Audio Engineering Society

Convention Paper

Presented at the 147th Convention
2019 October 16 – 19, New York

Measurement of oral-binaural room impulse response by singing scales

Munhum Park[1]

[1]*Institute of Music, Science and Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand*

Correspondence should be addressed to Munhum Park (munhum.pa@kmitl.ac.th)

ABSTRACT

Oral-binaural room impulse responses (OBRIRs) are the transfer functions from mouth to ears measured in a room. Modulated by many factors, OBRIRs contain information for the study of stage acoustics from the performer?s perspective and can be used for the auralization. Measuring OBRIRs on human is, however, a cumbersome and time-consuming process. In the current study, some issues of the OBRIR measurement on human were addressed in a series of measurement. With in-ear and mouth microphones, volunteers sang scales, and a simple post-processing scheme was used to refine the transfer functions. The results suggest that OBRIRs may be measured consistently by using the proposed protocol, where only 4~8 diatonic scales need to be sung depending on the target signal-to-noise ratio.

1 Introduction

Binaural room impulse responses (BRIRs) [or frequency responses (BRFRs)] refer to the head-related transfer functions (HRTFs) measured in a reverberant room from a sound source to two ears [1]. BRIRs contain all information about the direct and reflected sounds arriving at the listener's ears, the latter of which are heavily influenced by the room design and the materials used for the room interior. Therefore, BRIRs are typically measured from the most likely location of sound source (e.g., stage or podium) to one or more representative positions in the audience area, and when analyzed, the perceived acoustic properties of the room can effectively be investigated from the audience perspective.

Although the acoustic properties of a room as perceived by the audience may be of great importance, those perceived by the performers on the stage can never be discounted, who naturally adapt the manner of their speaking, singing and instrument-playing depending on the stage acoustics [2, 3]. Especially for speakers and singers, the transfer functions from the mouth to the ears characterize the airborne sounds that they themselves produce and hear, which are referred to as oral-binaural room impulse responses (OBRIRs) [or frequency responses (OBRFRs)] [4]. Once measured, OBRIRs can shed light on many aspects of stage acoustics from the performer's point of view, including, e.g., the perceived room size [5] and loudness of own voice [6]. As is the case with BRIRs, OBRIRs either measured or synthesized can also be used to *auralize* a particular stage, providing a virtual acoustic environment where, for example, the performer's preference of stage acoustics may effectively be studied [7].

Unlike BRIRs often measured on human, OBRIRs have mostly been obtained by using a head and torso simulator (HATS) [8]. Although the OBRIRs measured on

a HATS may be suitable to study the general acoustic properties of a stage, it may not effectively address the perception of individual performer whose OBRIRs (and HRTFs) are different from those of the HATS and others, if not unique. When used for the purpose of auralization, especially over headphones, the virtual acoustic scene of the stage may not be sufficiently convincing, which is one of the well-known problems with HRTF and its derivatives measured on a HATS [9].

Measuring OBRIRs on human head can be more time-consuming and cumbersome than on a HATS with some technical issues to be considered. To begin with, no clean source signal is available, and the voice recording made in the vicinity of mouth is the best alternative. Also, the spectrum of human voice is limited in frequency band, and more importantly, it consists of a fundamental with harmonic and non-harmonic components, which are distributed rather sparsely on the frequency axis. Unlike a short sine sweep played back on HATS, therefore, a rather long sequence of syllables at varying pitch might have to be spoken or sung to ensure that the input to the acoustic system (room) may have an appropriate level of signal-to-noise ratio over the frequency range of interest. Due to the inevitable movement of body parts, however, a relatively long period of measurement on participant may undesirably result in spatially-averaged OBRIRs. Accordingly, the key challenge for the OBRIR measurement on human head is to produce spectrally rich input sounds in the shortest possible time or to make a reasonable balance between 'spectrum' and time.

In the current study, a stepwise approach was taken to address the issues introduced in the preceding paragraph regarding the OBRIR measurement on human. In the first measurement described in section 2, recorded singing voices and a sine sweep were used and compared as the excitation signal for the impulse response measurement. Then, the OBRIR measurement was carried out on participants as described in section 3, where individual notes on four chromatic scales were sung in sequence. In the last measurement presented in section 4, a complete diatonic scale was sung at a time to shorten the recording period, and the results were analyzed in search of the optimal measurement protocol.

2 Singing voice for excitation signal

2.1 Methods

Four undergraduate students (2 males; 2 females) volunteered for the recording session in a quiet recording studio at the department. First, the lowest and the highest notes they could sing with comfort were identified. Assisted by the guide tone reproduced over headphones (SRH440, Shure) by a digital audio workstation (Logic Pro X, Apple), they then sang 'ah' for all notes of four chromatic scales within the range, of which the pitch differed by 25 cents from each other. In this way, all notes at 25-cent tone step could individually be recorded within the singers' vocal ranges. For the recording, a hands-free microphone (MX153, Shure) was positioned as close to the volunteer's lips as possible, and the output was saved at 44.1 kHz by using an audio interface (Fireface 802, RME).

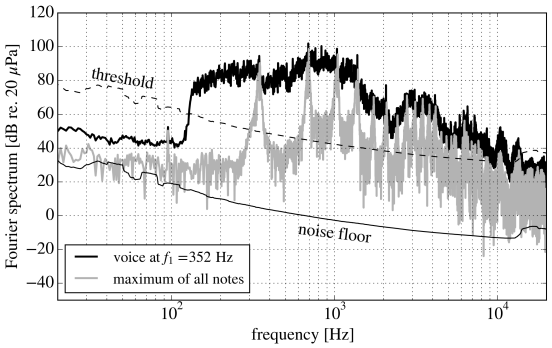


Fig. 1: An example spectrum of a singer's voice is shown in grey with the maximum obtained across all notes indicated in black. The threshold spectrum is determined at α dB above the noise floor.

The recorded voice was played back over a single-driver loudspeaker in one of the empty rooms in the recording studio, while the reproduced sound was being recorded by two microphones (M30, Earthworks), 'source mic' and 'room mic' placed at the distance of $10\,cm$ and $1\,m$ from the loudspeaker, respectively. A reference measurement was also made by playing back a sine sweep from 20 Hz to 20 kHz, recorded by the two microphones mentioned above.

Results were analyzed on Python. First, the spectrum of the noise floor was estimated from the quiet intervals

of the recordings. From the reference measurement, two different responses were determined depending on the input type: 1) The sine sweep as input and the room-mic recording as output; 2) the source-mic and the room-mic recordings as input and output, respectively. For the measurement made by using voice, the *raw* frequency responses from the source mic to the room mic were first determined for individual notes per singer, and one of the following three methods was used for the post-processing in the frequency domain:

- Process 1: The frequency responses were only averaged across all notes.
- Process 2: The maximum magnitude of the source-mic spectrum was calculated at each frequency bin across all notes, which was then compared to a threshold spectrum set at α dB above the spectrum of the noise floor (see Fig. 1). From this comparison, a usable frequency range with an appropriate level of signal-to-noise ratio ($> \alpha$ dB) was determined. The average frequency response (averaged across all notes) was considered to be valid only within the usable frequency range, and that out of this range was attenuated by applying a frequency-domain bandpass filter of which the magnitude response resembled that of the fourth-order Butterworth filter.
- Process 3: The threshold spectrum described in Process 2 was used to identify valid frequency *bins* (rather than a usable frequency range) in the source-mic spectrum for each note, which usually corresponded to those close to the harmonic frequencies of the note (see Fig. 1). The frequency responses estimated only in these frequency bins were considered valid and averaged across all notes. As a consequence, the frequency response estimate may not exist especially at very low or high frequencies, typically outside the usable frequency range described in Process 2. For these low and high frequency ranges, therefore, the frequency response estimated by using Process 2 was substituted.

After the post-processing, the time-domain impulse response was determined for each singer by using the inverse Fourier transform.

2.2 Results

Figure 2 shows four frequency responses measured by using a loudspeaker and one or two microphones, where the value of α in Process 3 was set at 20 dB. The reference response measured with the sine-sweep *signal* as input shows a gradual decay at low frequencies below ~80 Hz, attributed to the roll-off of the transducer (loudspeaker & microphone) response. The other reference response measured with the same signal but from the source mic to the room mic hardly decreases at low frequencies, which is obviously the result of having the same transducer response embedded in the input and the output spectra. When the singing voice was used for the excitation signal, the signal-to-noise ratio was very low outside the vocal range, and therefore, only averaging across the notes (Process 1) could not suppress the unlikely high response at low and high frequencies (see the lightest solid line in Fig. 2). Obviously, the frequency-domain bandpass filtering used in Process 2 and Process 3 could improve the response at these frequencies (the result of Process 2 not shown in the figure).

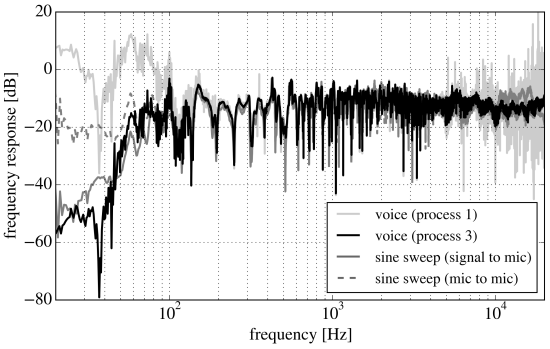


Fig. 2: Frequency responses compared between four cases.

From ~100 Hz up to ~5 kHz, it appears that the frequency responses did not depend much on the type of the excitation signal, the type of input (clean signal vs. the source-mic recording) or the post-processing method. If inspected with more care, however, it is noticed that the response with Process 3 agrees better with the reference measurement than Process 1 (thus Process 2 in this frequency range), where it is more stable with a lower variance, especially between ~100 Hz and ~400 Hz.

The findings described in the preceding text suggest that the human singing voice recorded near the singer's lips may be used as the excitation signal for the measurement of frequency/impulse response.

3 OBRIR measurement

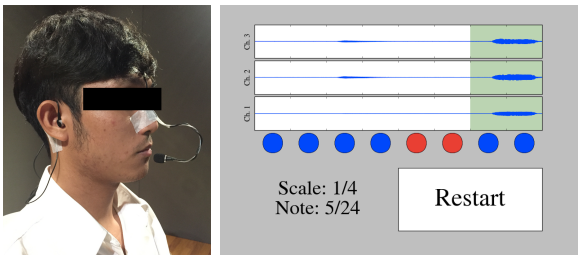
3.1 Methods

The same four volunteer singers were invited to the recording studio. In each recording session, a singer sat on a chair with backrest (without headrest) and wore two in-ear microphones (AT9905, Audio-Technica), and a lavalier microphone (AT9904, Audio-Technica) was also positioned just in front of ($< 1\,cm$) and at the center of the mouth opening when he/she sang 'ah' (see Fig. 3a). All microphones were powered by phantom-power converter (VXLR+, Rode) and connected to an audio interface (Fireface 802, RME).

A graphic user interface was created on Python, which generated the guide tone and also enabled the singer to control the progress of the recording (see Fig. 3b). Similar to the previous recording session as described in section 2.1, the singers sang the four chromatic scales (differing by 25 cents) one note after another within their vocal ranges. The participants were instructed to remember and keep the positions and shapes of their body parts, including jaw, tongue and lips as steady as possible, which was obviously intended to minimize the changes in the acoustic paths between the microphones and the reflective surfaces. All notes could be recorded typically within 10~15 minutes, and the singer could take break and restart at any time by using the graphical interface.

3.2 Results

Figure 4 shows an example of OBRFR and OBRIR estimated for a male singer (Male 2; left ear), where the results of the three post-processing methods are compared ($\alpha = 45\,$dB for Process 2 & Process 3). As was the case with the measurement using the pre-recorded voice played over loudspeaker, the averaged response (Process 1) does not decay at low and high frequencies due to the transducer response embedded both in the 'mouth-mic' and the 'ear-mic' signals [see panel (a)]. When a frequency-domain bandpass-filtering was applied (Process 2), the frequency response showed a more typical behavior, rolling off at low and high frequencies, which resulted in the impulse response with



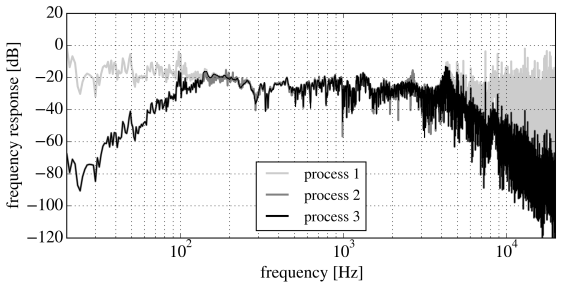
(a) (b)

Fig. 3: (a) Microphones worn by singer: Two at the entrance of the ear canal and another at the center of the mouth opening. (b) A graphical user interface which the singers used to listen to the guide tone and to control the progress.

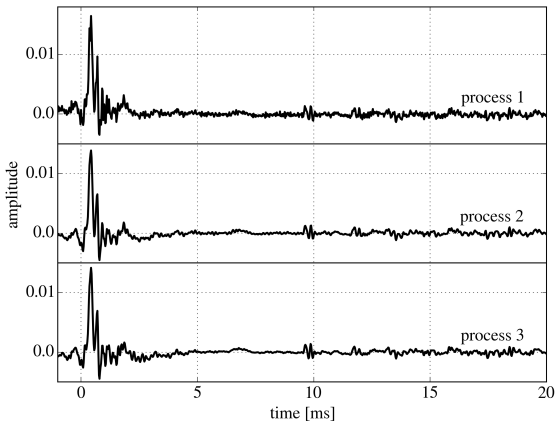
reduced high-frequency noise and a slightly lower DC offset compared to Process 1 [see panel (b)]. When the thresholding was additionally applied (Process 3), the magnitude of the frequency response seemed to be more stable than in Process 2, which is prominent from ~150 Hz to ~400 Hz and from ~2 kHz to ~4 kHz [see panel (a)]. Although clearly visible in the frequency domain, this improvement appears only to be subtle in the time-domain [see panel (b)].

Figure 5 shows the results of all four singers, where the responses (from the mouth to the left ear) were obtained by using Process 3. The fundamental frequencies of the lowest notes that Female 1 & 2 and Male 1 & 2 could sing were 168 Hz, 143 Hz, 84 Hz and 132 Hz, respectively, below which the signal-to-noise ratio was relatively low. In applying Process 3, therefore, the frequency responses were only averaged without thresholding below these frequencies (as in Process 2), and as a result, the variability of the OBRFRs appears to suddenly increase below these frequencies [see panel (a)]. Similarly, the OBRFRs tend to show higher variability above the fundamental frequency of the highest note each singer could sing, but the estimation may still be reasonable up to 4~5 kHz with a sufficient signal-to-noise ratio contributed by the harmonic components (not shown in the figure). Also shown in Fig. 5a is that OBRFR greatly differs between singers, which is obviously attributed to the differences in the head shape and in the acoustic paths from the mouth to the ears in the room.

Some common features can be observed in the OBRIRs (left ear) measured for all singers as shown in Fig. 5b.

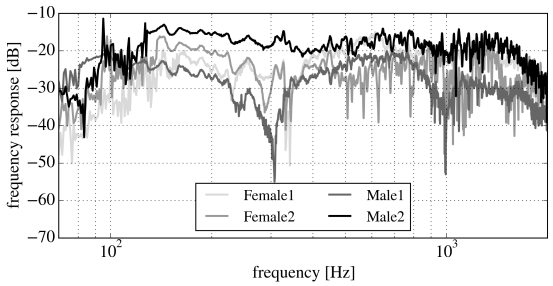


(a) Frequency responses

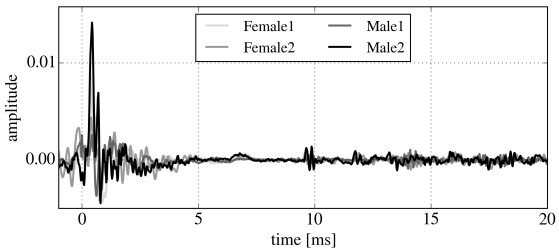


(b) Impulse responses

Fig. 4: OBRFRs and OBRIRs compared between three post-processing methods. (Left ear; Male 2)



(a) Frequency responses



(b) Impulse responses

Fig. 5: With Process 3 applied at $\alpha = 45\,\mathrm{dB}$, the OBRFRs and OBRIRs of all singers are compared (left ear only).

For example, the first peak of the response is positioned at 0.35~0.45 *ms*, equivalent to 12~15 *cm*, which seems to correspond well to the typical mouth-to-ear distance (no anthropometric measurement was made for the singers). Similarly, a rounded but prominent 'hill' is commonly found at ~7 *ms*, which appears to be the reflection from the floor somewhat occluded by the singer's body, and a stronger reflection from the ceiling may be associated with the peak at ~9 *ms*.

Given the results presented in the preceding text, it is suggested that OBRIR can be measured on human head by using the singing voice as the excitation signal. Seated on a chair with backrest, the movement of the singer's body parts could be limited to an extent so that the estimated OBRIRs show some features consistent across all singers. Process 3 appears to provide the best estimate of OBRIR, although the subtle time-domain differences observed between the three post-processing methods have yet to be investigated in terms of perceivable effects, for example, when the OBRIRs are used for auralization.

4 Measurement protocol for practicality

The OBRIR measurement described in section 3 typically took 10~15 min, which may be too long a period for singers to keep their posture restrained. Therefore, a shorter measurement procedure was conceived and tested for practicality and the consistency of the results.

4.1 Methods

Female 1 and the author (Male 3) volunteered in this part of the measurement, where the same measurement configuration was used as described in section 3.1. Instead of singing one note at a time, however, the participants sang eight notes sequentially from *Do* to the next *Do* in diatonic scale within a time interval of 1.2 *s* (in 4 beats at 100 beats/minute) following a 4.8-second guide sound reproduced by the graphical user interface. The vocal ranges of these singers were two octaves (Male 3) or slightly less (Female 1), and

therefore, by singing two scales ('low scale' & 'high scale') that differed in the pitch of the first *Do* by an octave or less, the whole vocal range could be covered. Once completed, the pitches of the first *Do*s in the low and high scales were raised by 25 cents at a time, and the measurement was repeated 8 times so that the reference *Do* may vary in one full tone (200 cents). In other words, the participants sang 16 scales, 8 pairs of low and high scales, where each pair differed in pitch by one-eight of full tone. Then, these 16 scales were sung three more times, resulting in a total of four sets of 16 recordings.

4.2 Results

OBRIR was estimated from each set of 16 recordings, where Process 3 was applied with $\alpha = 45\,\mathrm{dB}$. The results are shown in Fig. 6 for Male 3, in which the impulse response does not appear to vary much between the four sets of recordings, suggesting that OBRIR may be measured in a short time by using the proposed method with repeatability assured. Similar results were obtained from the recordings by Female 1 (data not shown).

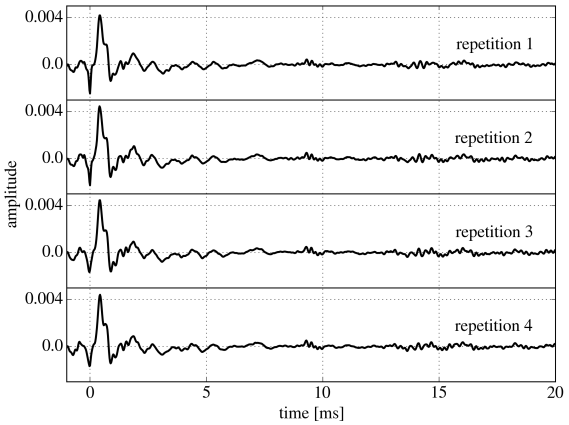


Fig. 6: The results of the OBRIR measurement repeated four times (left ear; Male 3). Each OBRIR was estimated from the recording of 16 scales.

Having found that recording a diatonic scale at once can shorten the time needed for the measurement, a further analysis was carried out to see if a smaller number of scales (less than 16) may be sufficient to estimate OBRIR. From the 16 scales recorded by Male 3 in the first repetition, two subsets of 8 scales and 4 scales were selected, of which the first *Do* differed by 50 and 100 cents, respectively. When estimating OBRIR only with 4 scales, the number of valid frequency bins was insufficient when $\alpha = 45\,\mathrm{dB}$, and therefore, a lower value, $\alpha = 35\,\mathrm{dB}$ was used. In Fig. 7, the OBRIRs estimated from the two subsets are compared to that from the 16 scales, where differences are hard to identify between the responses estimated from the 16 scales and from the 8 scales. Despite some additional but subtle 'jittering' around the first peak of the response, the OBRIR estimated only from the 4 scales appears to be similar to the former two. The results suggest that OBRIRs may reasonably be estimated by singing only 4 or 8 scales which can be completed in 1~2 minutes, although the parameter α might have to be adjusted.

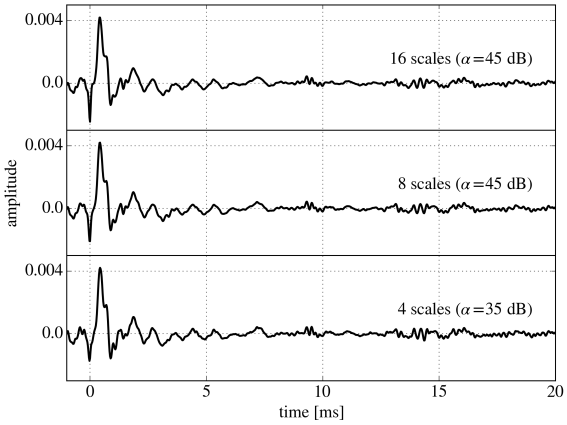


Fig. 7: The OBRIRs estimated from the recordings of 16, 8 and 4 scales (left ear; Male 3).

As a matter of fact, the value of α had to be adjusted manually in the current study: 20 dB in section 2, 45 dB in section 3 and 35 or 45 dB in section 4, depending on the signal-to-noise ratio of the source-mic or mouth-mic recordings. In the case of singing voice, the estimated OBRIR may be more prone to noise, unless the voice is sufficiently strong. When OBRIRs are to be used for the auralization of a concert venue or to investigate the perceived acoustic scene on stage, it is likely that such application will be made for professional singers with voice of sufficient volume, and therefore, the signal-to-noise ratio may not be an issue. Nevertheless, a more systematic method has yet to be established to find the optimal value of α for individual singers.

5 Summary

In the current study, a step-by-step approach was taken to investigate the possibility of measuring oral-binaural room impulse responses (OBRIRs) on human head by using the person's singing voice. First, it was shown that the recording taken at the proximity of a sound source (singer's mouth) could be used for the estimation of the frequency response. Three post-processing methods were compared, and it was found that averaging raw frequency responses only in valid frequency bins (where the input spectrum is above a predetermined threshold spectrum) may reduce the variability of the response in the frequency domain, thus resulting in the impulse response with least noise. With this post-processing scheme applied, OBRIRs were measured on volunteers, who sang one note after another in a chromatic scale or a diatonic scale at a time. The results showed that OBRIR may be estimated in 1~2 minutes by singing 4~8 scales within the singer's vocal range, and the response may be consistent when repeated.

Acknowledgment

This work was supported by King Mongkut's Institute of Technology Ladkrabang Research Fund [KREF186111]. The author thanks Kittitorn Himasuk, Kris Wannawong, Veerapat Pongyart and Watsaya Takkapaijit, who, within their final-year project, collected the data used in section 2. The author also thanks the four undergraduate students who volunteered to sing for the measurement.

References

[1] Møller, H., "Fundamentals of binaural technology," *Applied acoustics*, 36(3-4), pp. 171–218, 1992.

[2] Skirlis, K., Cabrera, D., and Connolly, A., "Spectral and temporal changes in singer performance with variation in vocal effort," *Proceedings of Acoustics 2005*, 2005.

[3] Kato, K., Ueno, K., and Kawai, K., "Effect of room acoustics on musicians' performance. part II: audio analysis of the variations in performed sound signals," *Acta Acustica united with Acustica*, 101(4), pp. 743–759, 2015.

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

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

[6] Yadav, M. and Cabrera, D., "Autophonic loudness of singers in simulated room acoustic environments," *Journal of Voice*, 31(3), pp. 388.e13–388.e25, 2017.

[7] Miranda Jofre, L. A., Cabrera, D., Yadav, M., Sygulska, A., and Martens, W., "Evaluation of stage acoustics preference for a singer using oral-binaural room impulse responses," *Proceedings of Meetings on Acoustics ICA 2013*, 19(1), 2013.

[8] Cabrera, D., Yadav, M., Miranda, L., Collins, R., and Martens, W. L., "The sound of one's own voice in auditoria and other rooms," in *Proceedings of International Symposium on Room Acoustics*, 2013.

[9] Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L., "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, 94(1), pp. 111–123, 1993.