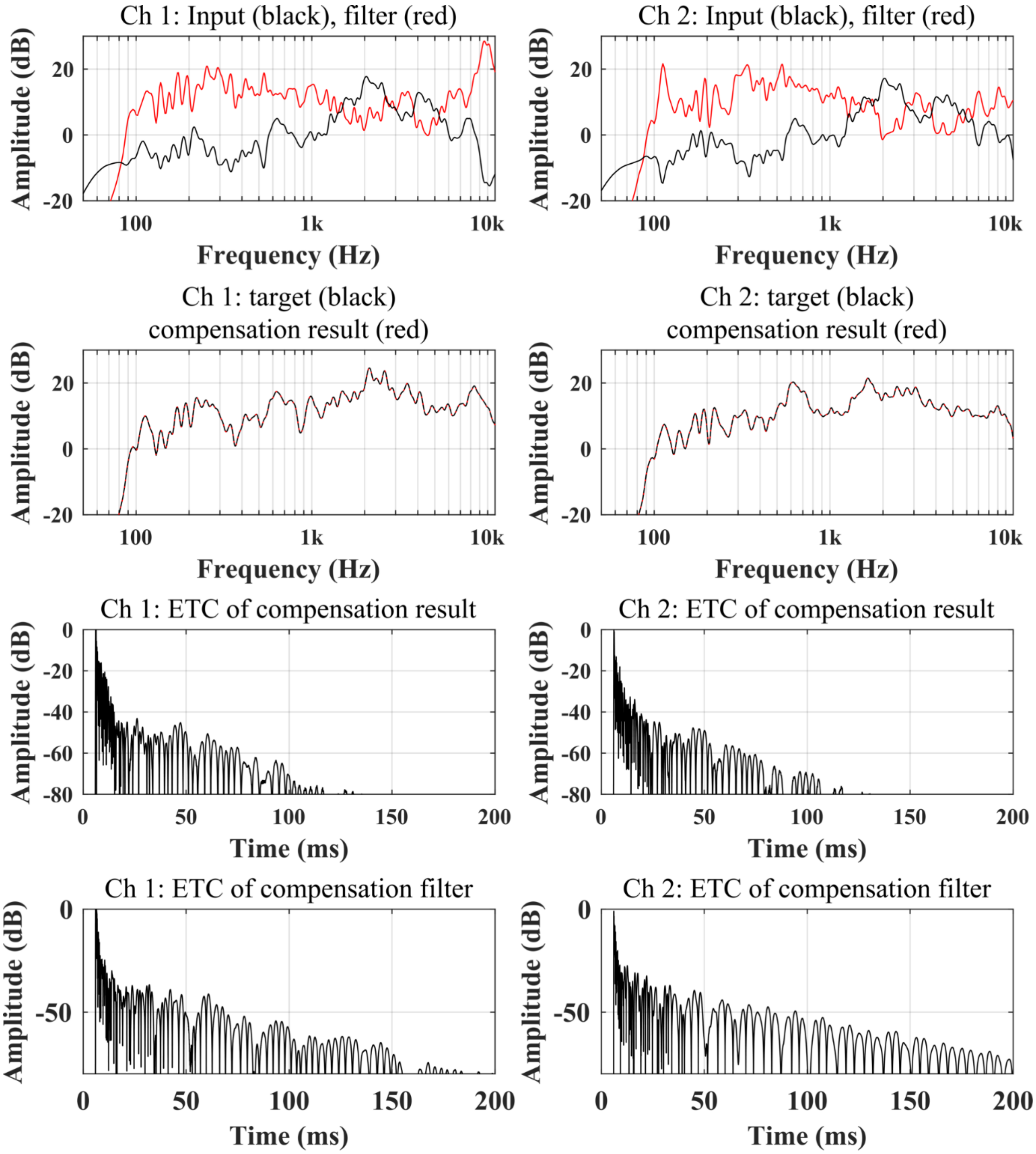


Figure D.17. Compensation filter for Simulated ESC C261.

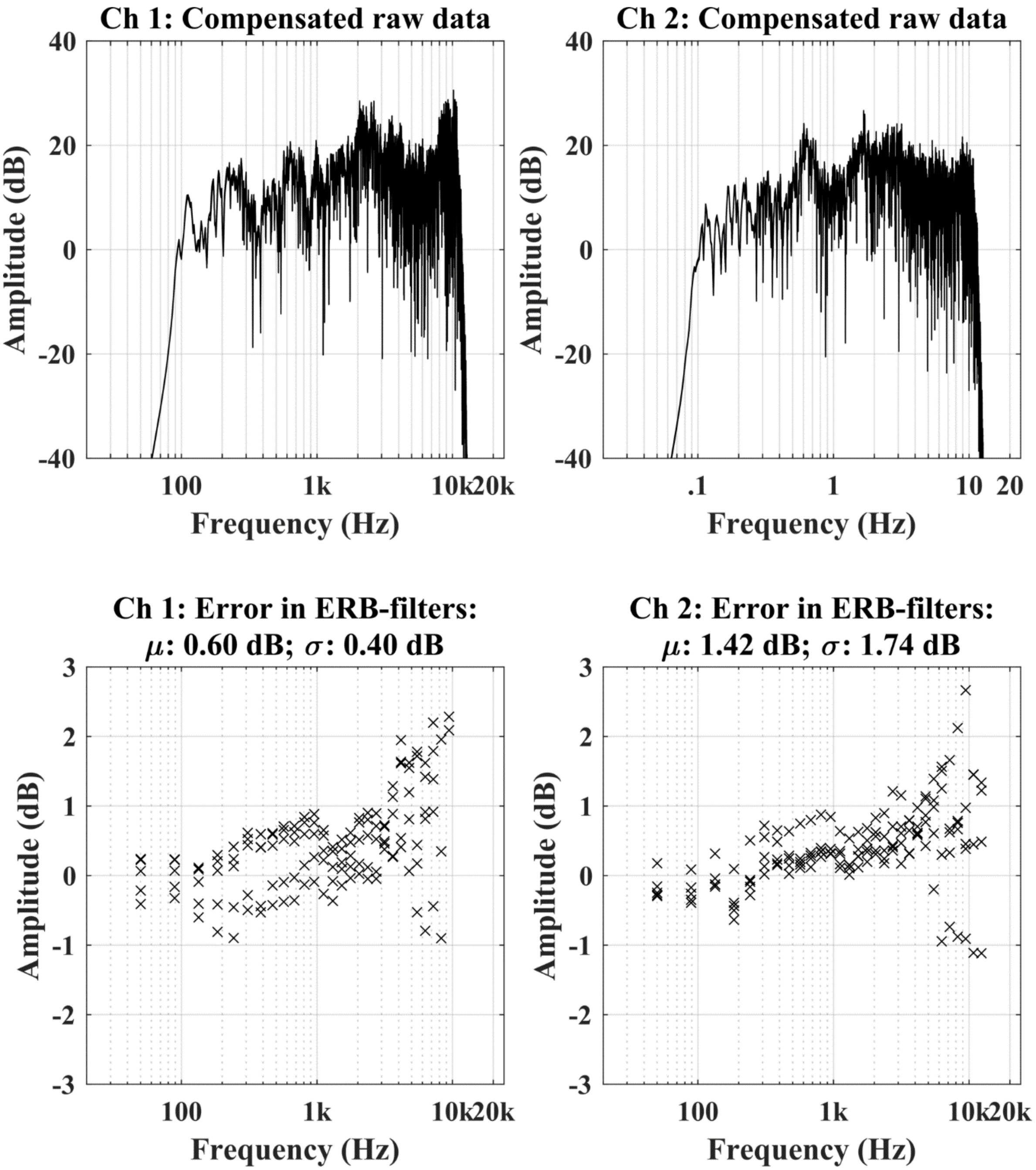


Figure D.18. Simulated Errors for Compensation Filter inclusion, simulated ESC C261.

D.10 Simulated ESC C261 Modified

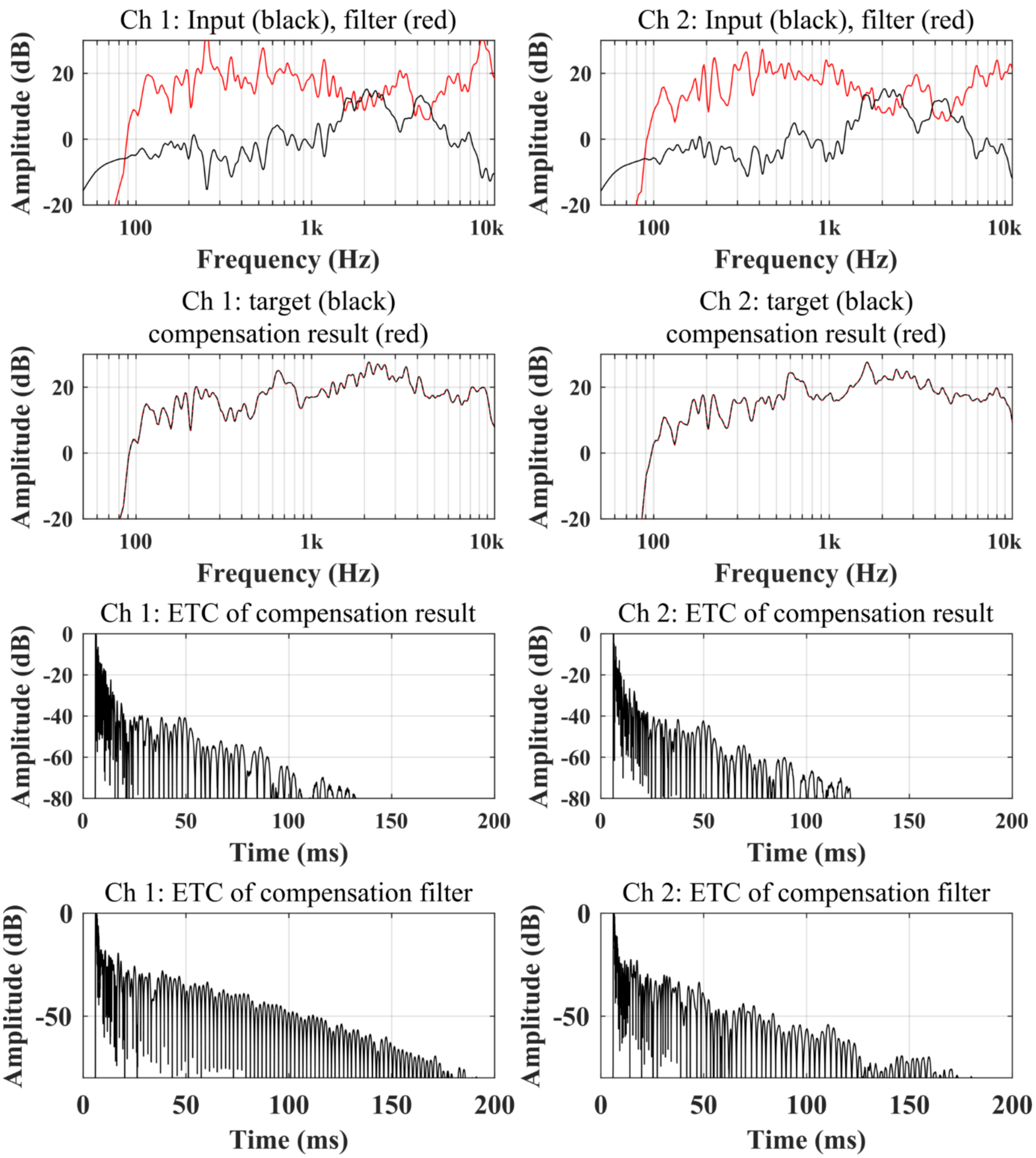


Figure D.19. Compensation filter for Simulated ESC C261, Modified.

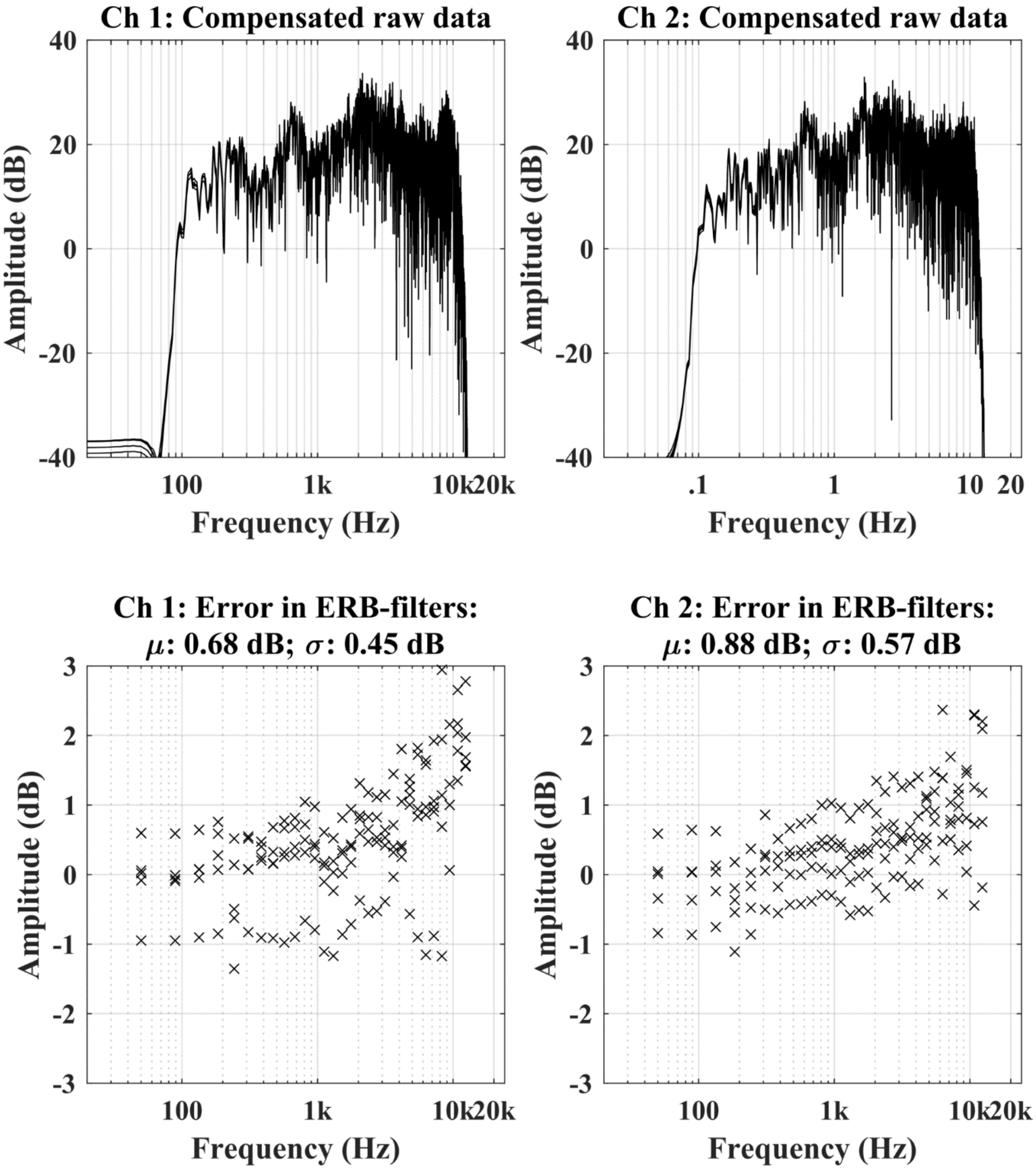


Figure D.20. Simulated Errors for Compensation Filter inclusion, Simulated ESC C261, Modified.

D.11 de Jong Concert Hall

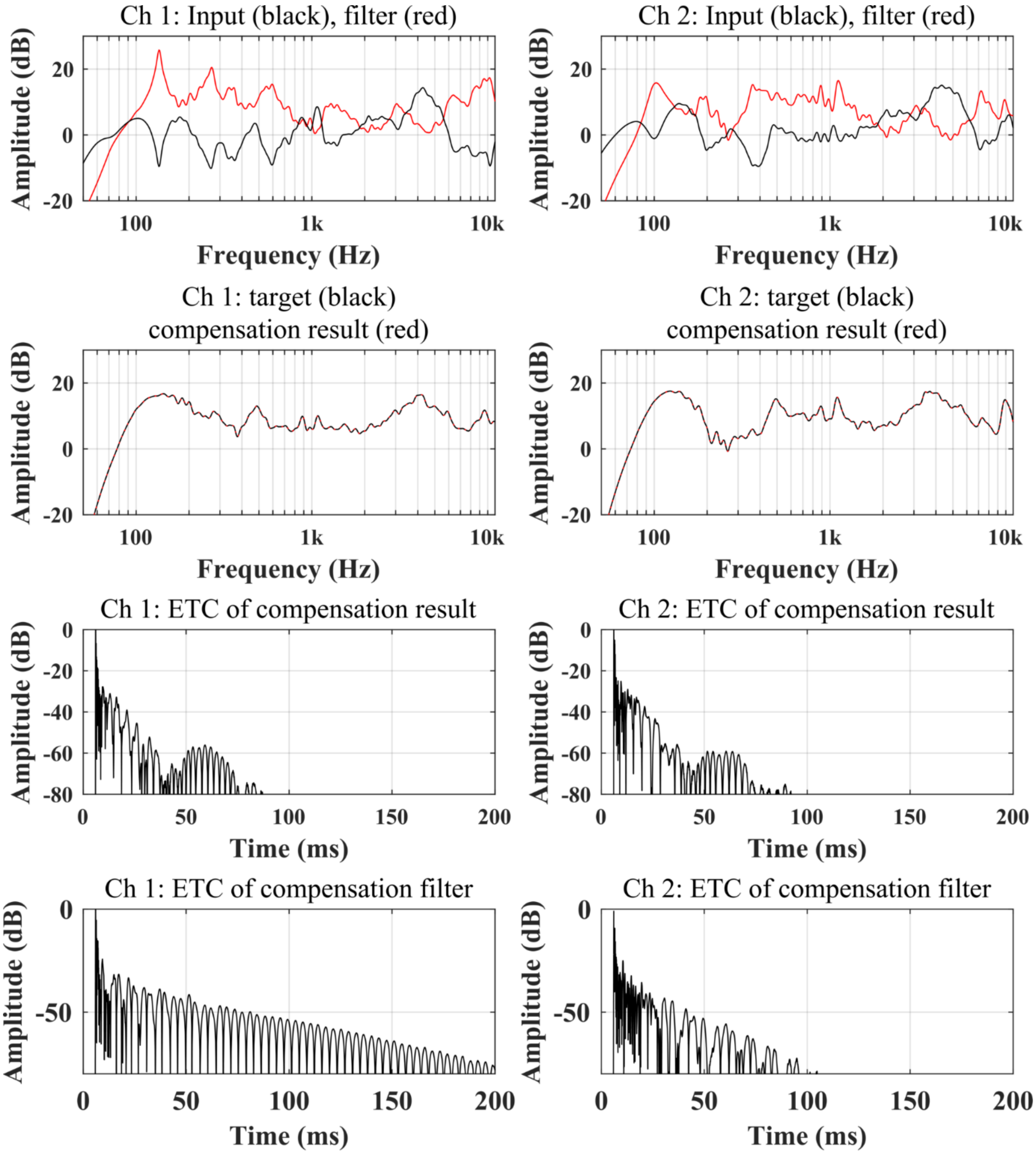


Figure D.21. Compensation filter for de Jong Concert Hall.

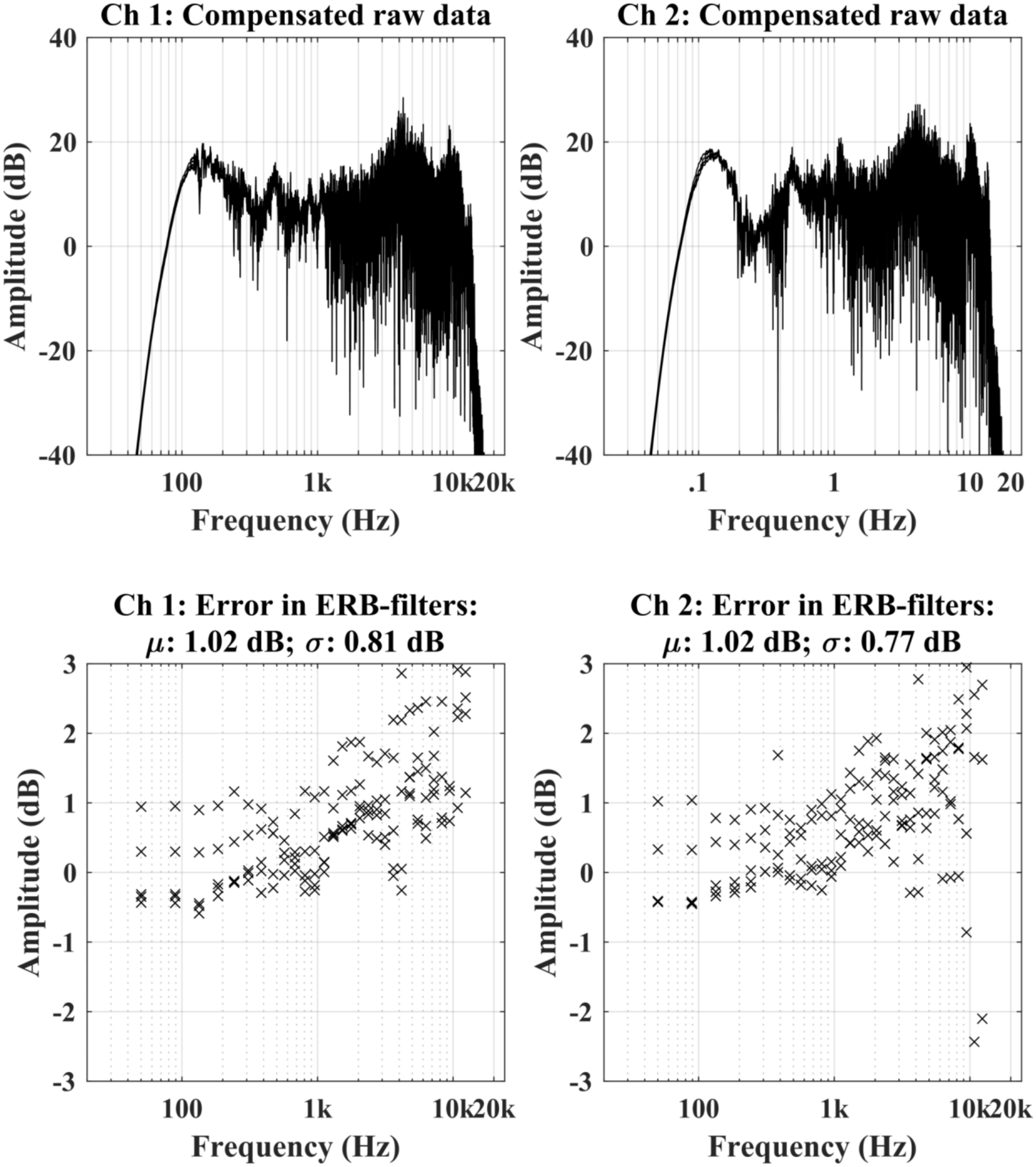


Figure D.22. Simulated Errors for Compensation Filter inclusion, de Jong Concert Hall.

D.12 Simulated de Jong Concert Hall

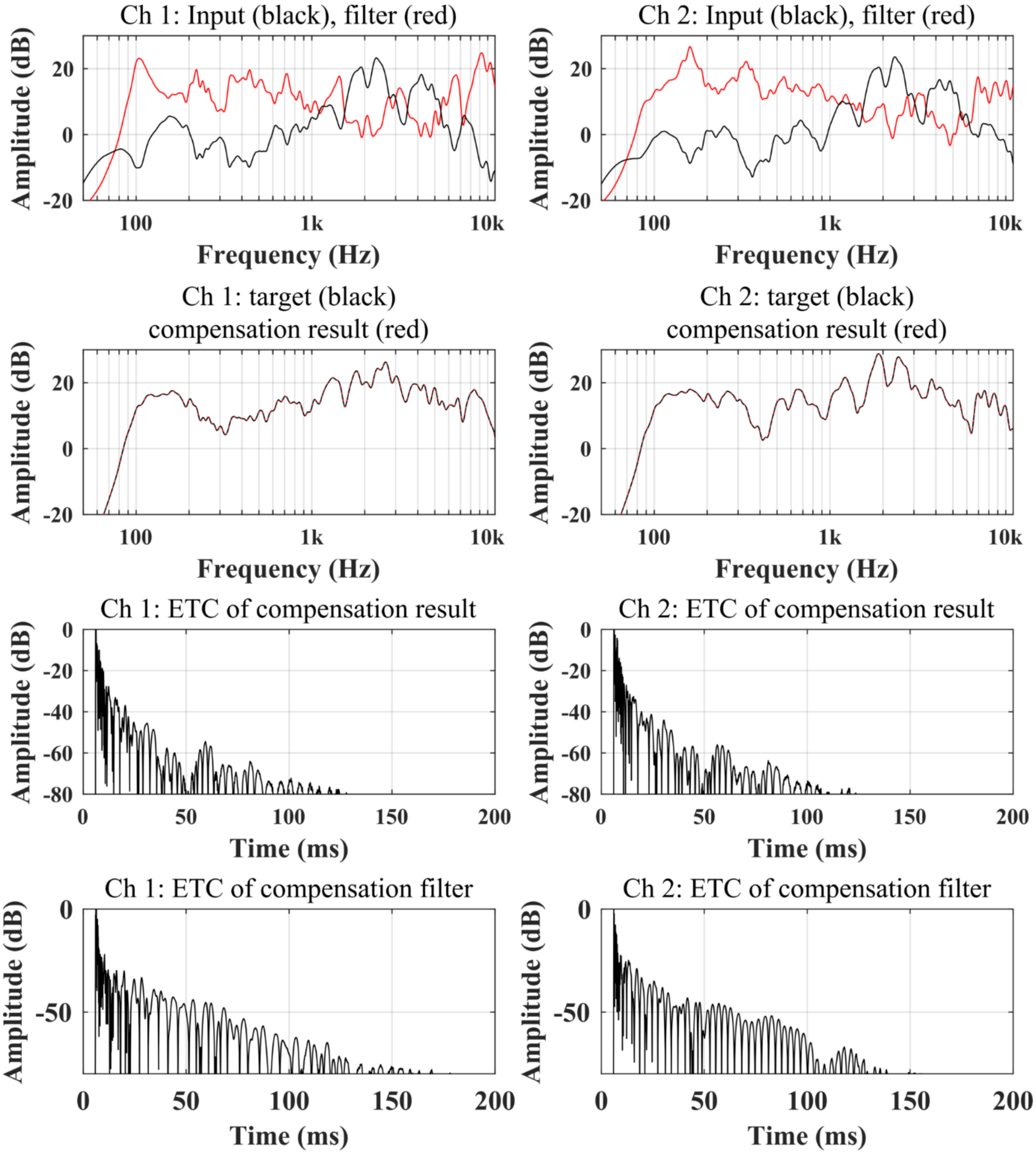


Figure D.23. Compensation filter for Simulated de Jong Concert Hall.

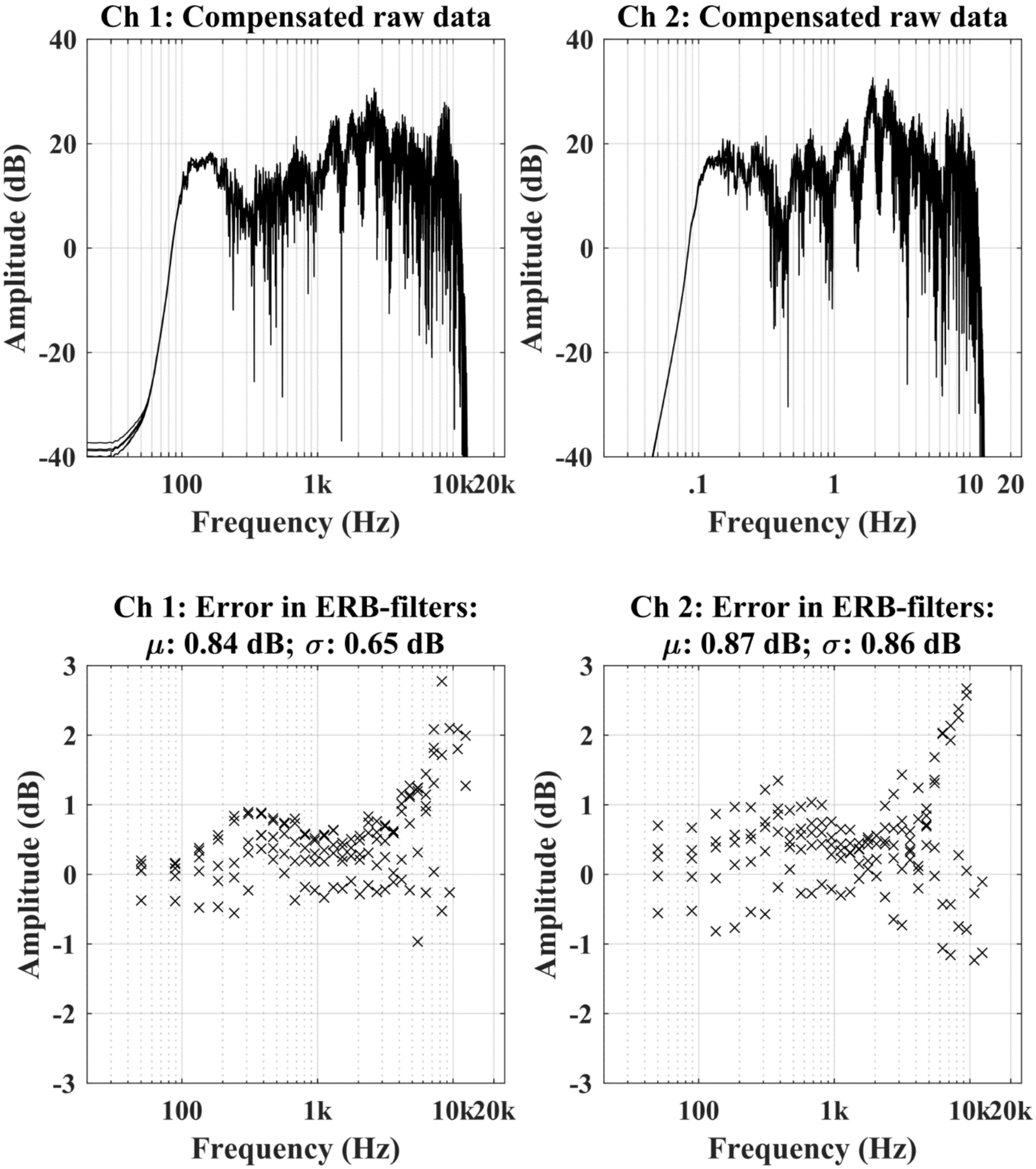


Figure D.24. Simulated Errors for Compensation Filter inclusion, Simuated de Jong Concert Hall.

Appendix E

Objective Evaluation of RTCS Performance

E.1 Introduction

In addition to the reverberation chamber with 32 absorbing wedges shown in Section 5.2, several more room conditions were prepared and evaluated. Table E.1 lists the main parameters for the filter computation and the RTCS level settings for the room OBRIR and the compensation filter for each condition. The simulated room conditions are different from the measured room in that calibrated pressure data is not available. Instead, the OBRIRs for the room simulations are stored as wav files, which have normalized units. During the modification process, a normalization factor is computed and saved. Then each of the RTCS OBRIR measurements are modified with the same normalization factor. Thus, even though wav files are being compared, the levels are comparable because the arbitrary normalization factor was removed and replaced with a known normalization factor. The compensation filter computation was performed in the same manner for both the simulated and measured OBRIR conditions.

Table E.1. OBRIRs in the RTCS. Filter computation information and level settings for use in the RTCS.

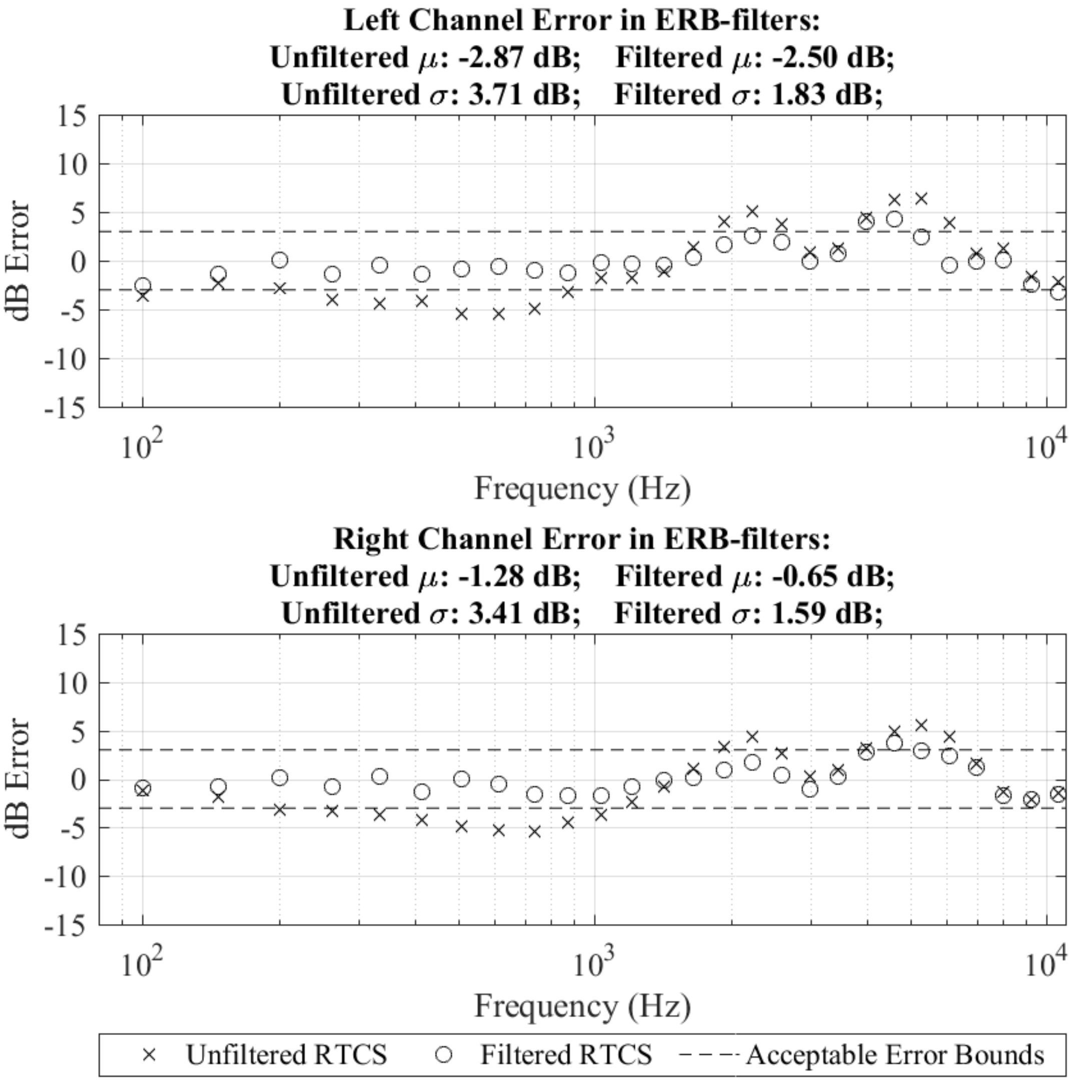
Condition	OBRIR Length (seconds)	Filter Length (samples)	Filter Length (seconds)	Room Level (dB)	Filter Level (dB)
RE00M	10	144000	3	-18	-5.8
RE02M	7.8	2^{16}	1.36	-18.72	-6.8
RE04M	2.8	2^{16}	1.36	-15.9	-5.8
RE08M	2.8	2^{16}	1.36	-16.5	-3.3
RE16M	1.6	2^{16}	1.36	-17.5	-3.3
RE24M	1.2	2^{16}	1.36	-15.9	-3.3
RE32M	1.2	2^{16}	1.36	-10.4	-5.3
C215M	2.2	62400	1.3	-13.62	-6.8
C261M	1.6	43200	0.9	-17.52	-4.3
C261S	1.2	24000	0.5	-8.4	-5.8
C261ABS	1.2	2^{14}	0.34	-17.52	-9.9
DJCHM	1.3	48000	1	-14.51	-13
DJCHS	2.8	2^{15}	0.68	-17	-18

Table E.2 summarizes the results of the error measurements performed in the time-and frequency-domains for each condition. A more detailed figure is presented for each condition in each subsection below.

Table E.2. RTCS Validation. Objective error measurement summary for each condition tested.

		Frequency-Domain				Time-Domain	
Condition		L Mean ($\boldsymbol{\mu}$) (dB)	L $\boldsymbol{\sigma}$ (dB)	R Mean ($\boldsymbol{\mu}$) (dB)	R $\boldsymbol{\sigma}$ (dB)	L Level (dB)	R Level (dB)
RE00M	Unfiltered	-2.87	3.71	-1.28	3.41	-0.54	43
	Filtered	-2.5	1.83	-0.65	1.59	0.23	0.08
RE02M	Unfiltered	-6.51	2.03	-5.24	2.56	-3.13	-3.31
	Filtered	-7.02	1.85	-4.36	2.22	-2.77	-1.68
RE04M	Unfiltered	4.12	4.13	7.71	4.04	4.51	4.32
	Filtered	-1.96	1.94	0.5	2.35	-0.14	0.89
RE08M	Unfiltered	-2.42	4.91	-1.22	3.94	0.37	0.14
	Filtered	0.2	1.13	2.61	1.57	-0.09	1.15
RE16M	Unfiltered	0.04	4.11	0.51	3.68	1.99	1.57
	Filtered	0.12	1.48	3.68	0.72	0.23	0.35
RE24M	Unfiltered	-0.86	4.72	0.03	3.77	2.4	1.9
	Filtered	-0.39	1.04	0.71	1.23	-0.45	0.01
RE32M	Unfiltered	1.31	4.73	2.15	3.45	3.17	2.65
	Filtered	-0.86	1.48	0.3	1.51	-0.45	-0.25
C215M	Unfiltered	-1.5	4.33	1.44	3.89	6.9	6.69
	Filtered	-1.73	1.18	1.73	1.67	3.55	5.11
C261M	Unfiltered	0.03	4.13	1.69	4.07	4.94	4.73
	Filtered	-1.1	1.64	1.1	1.44	1.47	2.81
C261S	Unfiltered	-0.88	4.11	2.84	4.43	1.5	1.93
	Filtered	-2.27	0.86	2.22	3.05	-0.86	1.63
C261ABS	Unfiltered	-3.15	4.11	-0.46	4.09	0.31	0.44
	Filtered	-2.34	1.13	0.47	1.5	-0.96	0.25
DJCHM	Unfiltered	5.05	4.19	5.57	3.83	4.05	3.41
	Filtered	-0.93	2.68	-0.39	2.83	-1.7	-0.97
DJCHS	Unfiltered	-7.79	4.76	-6.6	5.13	-2.4	-2.42
	Filtered	-0.59	2.14	0.35	1.92	-0.47	0.59

E.2 Reverberation Chamber, 0 wedges



Left Channel Error in ERB-filters:
Unfiltered μ: -2.87 dB; Filtered μ: -2.50 dB;
Unfiltered σ: 3.71 dB; Filtered σ: 1.83 dB;

Right Channel Error in ERB-filters:
Unfiltered μ: -1.28 dB; Filtered μ: -0.65 dB;
Unfiltered σ: 3.41 dB; Filtered σ: 1.59 dB;

Figure E.1. Frequency-domain error results for RTCS representing Reverberation Chamber with 00 wedges. Each of the OBRIRs were modified to remove the effects of the KEMAR mouth simulator used in making the measurements so as to only show the part of the OBRIR indicative of the performance of the RTCS.

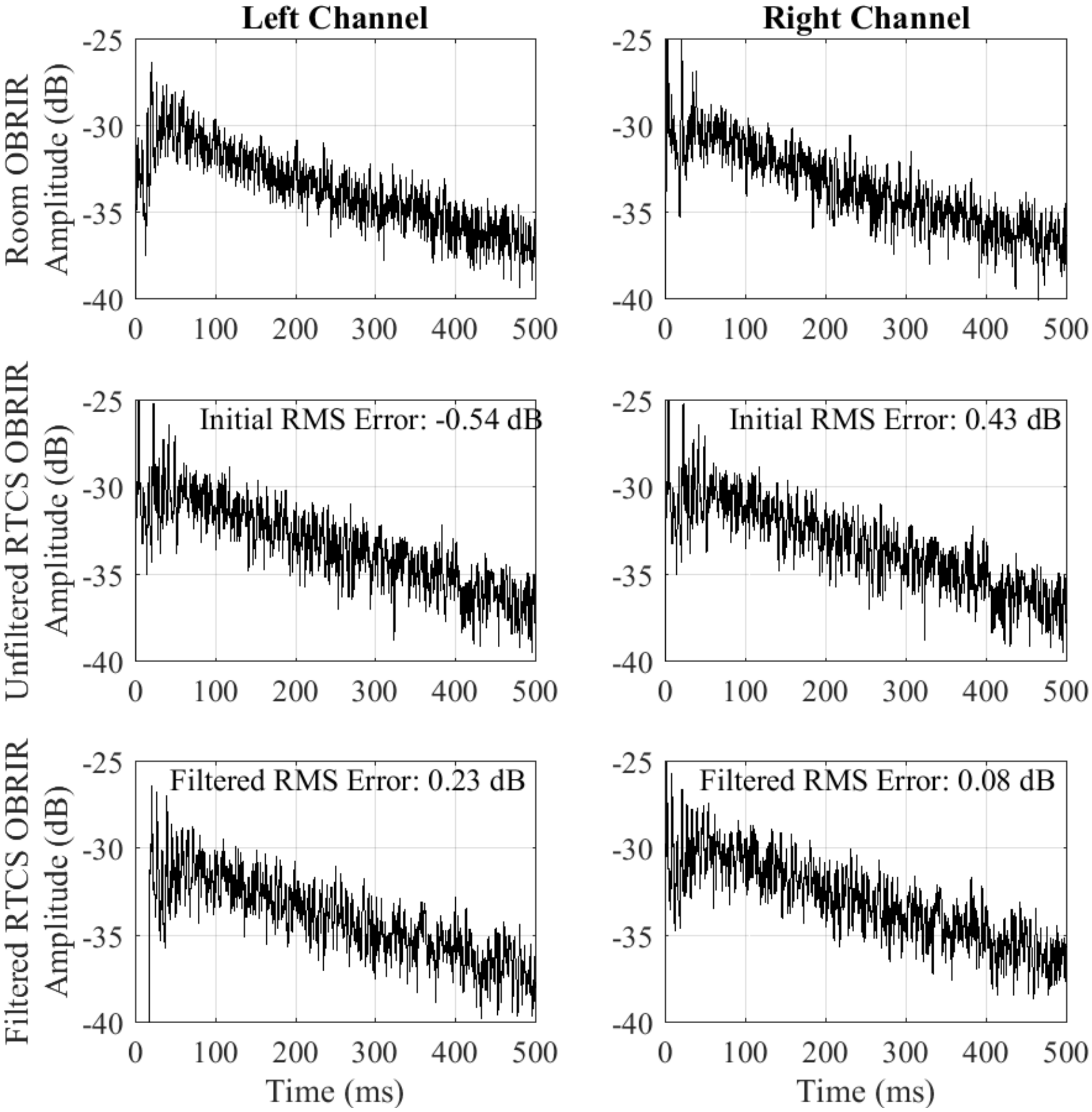


Figure E.2. Time-domain error results for the RTCS representing Reverberation Chamber with 00 wedges.

E.3 Reverberation Chamber, 2 wedges

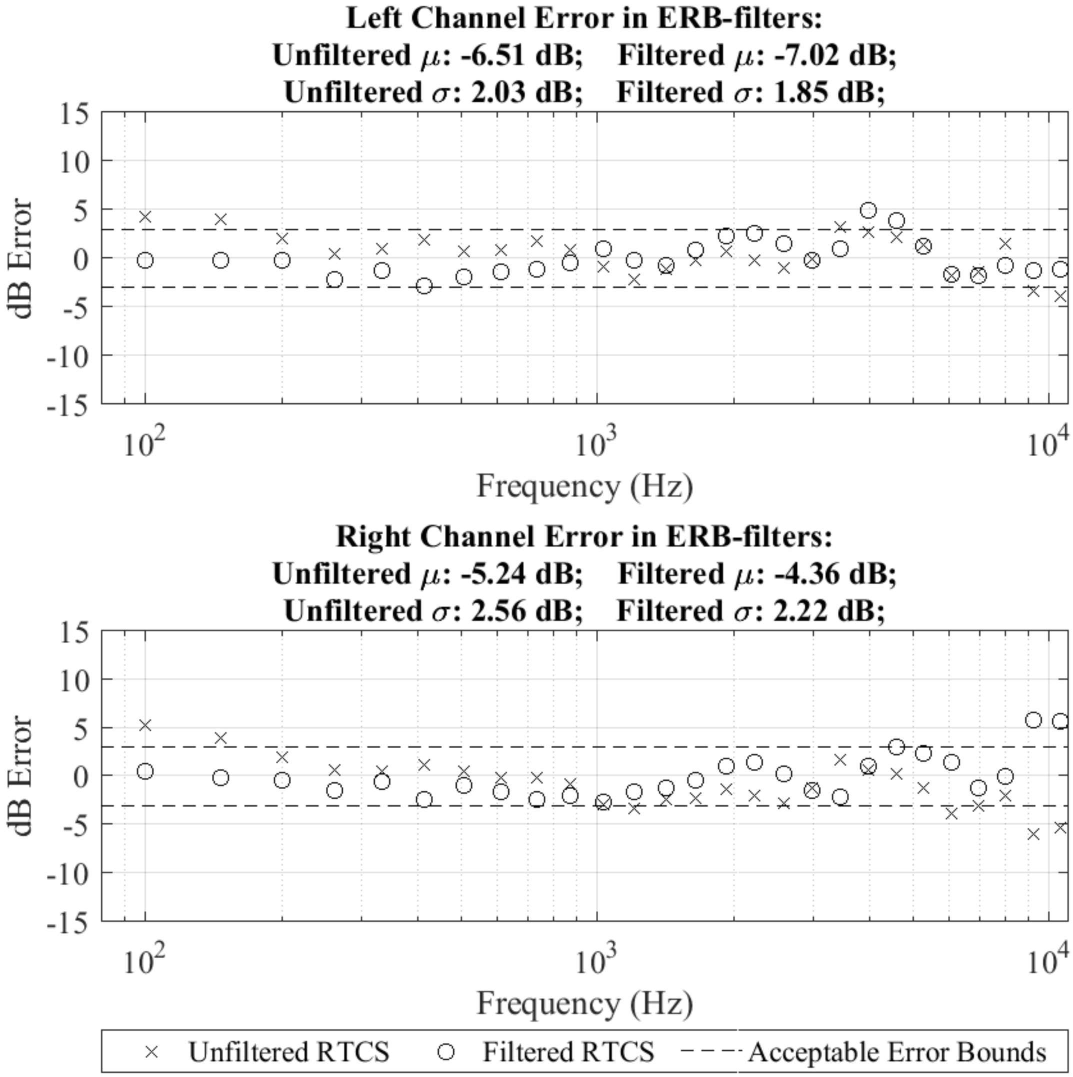


Figure E.3. Frequency-domain error results for RTCS representing Reverberation Chamber with 02 wedges

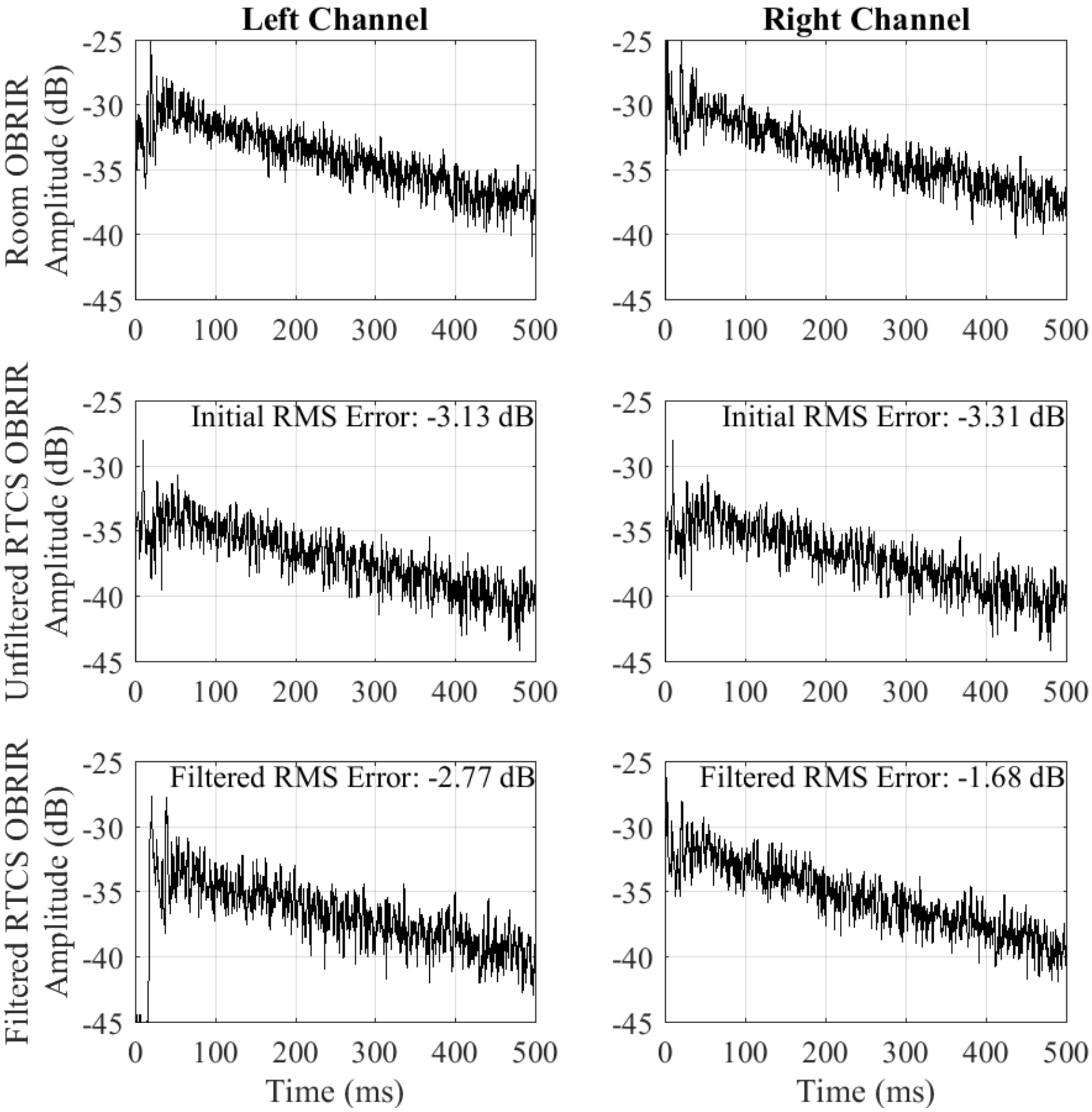


Figure E.4. Time-domain error results for RTCS representing Reverberation Chamber with 02 wedges.

E.4 Reverberation Chamber, 4 wedges

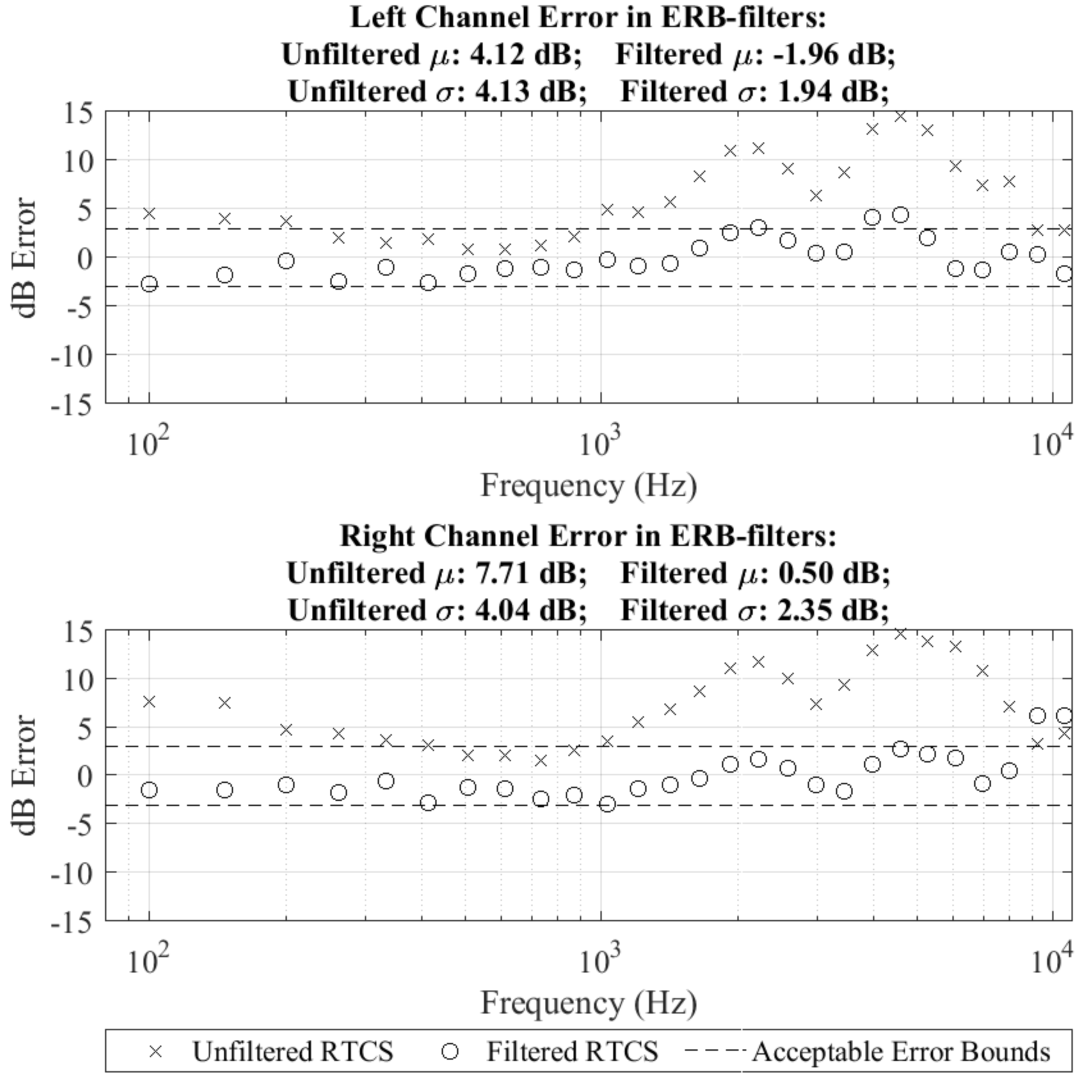


Figure E.5. Frequency-domain error results for RTCS representing Reverberation Chamber with 04 wedges. The filtered RTCS mean error was nearly completely contained in the acceptable error bounds of +/- 3 dB on both channels.

.

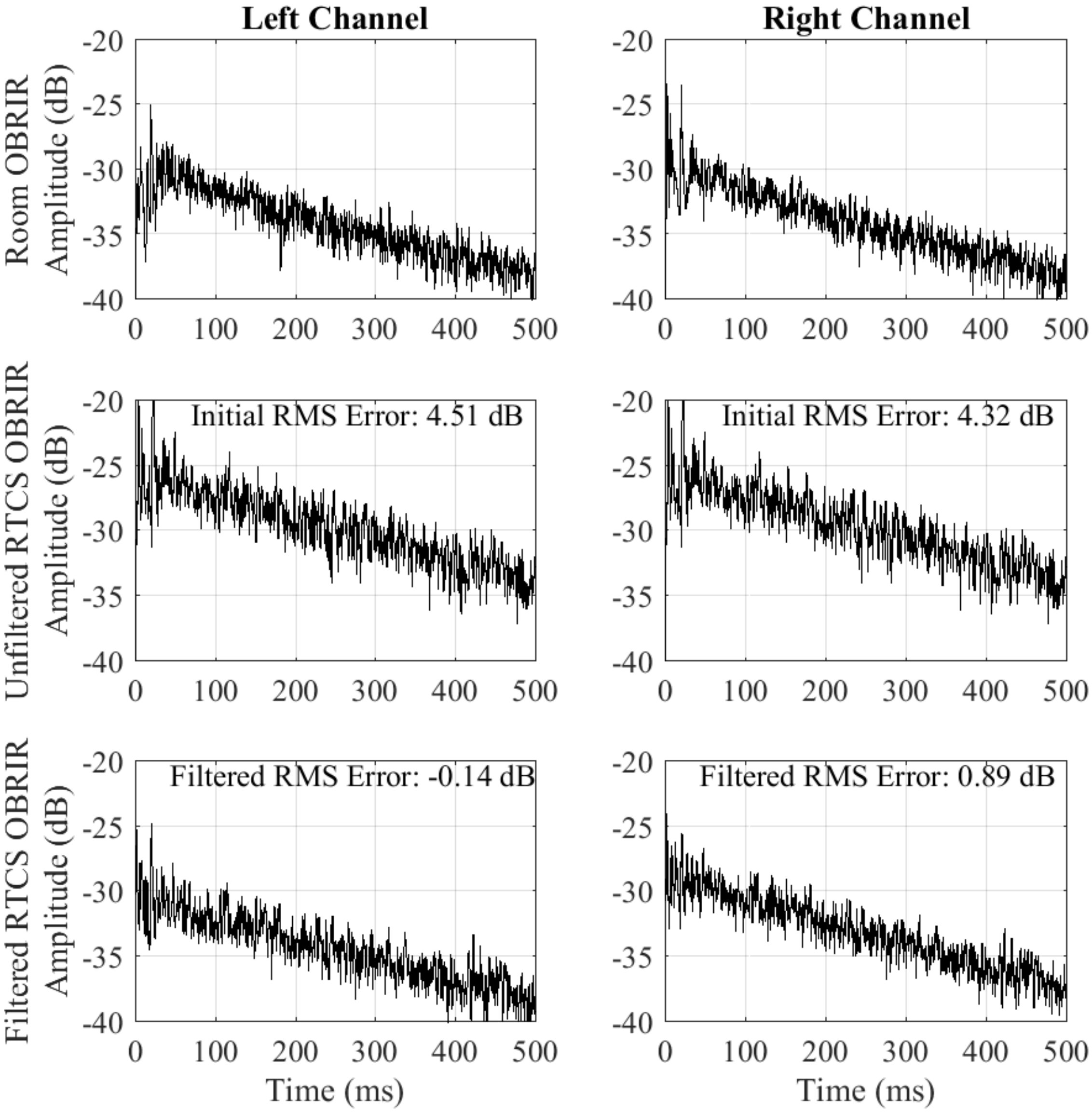
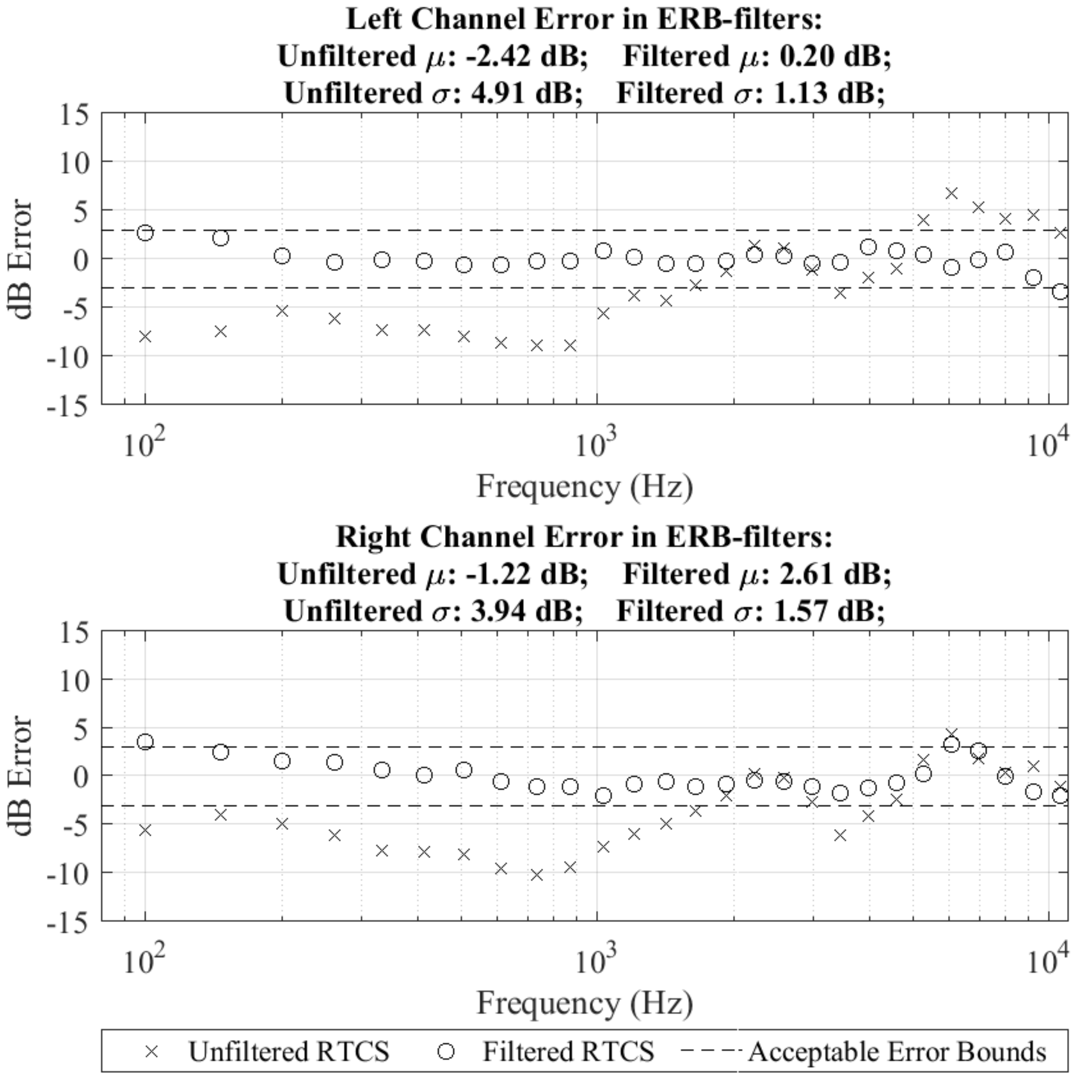


Figure E.6. Time-domain error results for RTCS representing Reverberation Chamber with 04 wedges.

E.5 Reverberation Chamber 8 wedges



Left Channel Error in ERB-filters:
Unfiltered μ: -2.42 dB; Filtered μ: 0.20 dB;
Unfiltered σ: 4.91 dB; Filtered σ: 1.13 dB;

Right Channel Error in ERB-filters:
Unfiltered μ: -1.22 dB; Filtered μ: 2.61 dB;
Unfiltered σ: 3.94 dB; Filtered σ: 1.57 dB;

Figure E.7. Frequency-domain error results for RTCS representing Reverberation Chamber with 08 wedges

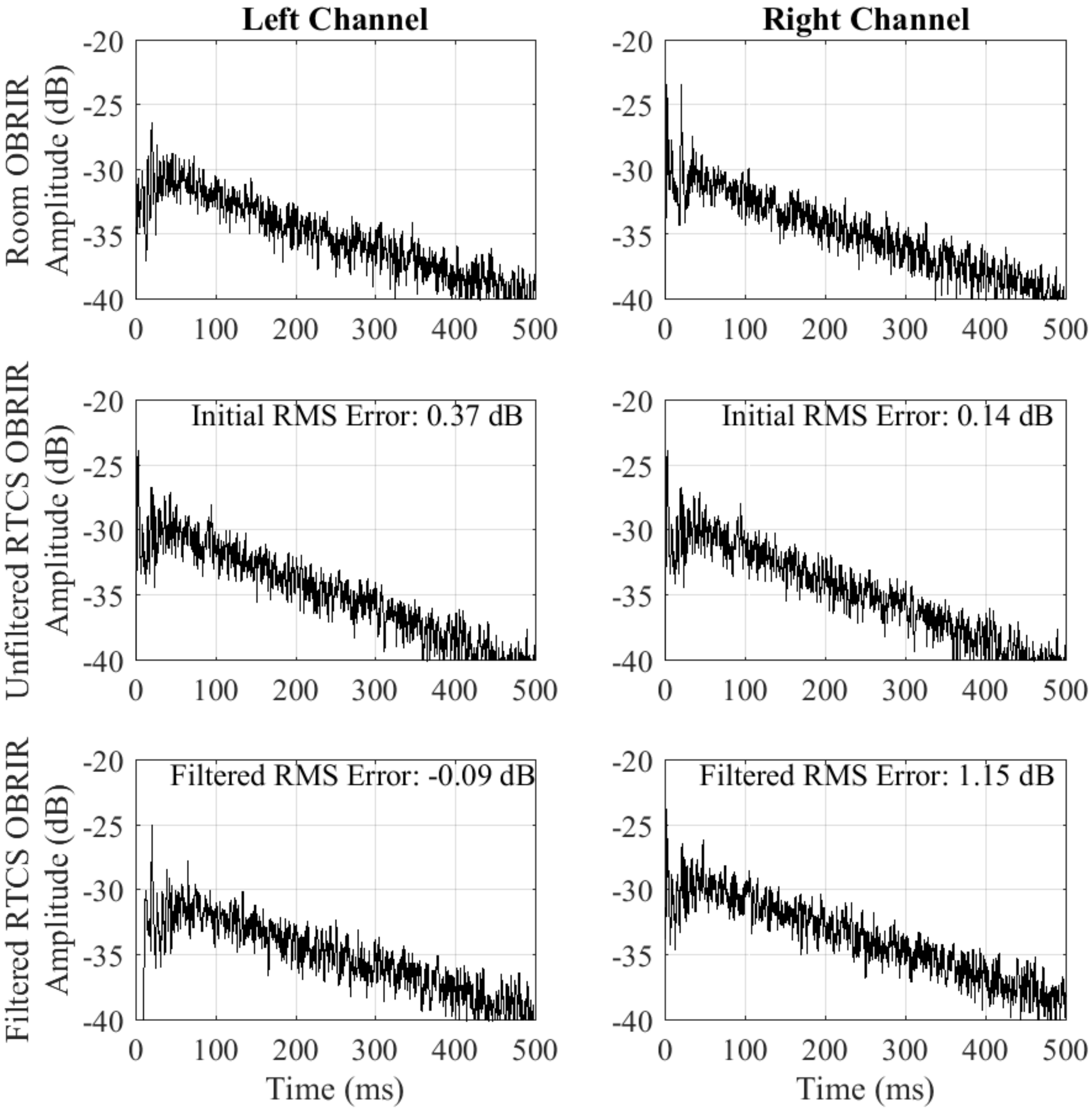


Figure E.8. Time-domain error results for RTCS representing Reverberation Chamber with 08 wedges.

E.6 Reverberation Chamber, 16 wedges

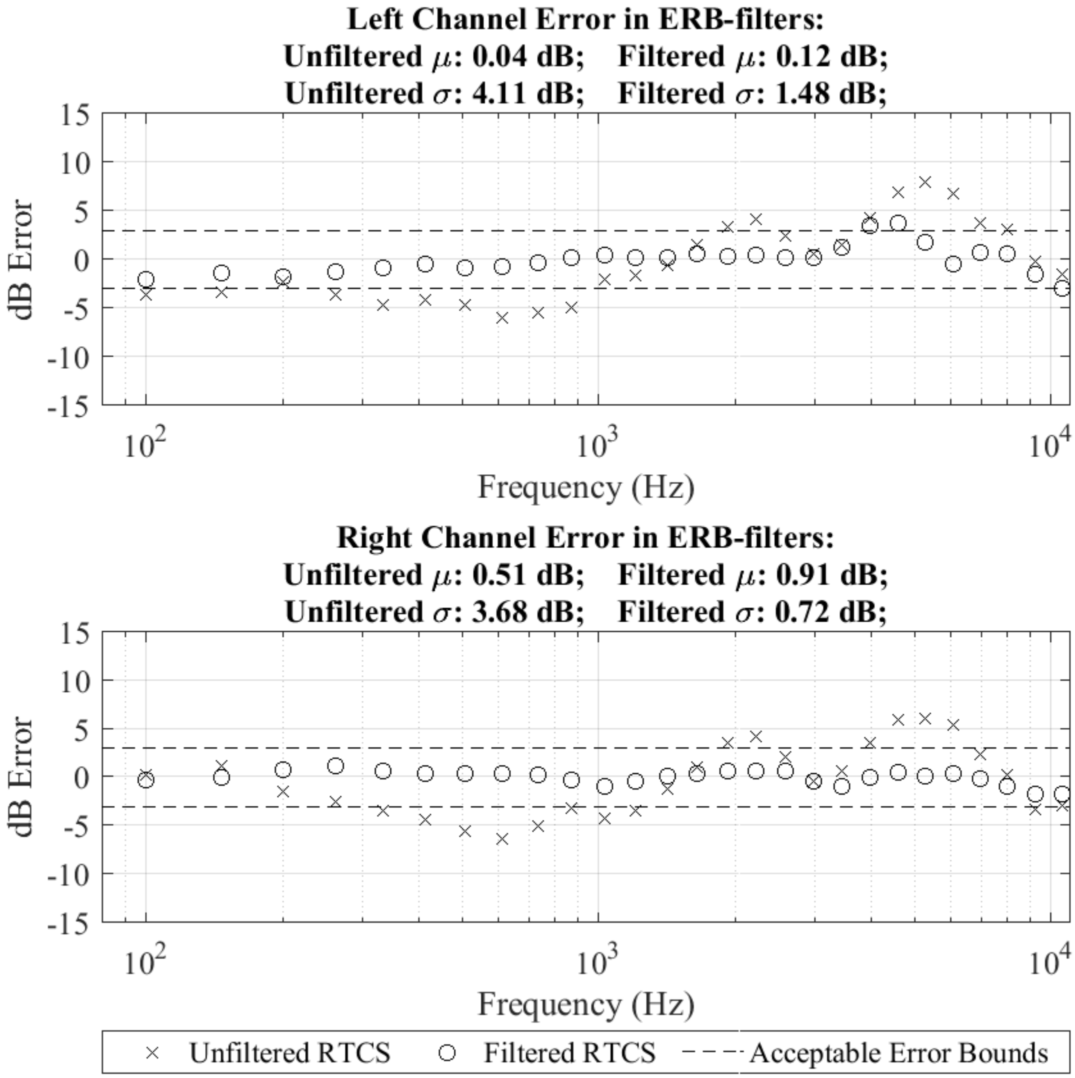


Figure E.9. Frequency-domain error results for RTCS representing Reverberation Chamber with 16 wedges.

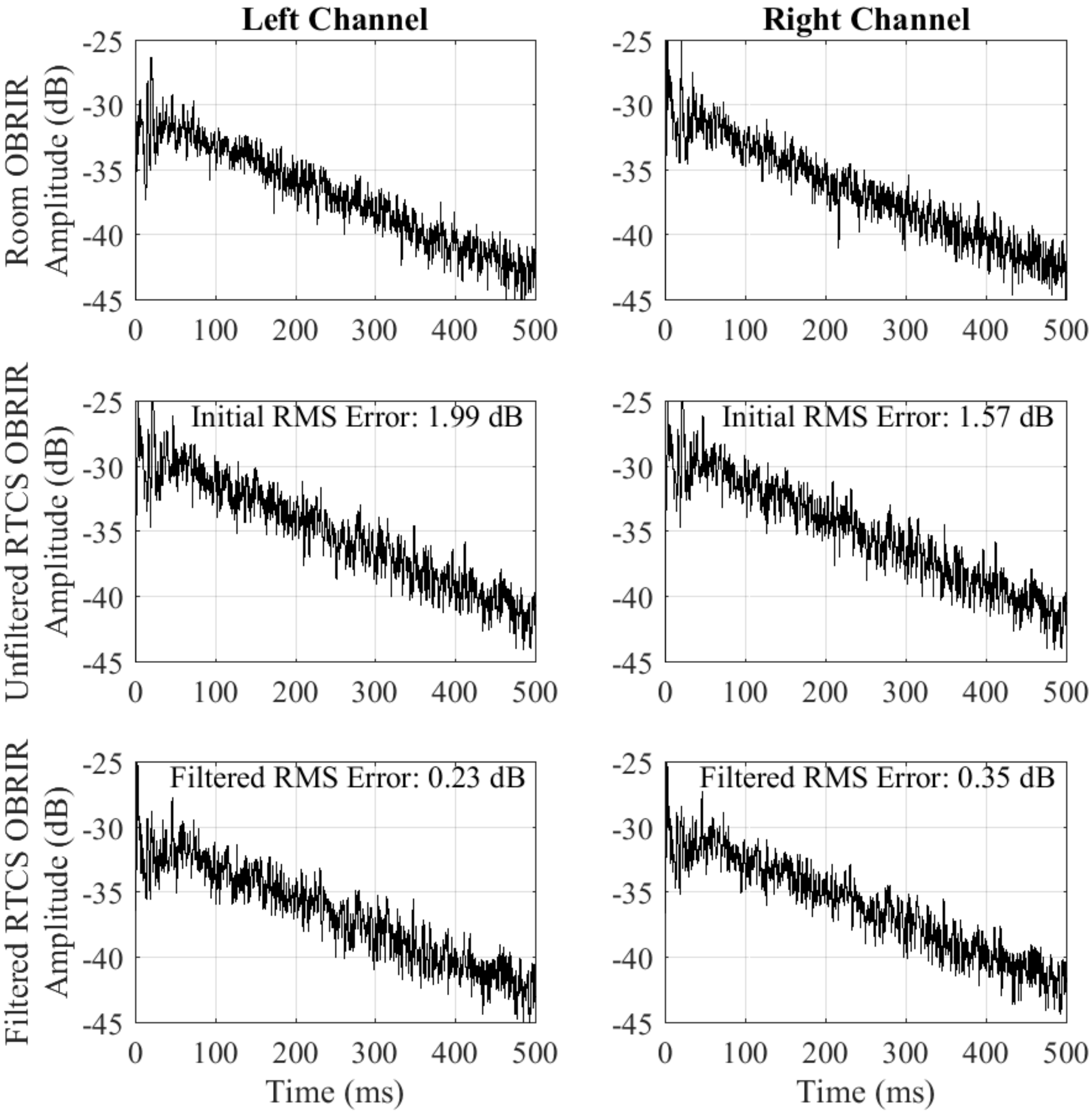


Figure E.10. Time-domain error results for RTCS representing Reverberation Chamber with 16 wedges.

E.7 Reverberation Chamber, 24 wedges

.

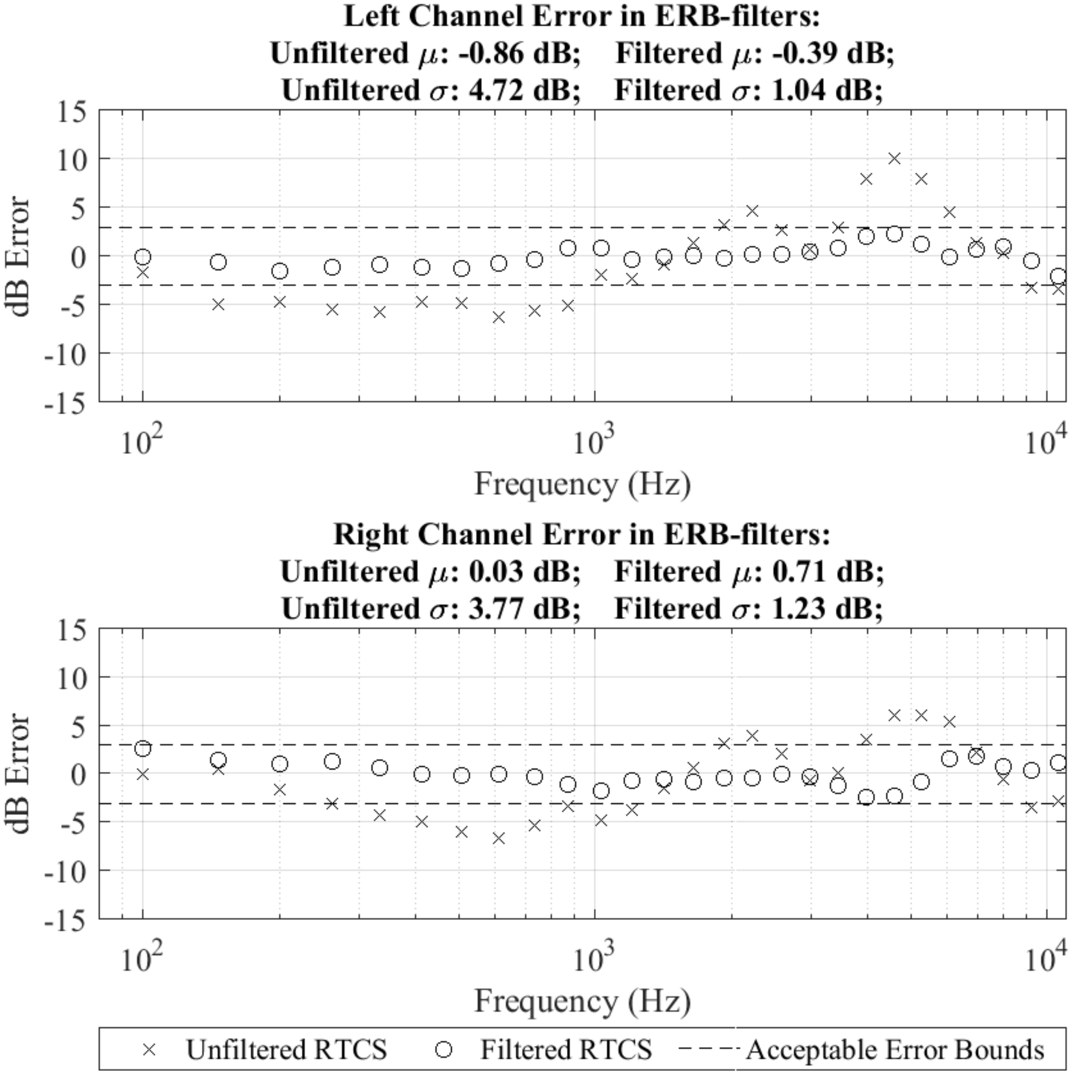


Figure E.11. Frequency-domain error results for RTCS representing Reverberation Chamber with 24 wedges.

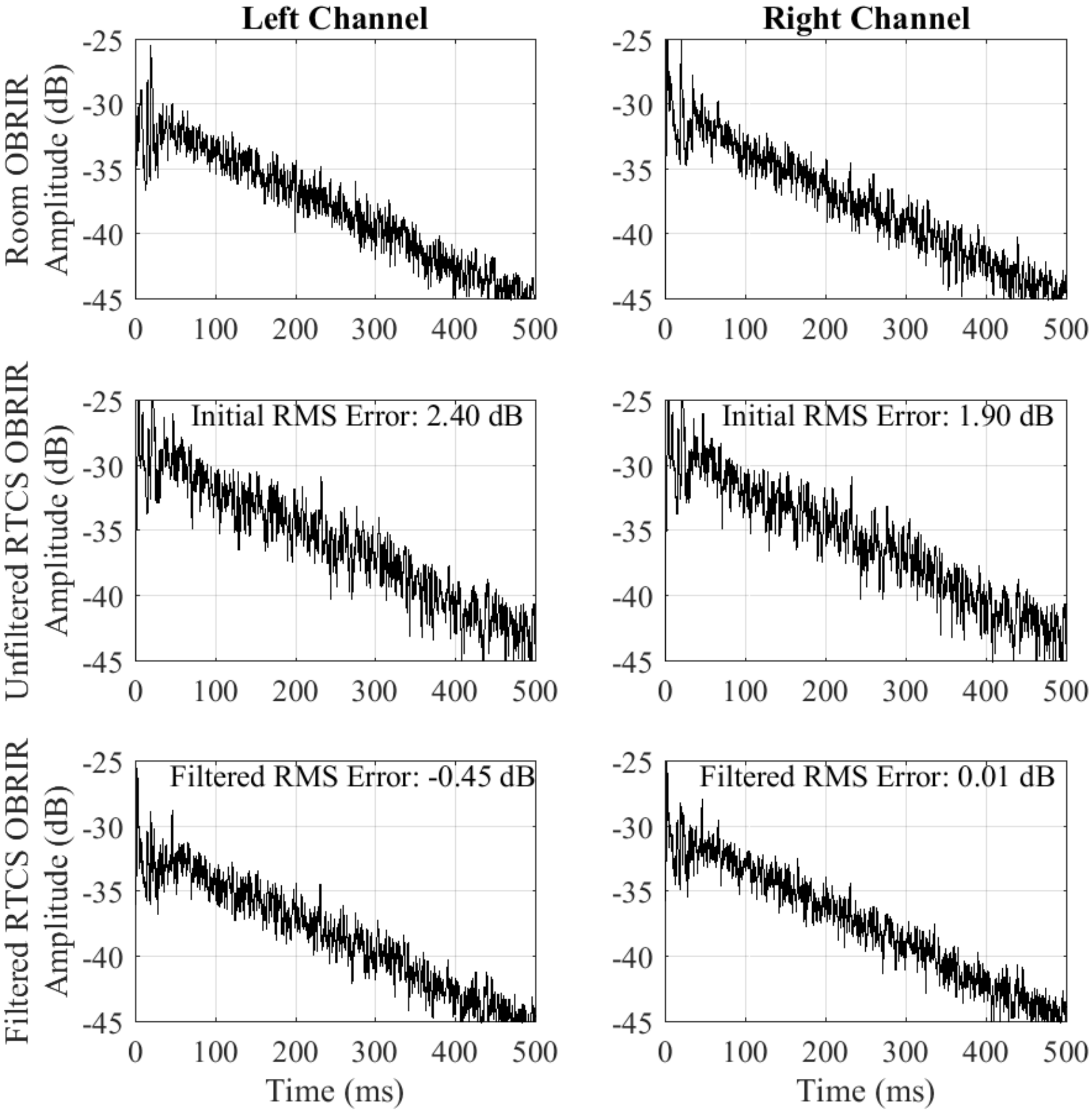


Figure E.12. Time-domain error results for RTCS representing Reverberation Chamber with 24 wedges.

E.8 de Jong Concert Hall

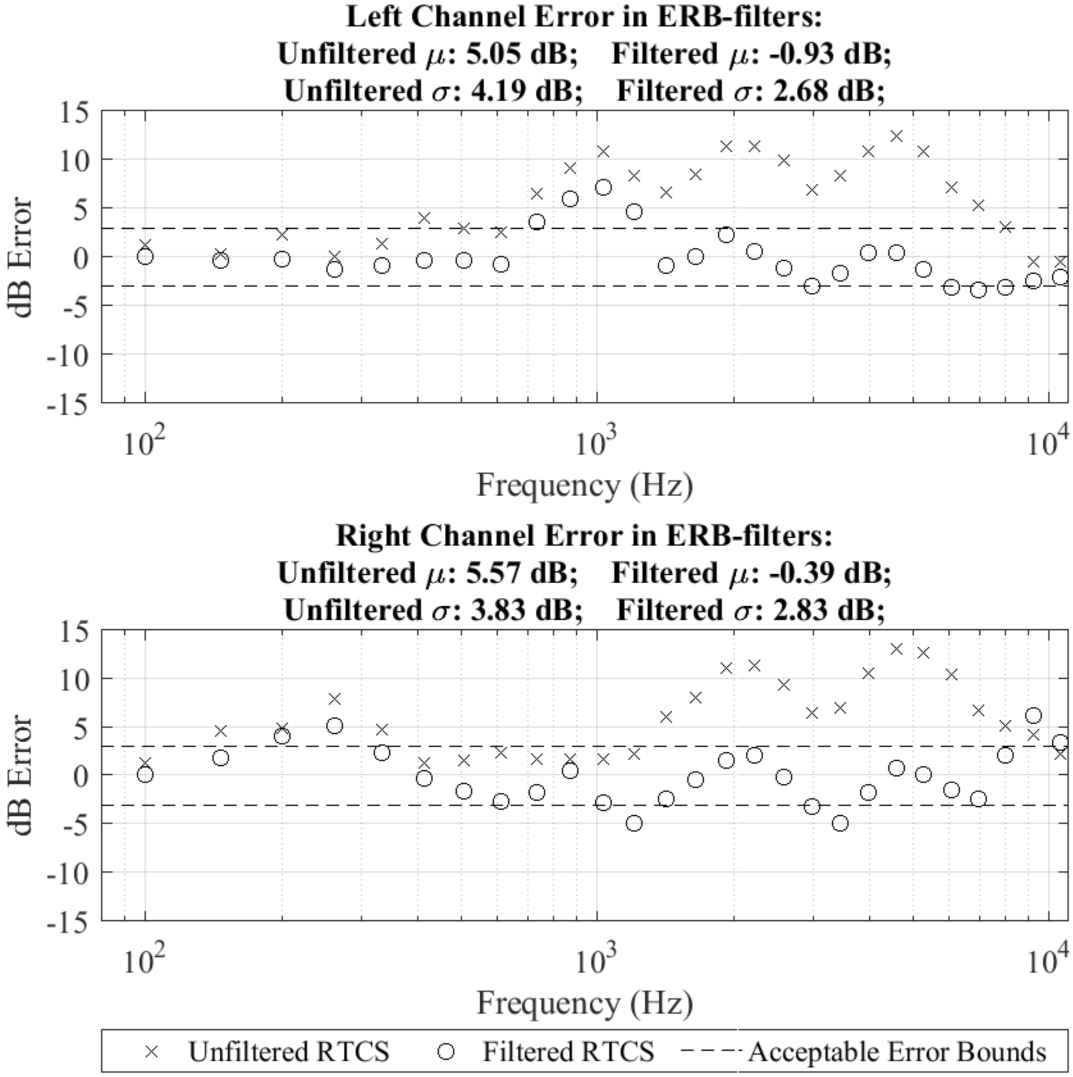


Figure E.13. Frequency-domain error results for RTCS representing the de Jong Concert Hall. The standard deviation was reduced with the filter, but results as good as the reverberation chamber cases were not achievable. This could be due to the nature of the OBRIR, the filter length, or the RTCS level settings during the OBRIR measurements.

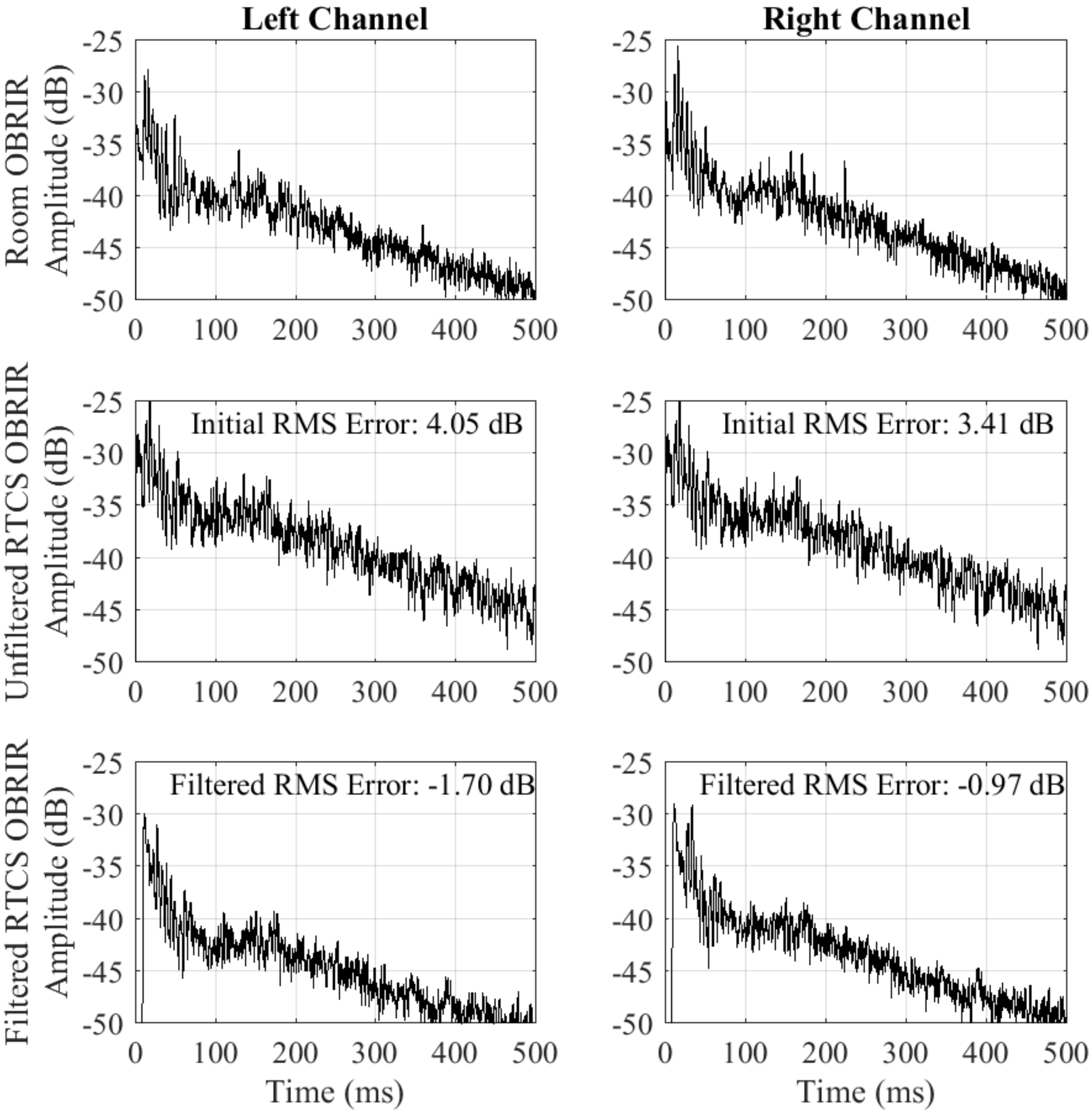


Figure E.14. Time-domain error results for RTCS the de Jong Concert Hall.

E.9 Simulated de Jong Concert Hall

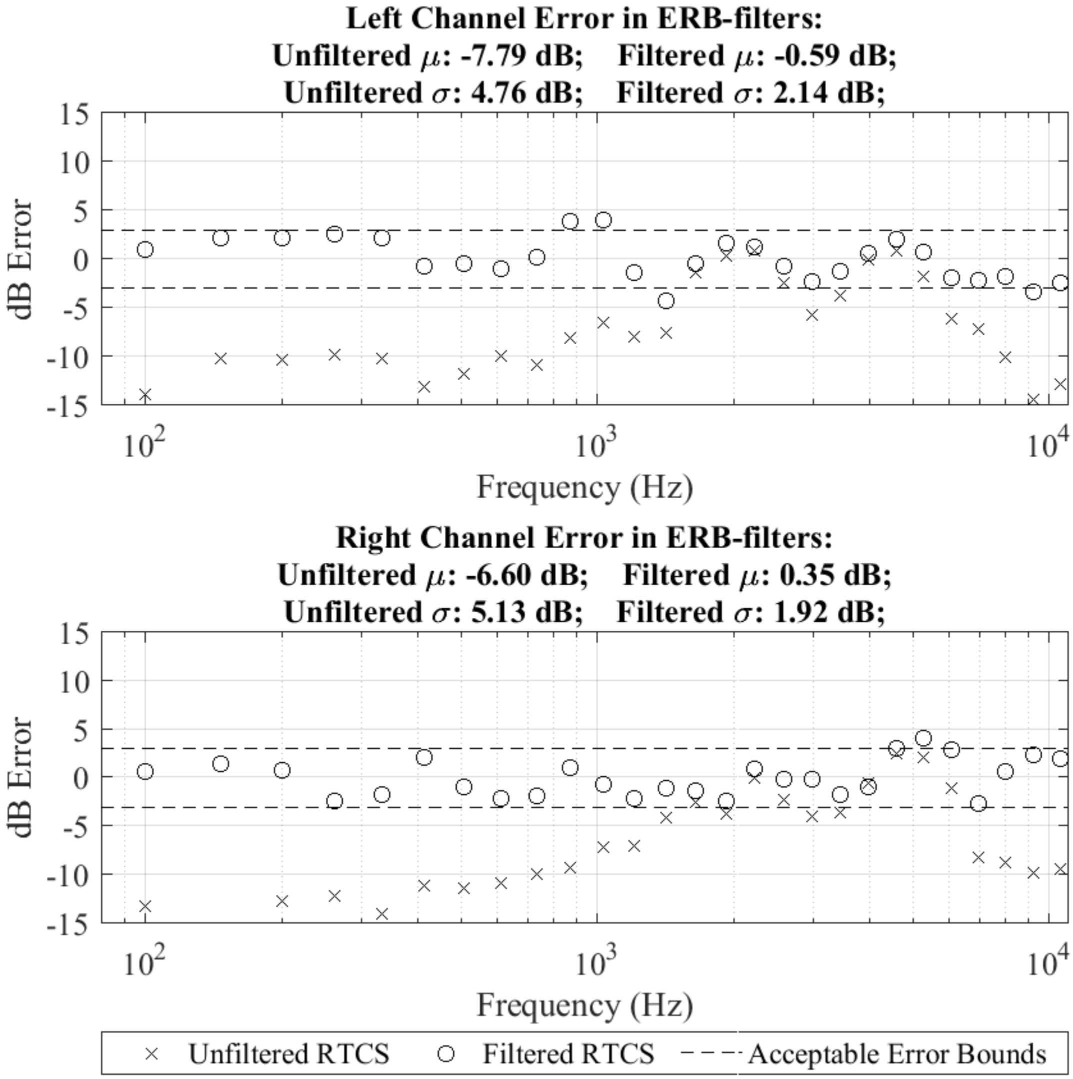


Figure E.15. Frequency-domain error results for RTCS representing the simulated de Jong Concert Hall.

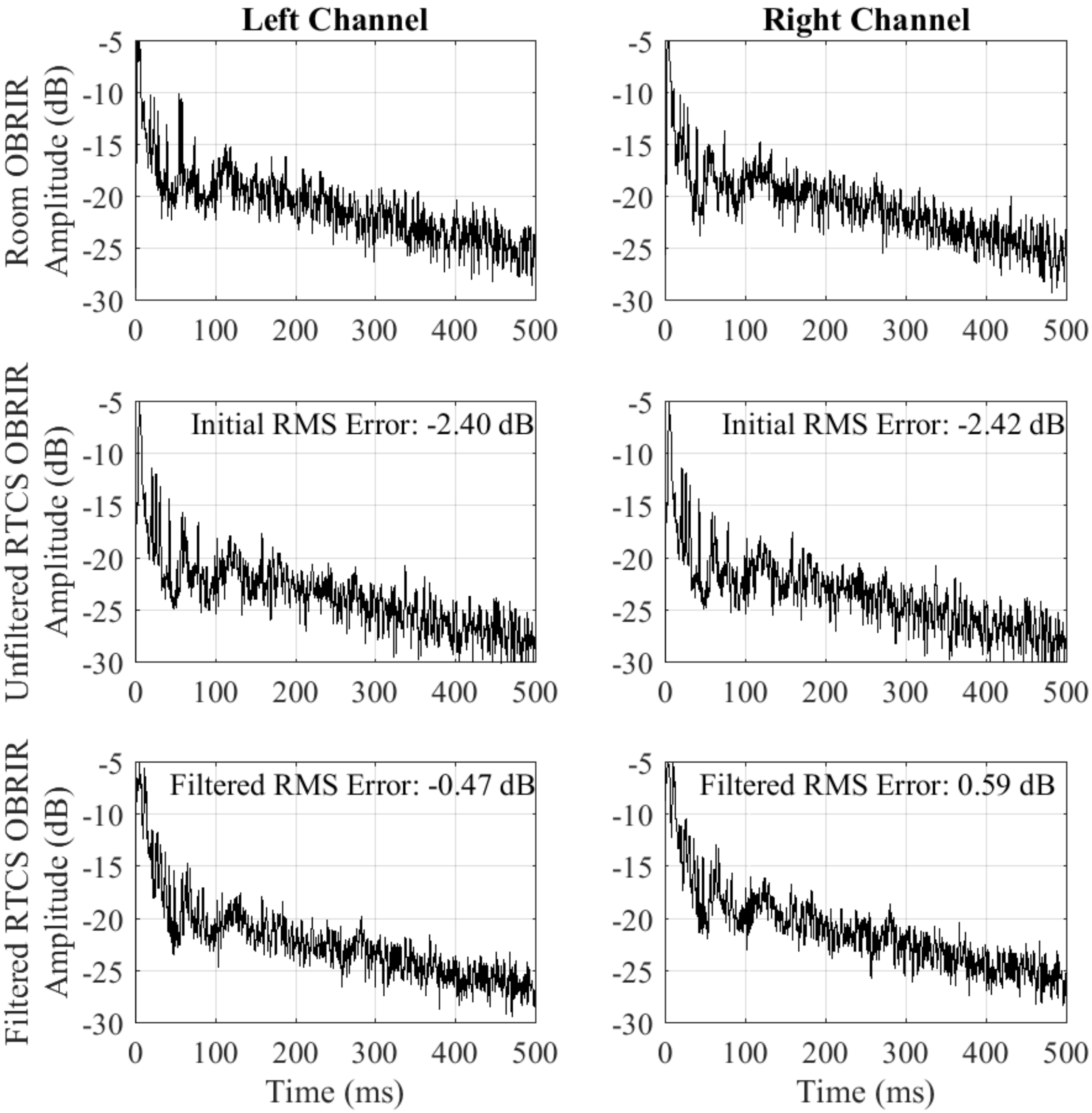
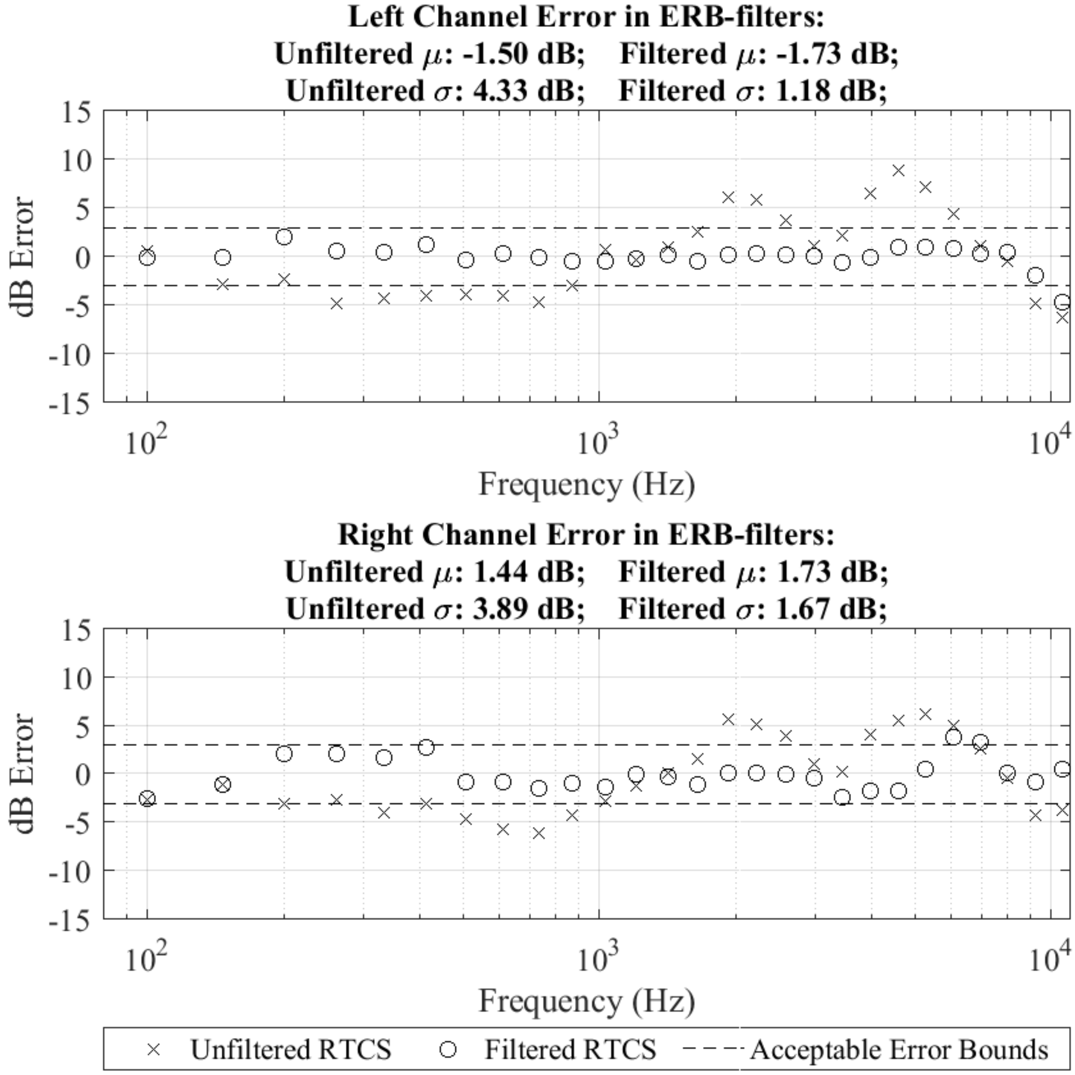


Figure E.16. Time-domain error results for the RTCS representing the simulated de Jong Concert Hall.

E.10 ESC C215

.


Left Channel Error in ERB-filters:
Unfiltered μ: -1.50 dB; Filtered μ: -1.73 dB;
Unfiltered σ: 4.33 dB; Filtered σ: 1.18 dB;

Right Channel Error in ERB-filters:
Unfiltered μ: 1.44 dB; Filtered μ: 1.73 dB;
Unfiltered σ: 3.89 dB; Filtered σ: 1.67 dB;

Figure E.17. Frequency-domain error results for RTCS representing ESC C215.

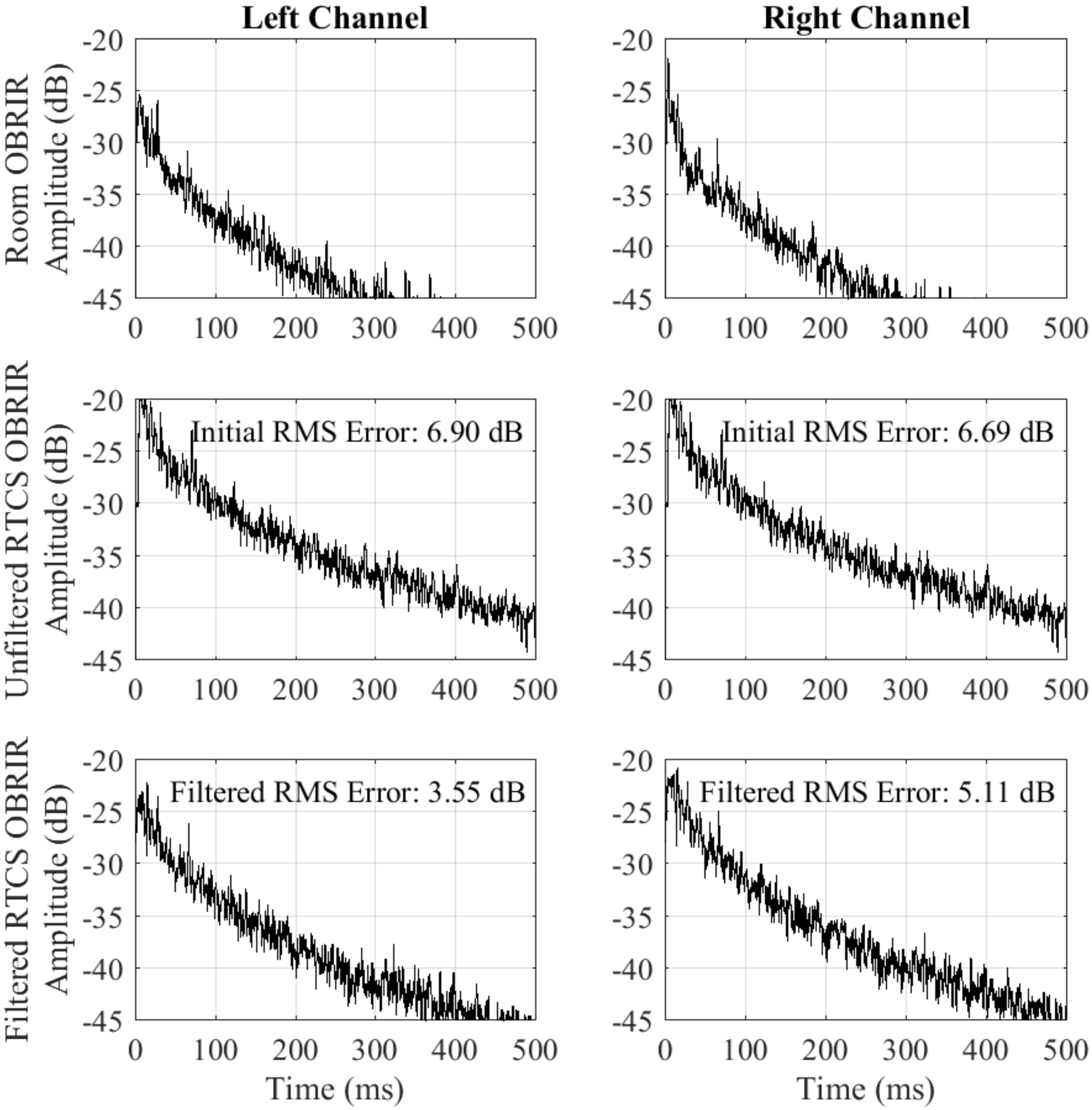


Figure E.18. Time-domain error results for the RTCS representing ESC C215. The discrepancy in the levels may be due to the tail after 500 ms, which does not have the same steep drop off rate.

E.11 ESC C261

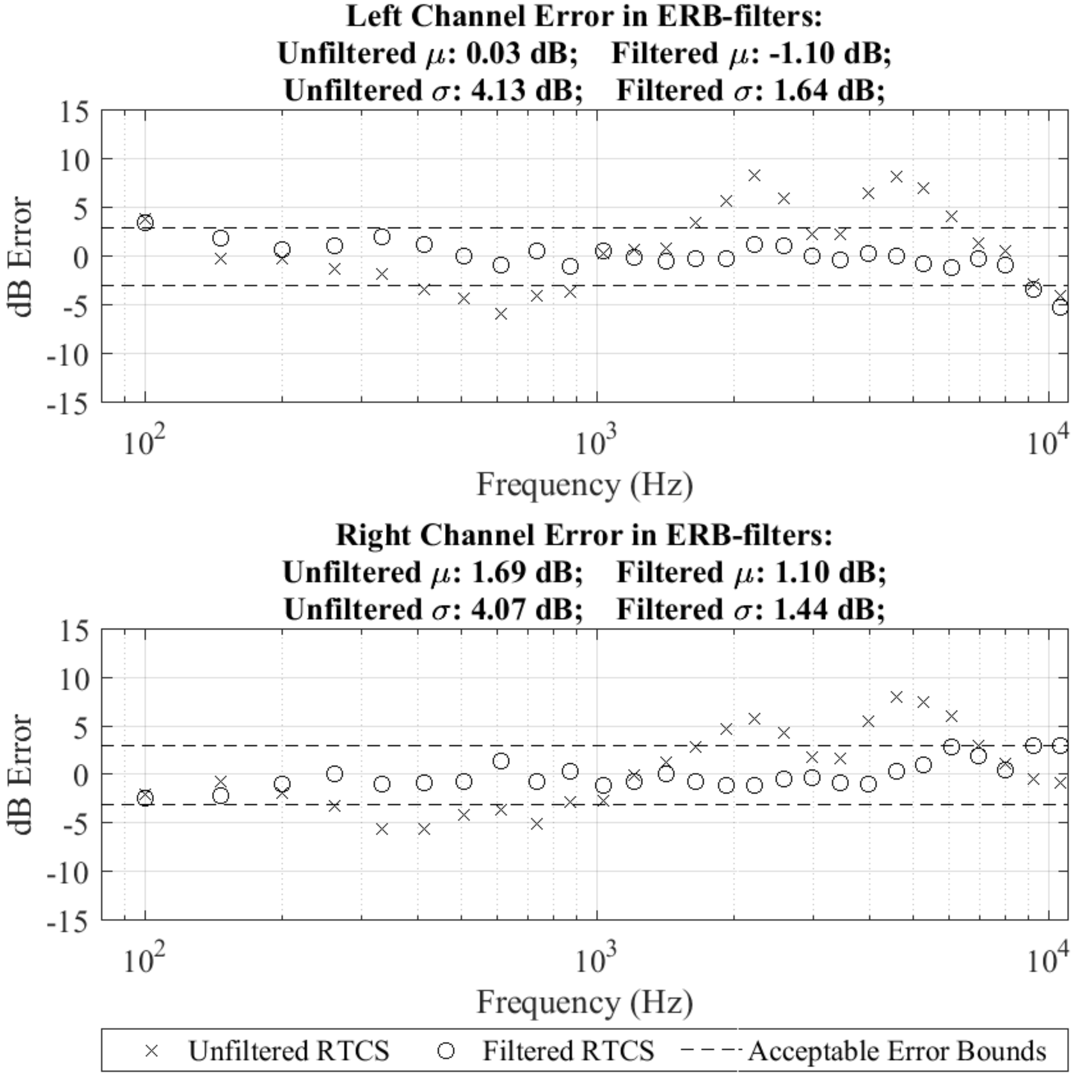


Figure E.19. Frequency-domain error results for RTCS representing ESC C261.

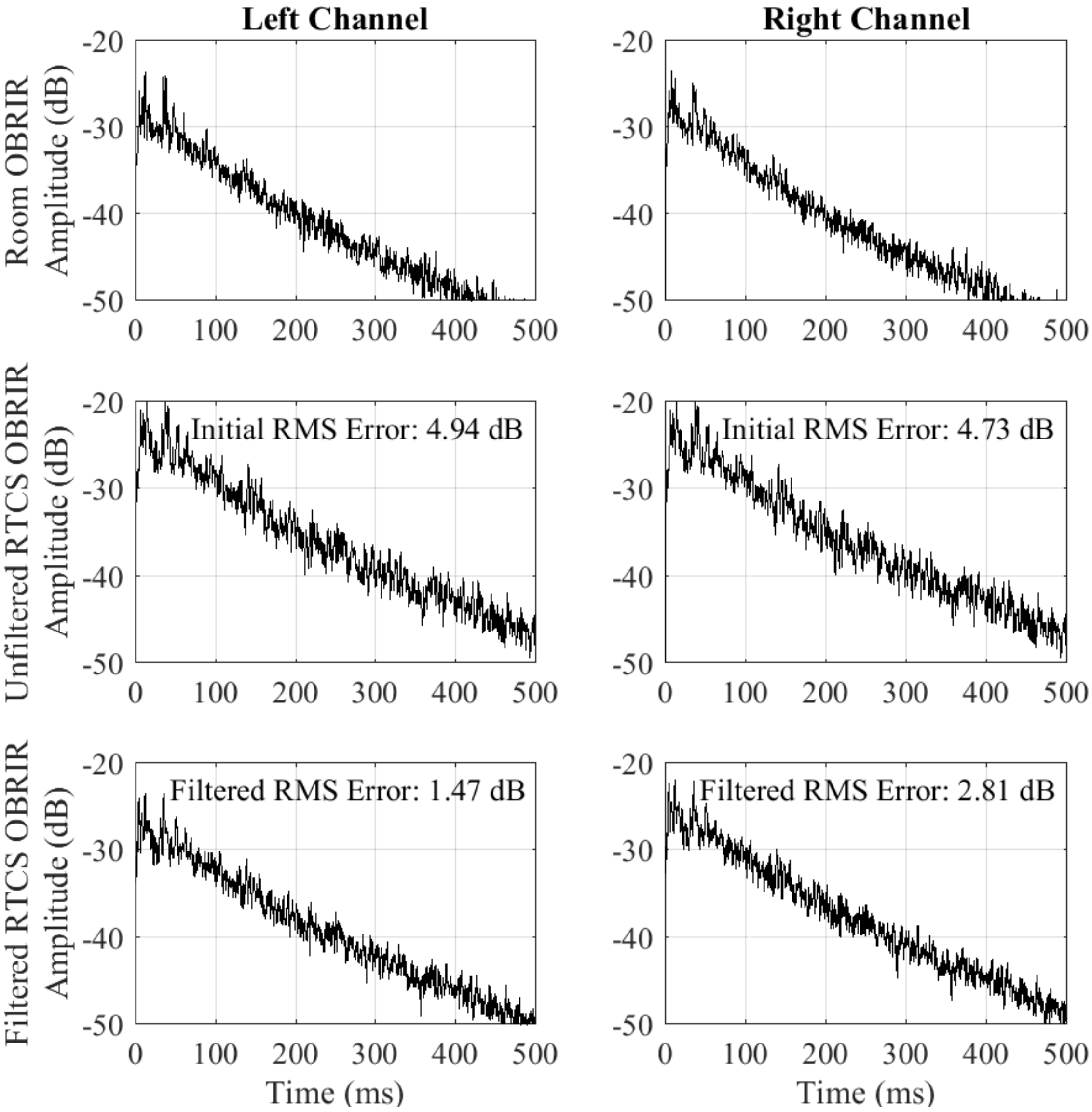


Figure E.20. Time-domain error results for the RTCS representing ESC C261 The decay profile after 100 ms is especially close to the room OBRIR for the filtered RTCS case.

E.12 Simulated ESC C261

.

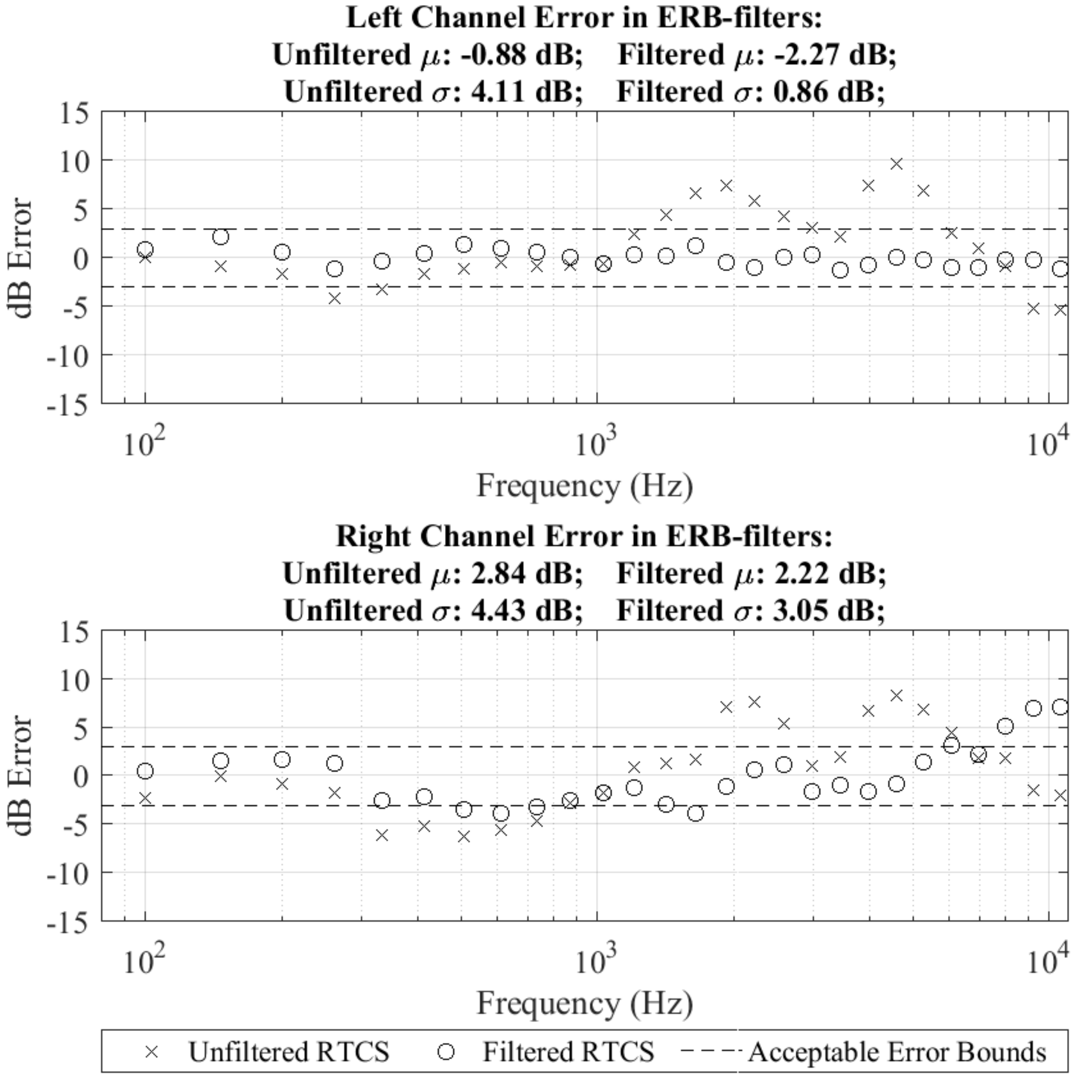


Figure E.21. Frequency-domain error results for RTCS representing the simulated classroom C261.

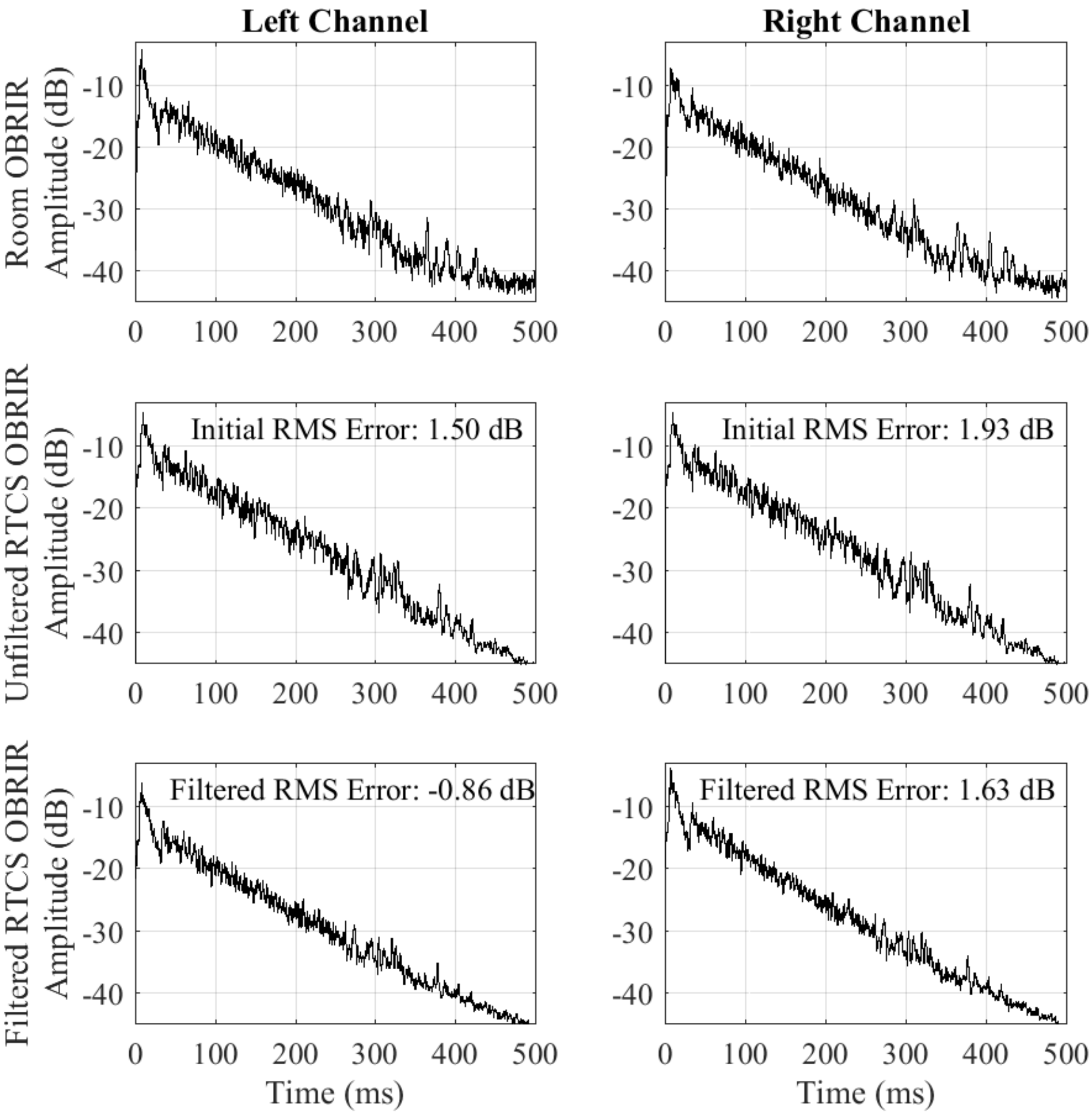


Figure E.22. Time-domain error results for the RTCS representing ESC C261.

E.13 Simulated ESC C261 with Early Reflections Removed

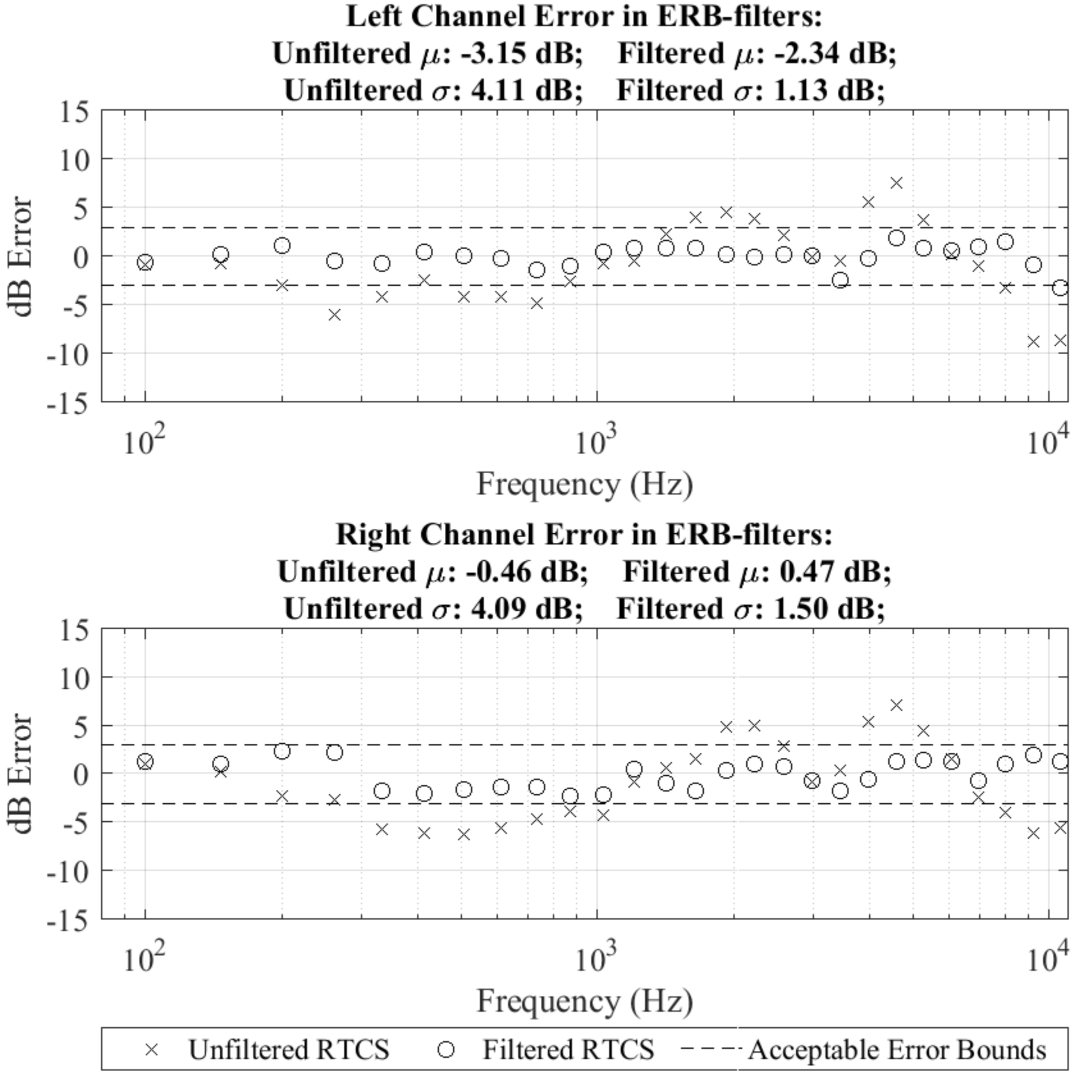


Figure E.23. Frequency-domain error results for RTCS representing the simulated classroom C261 with early reflections removed. As described in Section 3.5.2, this OBRIR was a modification of the simulated ESC C261 OBRIR to remove the earliest reflections.

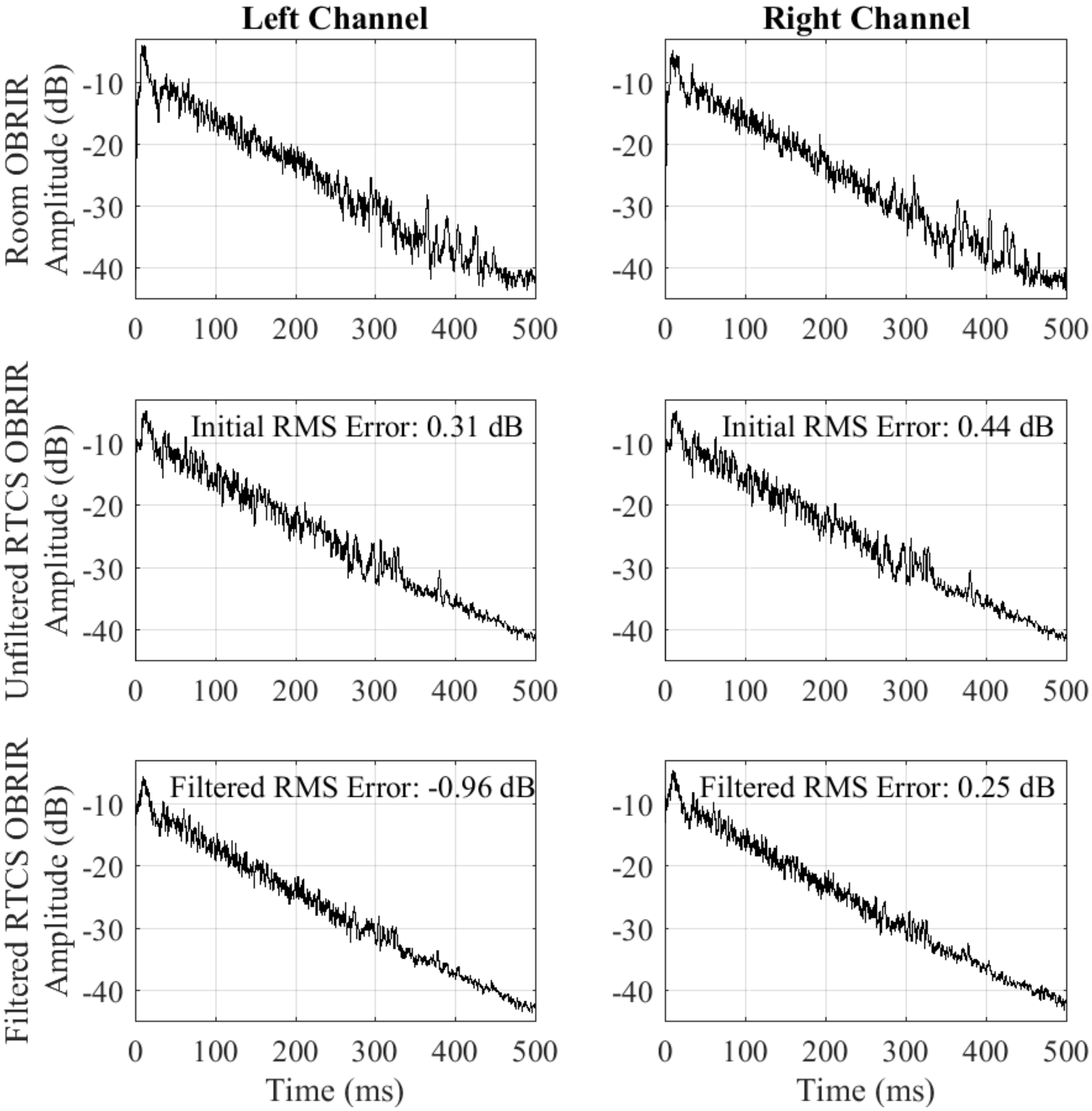


Figure E.24. Time-domain error results for the RTCS representing ESC C261 with early reflections removed. As described in Section 3.5.2, this OBRIR was a modification of the simulated ESC C261 OBRIR to remove the earliest reflections

Appendix F

RTCS Settings and Randomization

Room OBRIR	SIR2 Level (dB)	Compensation Filter	SIR2Level (dB)
Reverb 00	-18	Filter_Reverb_00	-5.8
Reverb 02	-18.72	Filter 11.7 Reverb 02	-6.8
Reverb 04	-15.9	Filter_Reverb04	-5.8
Reverb 08	-16.5	Filter_Reverb_08	-3.3
Reverb 16	-17.5	Filter_Reverb_16	-3.3
Reverb 24	-15.9	Filter_Reverb_24	-3.3
Reverb 32	-10.4	Filter_Reverb_32	-5.3
C215	-13.62	Filter_C215	-6.8
de Jong	-14.51	Fliter_de Jong 10.31	-13
de JongSF	-17	Filter11.1 DEJOSF	-18
C261SF	-8.4	Filter_C261SF_01	-5.8
C261ABSF	-21	Filter11.1	-9.9
C261	-17.52	Filter_C261	-4.3

Participant # Trial #									09	10
01									RE00M	RE00M
02									RE16M	RE16M
03									RE08M	RE32M
04									RE32M	RE04M
05									C215M	DJCHS
06									RE02M	DJCHM
07									RE24M	RE02M
08									DJCHS	RE08M
09									DJCHM	C215M
10									RE04M	RE24M
11									ANCH	ANCH

Participant # Trial #	11	12	13	14	15	16	17	18	19	20
01	RE16M	RE00M	RE04M	RE32M	RE32M	DJCHM	RE08M	RE02M	RE16M	RE16M
02	RE24M	DJCHM	DJCHS	DJCHM	DJCHS	RE32M	RE32M	RE16M	RE32M	DJCHS
03	RE04M	RE16M	RE02M	RE24M	RE16M	RE24M	C215M	DJCHS	RE00M	RE04M
04	DJCHS	RE02M	RE00M	RE16M	RE04M	C215M	DJCHS	RE04M	RE08M	DJCHM
05	RE32M	DJCHS	RE32M	DJCHS	RE08M	RE04M	DJCHM	C215M	RE24M	RE00M
06	RE02M	RE04M	RE08M	C215M	RE00M	RE08M	RE04M	RE08M	C215M	C215M
07	C215M	RE08M	DJCHM	RE02M	RE02M	DJCHS	RE24M	RE24M	RE02M	RE24M
08	RE08M	RE32M	RE24M	RE04M	C215M	RE16M	RE02M	RE00M	RE04M	RE32M
09	RE00M	C215M	C215M	RE08M	RE24M	RE02M	RE00M	DJCHM	DJCHS	RE02M
10	DJCHM	RE24M	RE16M	RE00M	DJCHM	RE00M	RE16M	RE32M	DJCHM	RE08M
11	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH

Participant # Trial #	21	22	23	24	25	26	27	28	29	30
01	RE02M	RE32M	C215M	DJCHM	RE08M	DJCHM	RE24M	RE08M	RE02M	RE32M
02	RE16M	RE08M	RE02M	RE02M	RE32M	RE08M	C215M	RE32M	RE00M	DJCHM
03	RE32M	DJCHM	RE24M	RE16M	RE04M	RE04M	RE32M	DJCHM	RE08M	RE00M
04	RE04M	DJCHS	DJCHM	DJCHS	RE24M	RE24M	DJCHM	RE00M	RE16M	RE02M
05	RE24M	RE24M	RE00M	RE24M	RE02M	RE02M	RE02M	RE02M	RE32M	RE08M
06	DJCHS	C215M	RE04M	RE04M	RE00M	DJCHS	RE08M	RE24M	C215M	RE24M
07	C215M	RE04M	RE16M	RE00M	DJCHM	RE16M	RE04M	C215M	DJCHM	C215M
08	RE00M	RE00M	RE08M	RE08M	RE16M	RE32M	DJCHS	DJCHS	RE04M	RE16M
09	DJCHM	RE16M	DJCHS	RE32M	C215M	C215M	RE00M	RE04M	DJCHS	DJCHS
10	RE08M	RE02M	RE32M	C215M	DJCHS	RE00M	RE16M	RE16M	RE24M	RE04M
11	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH

Participant # Trial #	31	32	33	34	35	36	37	38	39	40
01	RE24M	DJCHM	DJCHM	RE24M	RE08M	RE32M	RE16M	DJCHS	C215M	RE08M
02	RE02M	RE32M	DJCHS	RE16M	C215M	RE02M	RE02M	RE02M	RE00M	C215M
03	RE04M	RE24M	RE08M	DJCHM	RE04M	RE16M	RE08M	RE00M	RE02M	DJCHS
04	DJCHM	RE00M	RE02M	RE00M	RE32M	RE00M	RE32M	RE08M	RE24M	RE04M
05	DJCHS	RE16M	RE16M	RE08M	DJCHM	C215M	DJCHM	C215M	RE16M	RE16M
06	RE08M	RE02M	RE24M	RE04M	RE00M	RE04M	RE04M	RE32M	DJCHS	RE02M
07	RE00M	DJCHS	C215M	RE32M	DJCHS	DJCHS	DJCHS	RE16M	RE08M	RE32M
08	RE32M	C215M	RE04M	RE02M	RE16M	DJCHM	RE00M	RE04M	DJCHM	DJCHM
09	C215M	RE04M	RE00M	DJCHS	RE02M	RE08M	RE24M	DJCHM	RE04M	RE24M
10	RE16M	RE08M	RE32M	C215M	RE24M	RE24M	C215M	RE24M	RE32M	RE00M
11	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH

Participant # Trial #	41	42	43	44	45	46	47	48	49	50
01	C215M	RE00M	RE04M	DJCHM	RE08M	RE08M	RE08M	DJCHM	RE04M	C215M
02	RE04M	RE04M	RE16M	RE00M	DJCHM	RE24M	RE04M	RE24M	RE08M	RE02M
03	RE02M	RE02M	RE02M	DJCHS	RE04M	RE04M	RE00M	RE08M	RE24M	RE16M
04	RE16M	RE24M	RE00M	RE16M	C215M	RE16M	RE32M	RE04M	RE16M	DJCHM
05	DJCHM	RE08M	C215M	RE02M	DJCHS	DJCHM	RE02M	RE00M	C215M	RE04M
06	DJCHS	RE32M	RE32M	C215M	RE24M	DJCHS	RE24M	C215M	DJCHS	RE00M
07	RE32M	DJCHS	RE08M	RE04M	RE00M	RE02M	C215M	DJCHS	RE00M	RE24M
08	RE08M	DJCHM	DJCHS	RE08M	RE16M	RE32M	RE16M	RE02M	DJCHM	RE08M
09	RE00M	C215M	DJCHM	RE24M	RE02M	RE00M	DJCHS	RE16M	RE32M	RE32M
10	RE24M	RE16M	RE24M	RE32M	RE32M	C215M	DJCHM	RE32M	RE02M	DJCHS
11	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH	ANCH

Appendix G

Vocal Effort Study Filenaming Protocol

Reaper Project Name Set project settings → media → path to save in a folder of the same name as the project

1-2:	**gd**	(gender differences grant)
3:	**B**	(BYU study)
4:	M *or* F	(gender)
5-6:	##	(2 digit ID number) (01-32)

Reaper Track Name

7:	**h** ***or*** **n** ***or*** **a**	(head-worn mic / neck mic / accelerometer)
8-11:	####	(room code, see list below)
12:	**M** ***or*** **S**	(Measured or Simulated OBRIR in use)
13-14:	##	(trial number, 01-13)

Examples of raw recording filenames:
gdBF01hRE00M01.wav
gdBM31aC215S05.wav

Trimmed recordings Two more letters will be appended, indicating which speech task was included

15-16: two capital letter indicating the speech task.
AH - sustained /ɑ/ (middle 3 seconds of the second recording unless exception)
DE - describing a picture, ~45-60 seconds (spontaneous speech)
OQ - answering an open question, ~45-60 seconds (spontaneous speech)
R2 - rainbow passage sentences 2-3
RB - rainbow passage (sentences 1-6)

Example of trimmed recording filename:
gdBM31aC215S05R2.wav

Room Name	Room Code
Reverberation Chamber, 0 wedges	RE00
Reverberation Chamber, 2 wedges	RE02

Reverberation Chamber, 4 wedges	RE04
Reverberation Chamber, 8 wedges	RE08
Reverberation Chamber, 16 wedges	RE16
Reverberation Chamber, 24 wedges	RE24
Reverberation Chamber, 32 wedges	RE32
de Jong Concert Hall	DJCH
ESC Lecture Hall C215	C215
ESC Classroom C261	C261
ESC Classroom C261 with absorbing panels	C2AB

Appendix H

Speech Parameters in Vocal Effort Study

Fo_mean_RB:	Mean fundamental frequency as measured in Rainbow Passage reading. Unit: Hz.
Fo_std_RB	Standard deviation in fundamental frequency as measured in Rainbow Passage reading. Unit: Hz.
Ps_mean_RB	Mean Pitch Strength as measured in Rainbow Passage Reading. Unit: dB.
Ps_std_RB	Standard deviation of pitch strength as measured in rainbow passage reading. Unit: dB.
dB_mean_RB	Mean decibel level during rainbow passage reading. Unit: dB.
dB_mean_RB norm to ANCH	Mean decibel level normalized to that of each subject’s anechoic chamber, 11th trial. Unitless
dB_mean_RB norm to C215	Mean decibel level normalized to that of each subject’s C215 trial. Unitless
dB_std_RB	Standard deviation of decibel level as measured in rainbow passage reading. Unit: dB.
ActyFact_RB	Activity factor during rainbow passage reading. Measurement of speech-to-silence ratio. Unitless, 0-1 scale.
AlphaRto_RB	Ratio of the spectral energy above 1kHz and below 1kHz
dBspcSpF_RB	Spectral slope from fundamental frequency
STSD_RB	Semi-tone standard deviation. Another measure of spectral deviations, but comparable amongst males and females.
DurOTas2_R2	Task Duration of the rainbow passage sentences 2 and 3. Units: seconds
syl_rate_R2	Syllable rate: 29 syllables / DurOTas2_R2 provides syllable rate for rainbow passage sentences 2 and 3. Units: syllables / second
STSD_R2	Semi-tone standard deviation for rainbow passage sentences 2 and 3

CCPS_R2AH	Smoothed Cepstral Peak Prominence. Used in Calculating AVQI. Unit: dB
AVQI_R2AH	Acoustic Voice Quality Index: Calculated from concatenated Rainbow Passage sentences 2 and 3 and sustained vowel AH speech tasks. According to Reynolds, a value greater than 3.5 is indicative of voice dysphonia [37,122,123].
Fo_mean_AH	Mean fundamental frequency as measured in sustained vowel AH. Unit: Hz.
Fo_std_AH	Standard deviation in fundamental frequency as measured in sustained vowel AH. Unit: Hz.
Ps_mean_AH	Mean Pitch Strength as measured in sustained vowel AH. Unit: dB.
Ps_std_AH	Standard deviation of pitch strength as measured in sustained vowel AH. Unit: dB.
dB_mean_AH	Mean decibel level during sustained vowel AH. Unit: dB.
dB_std_AH	Standard deviation of decibel level as measured in sustained vowel AH. Unit: dB.
jitter_AH	Jitter is the deviation from true periodicity of a presumably periodic signal. This is the average absolute difference between consecutive periods of the sustained vowel, divided by the average period.
shimmer_AH	This is the average absolute difference between the amplitudes of consecutive periods of the sustained vowel, divided by the average amplitude [125].
HNR_AH	A Harmonicity object represents the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR). Harmonicity is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is 10*log10(99/1) = 20 dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise.
dB_mean_DE	Mean decibel level during spontaneous speech task. Unit: dB.
dB_std_DE	Standard deviation of decibel level during spontaneous speech task. Unit: dB.
ActyFact_DE	Activity Factor during spontaneous speech task