รายงานการวิจัยฉบับสมบูรณ์

ระบบห้องฝึกซ้อมเสมือนจริงสำหรับนักดนตรี

Virtual rehearsal suite for musicians

ผศ.ดร. มูนฮีม บัค

Asst. Prof. Munhum Park

ได้รับการสนับสนุนทุนวิจัย กองทุนวิจัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (KREF186111)

This work is supported by
King Mongkut's Institute of Technology Ladkrabang (KREF186111)

รายงานการวิจัยฉบับสมบูรณ์

ระบบห้องฝึกซ้อมเสมือนจริงสำหรับนักดนตรี

Virtual rehearsal suite for musicians

ผศ.ดร. มูนฮีม บัค

Asst. Prof. Munhum Park

ได้รับการสนับสนุนทุนวิจัย กองทุนวิจัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (KREF186111)

This work is supported by
King Mongkut's Institute of Technology Ladkrabang (KREF186111)

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.

Acknowledgement

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.

Index

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 timeconsuming 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.

1.1 Objectives

1.1.1 Main objective: In this research project, advanced sound reproduction technology will be studied, improved and employed to create a "virtual rehearsal suite" with which musicians can experience practicing in the music hall of their choice, listening to the sound of their own performance and to the sound of the accompanying musicians (if available).

1.1.2 Secondary objectives: Given the system of virtual rehearsal suite developed and tested objectively, a formal subjective test with musicians will be carried out to evaluate the system for the similarity to the real auditory environment. In the sister project (to be submitted for the similar funding opportunity to KMITL), the influence of acoustic properties on the musicians' performance style is to be investigated, where visual sensation will also be presented.

1.2 Literature reviews

It is well known that the recordings made at human's ear drums on both sides contain all the information that our brain requires to reconstruct the three dimensional auditory scene, and therefore, carefully replicating (binaural reproduction) or synthesizing (binaural synthesis) these recordings (signals) can create virtual auditory environment (see, e.g., (1,2)). Since the sounds arriving at our ears are modified not only by the surrounding (room) but also by the listener's body (mostly the shape of the ears and head), the binaural room impulse response (BRIR) represents the complete transfer functions from a sound source to the receiver position in the room, and if known (measured) and convolved with a clean source signal, virtual acoustic images in the same room can be created and presented over headphones or loudspeakers (3,4). For concert hall acoustics, BRIR is usually measured between a representative source location on the stage and a selected receiver position in the audience. In order to evaluate the stage acoustics (the quality of the sound as heard from the musicians' perspective), however, both source and the receive must be on the same stage, and especially for singers, the source-receiver distance is very short, only ~10cm (from mouth to ears). A specific method to measure the so-called oral binaural room impulse response (OBRIR) has been developed relatively recently (5), which paved the way to investigate musicians' perception of the stage acoustics or to synthesize the binaural sounds that represent the sounds of their own on the stage (6). Virtual acoustic imaging systems for musicians were implemented mostly by using loudspeaker arrays in an anechoic chamber. For example, Brereton et al. (7) created a loudspeaker-based real-time room acoustics simulation system for musicians and evaluated the system in comparison to off-line auralization techniques. Kato et al. (8) also used a virtual environment for musicians where the influence of room acoustics on the musicians' performance styles was investigated. Since the systems described in the preceding text requires loudspeakers configured in an anechoic chamber, it is almost impossible for musicians to actually use the system for their daily practice and training. In the current project, we aim at developing a headphone-based system that can present a virtual auditory environment for musicians in an ordinary (e.g. practice) room, for which additional signal processing modules are essential to develop for the equalization of the room characteristics (2).

1.3 Scope of Thesis

With the conventional system for virtual acoustic environment, the user can only passively receive the sound signals over loudspeakers or headphones. For the implementation of the virtual rehearsal suite as described above, however, the acoustic scene must be synthesized in real time, also incorporating the musicians' own sound (either singing or playing instrument), which may be the first challenge. The second challenge may be to implement the system NOT in an anechoic chamber (where reflections do not exist, therefore acoustically very well controlled), but in an ordinary practice room (where the response of the room must be equalized). Overcoming these technical challenges, we aim at building a lab prototype of the virtual rehearsal suite, which consists of headphones and microphones connected to high-performance computer (for processing the audio signals in real time). Although the visual information may influence the musicians' perception (also affecting the auditory perception), only the auditory environment will be realized for musicians, where adding the visual display to the virtual rehearsal suite is the topic of the sister project (to be submitted for the similar funding opportunity to KMITL).

2

Theory

2.1 Conceptual Framework

2.1.1 Introduction

The main theoretical framework used for the system is "binaural hearing/reproduction," which states that humans perceive three-dimensional auditory scenes only by the sounds heard at two ears, and therefore, a virtual sound scene can be created, if these signals are exactly replicated. Another framework used for the system is the principles in room acoustics and stage acoustics. The sound generated at the source (musician) arrives at the receiver position (audience), where the sound is obviously modified by the acoustical characteristics of the room (concert hall). Binaural room impulse responses (the response of the room to the source sound as measured at the listener's two ears) can capture all the transfer characteristics from the source to the receiver, with which any virtual acoustic scene may be synthesized on computer. Stage acoustics is a sub-topic of room acoustics where the receiver (listener) is not any more the audience, but the musicians themselves on the stage, where the perceived quality (including intelligibility) of the sound generated by the colleague musicians on the stage or by the musician him/herself are the most important aspects. Past findings in these research fields will be the bases of the virtual rehearsal suite to be developed in the current project.

3

Methodology

3.1 Singing voice for excitation signal

3.1.1 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 each note of four chromatic scales within the range, of which the pitch differed by 25 cents from each other. ('ah' was chosen since it is the phoneme that can be pronounced and sung easily, if not most easily.) 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).

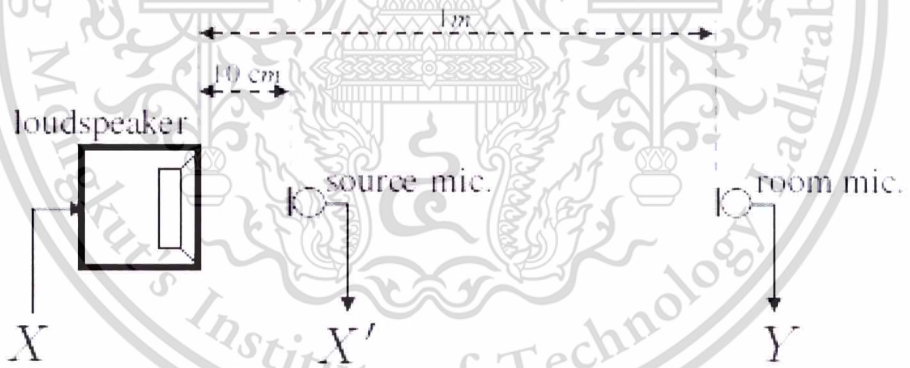


Fig. 1: Configuration for the impulse response measurement.

For the impulse response measurement, two microphones (M30, Earthworks) and a single-driver loudspeaker were placed in one of the empty rooms in the recording studio (_ 4:5m_5:5m_2:5m) as shown in Fig. 1. 'Source mic' and 'room mic' were positioned at the distance of 10cm and 1m from the loudspeaker, respectively, where 1m was considered to be reasonably sufficiently close to the sound source. First, a sine sweep from 20 Hz to 20 kHz (with spectrum X) was played back over the speaker, and the sounds recorded at the

source- and room-mics were Fourier-transformed (labeled as X, X0 and Y in Fig. 1), and the two frequency responses H =Y=X and H0 =Y=X0 were estimated, which served as the reference responses. Then, the voice recorded for each note was played back one

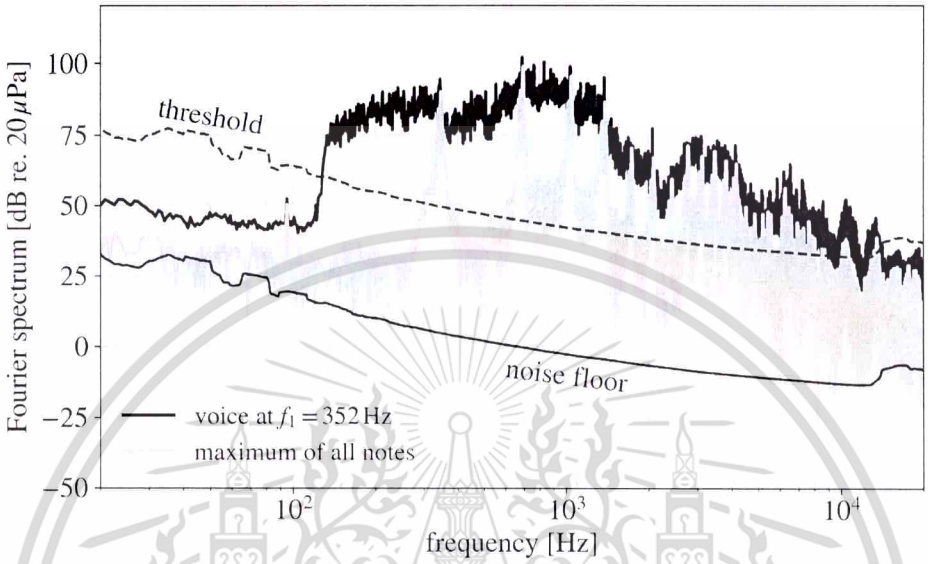


Fig. 2: An example spectrum of a singer's voice is shown in grey with the maximum obtained across all notes indicated in black. The threshold spectrum is determined at a dB above the noise floor.

by one over the speaker, and, similar to the sine-sweep case, H0 = Y=X0 was estimated for each note. These 'raw' responses were further refined by using one of the following three methods:_ Process 1: The frequency responses were only averaged across all notes._ Process 2: The maximum magnitude of the sourcemic spectrum was calculated at each frequency bin across all notes, which was then compared to a threshold spectrum set at a dB above the spectrum of the noise floor (see Fig. 2). (The noise floor was estimated from the quite intervals of the voice recording at the source mic.) From this comparison, a usable frequency range with an appropriate level of signal-to-noise ratio (> a dB) was determined.

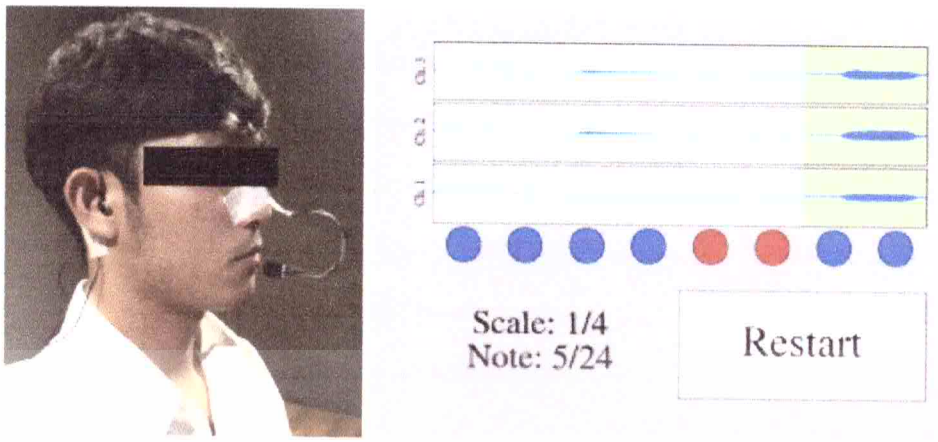
The average frequency response (averaged across all notes) was considered to be valid only within the usable frequency range, and that out of this range was attenuated by applying a frequencydomain bandpass filter, of which the magnitude response resembled that of the fourth-order Butterworth filter. _ Process 3: The threshold spectrum described in Process 2 was used to identify usable frequency bins (rather than range) in the source-mic

spectrum for each note, which usually corresponded to those close to the harmonic frequencies of the note (see Fig. 2). 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. All post-processings and analyses described in the preceding text were performed on Python.

3.2 OBRIR measurement

3.2.1 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. 4(a)]. All microphones were powered by phantompower 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. 4(b)]. 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.



(a) (b)

Fig. 4 : (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.

3.3 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.

3.3.1 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 1.2s (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 Dos 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

Results & Discussions

4.1 Singing voice for excitation signal

4.1.1 Results

Figure 3 shows four frequency responses measured by using a loudspeaker and one or two microphones, where the value of a 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. 3). 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). 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.

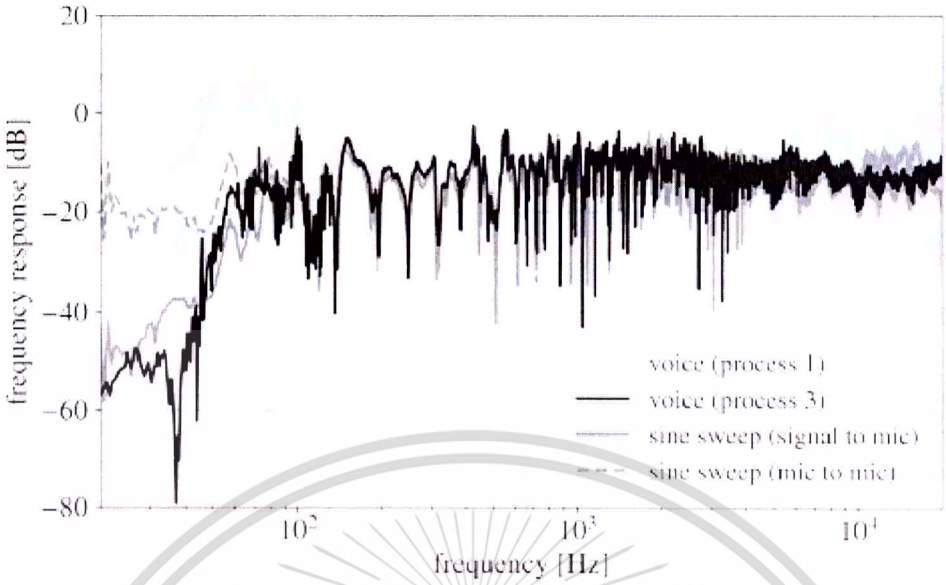


Fig. 3 : Frequency responses compared between four cases.

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.

4.2 OBRIR measurement

4.2.1 Results

Figure 5 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 (a = 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 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 6 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 signalto-noise ratio contributed by the harmonic components (not shown in the figure).

Also shown in Fig. 6(a) 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. 6(b). 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 ~7ms, 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.

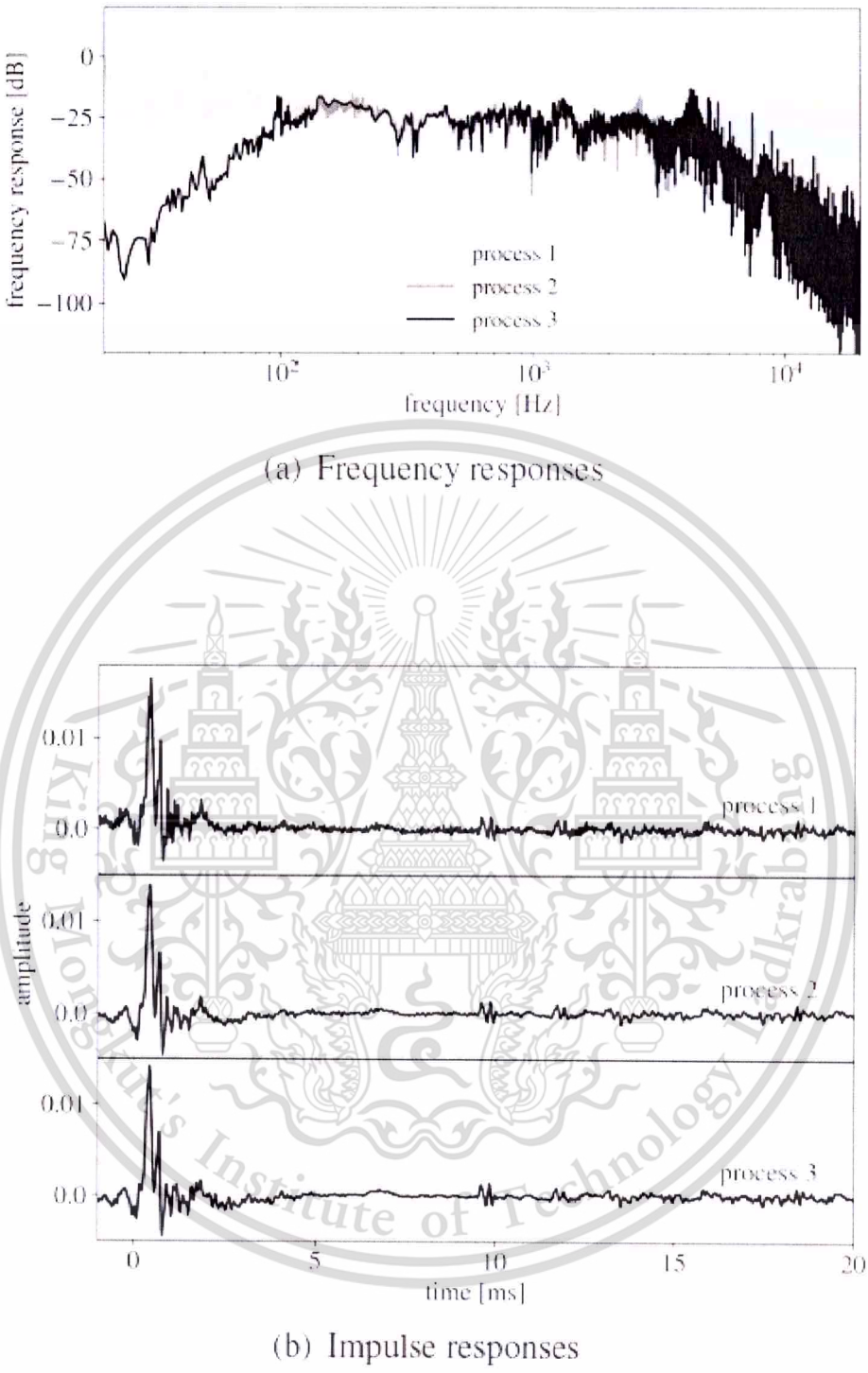


Fig. 5 : OBRFRs and OBRIRs compared between three post-processing methods. (Left ear; Male 2)

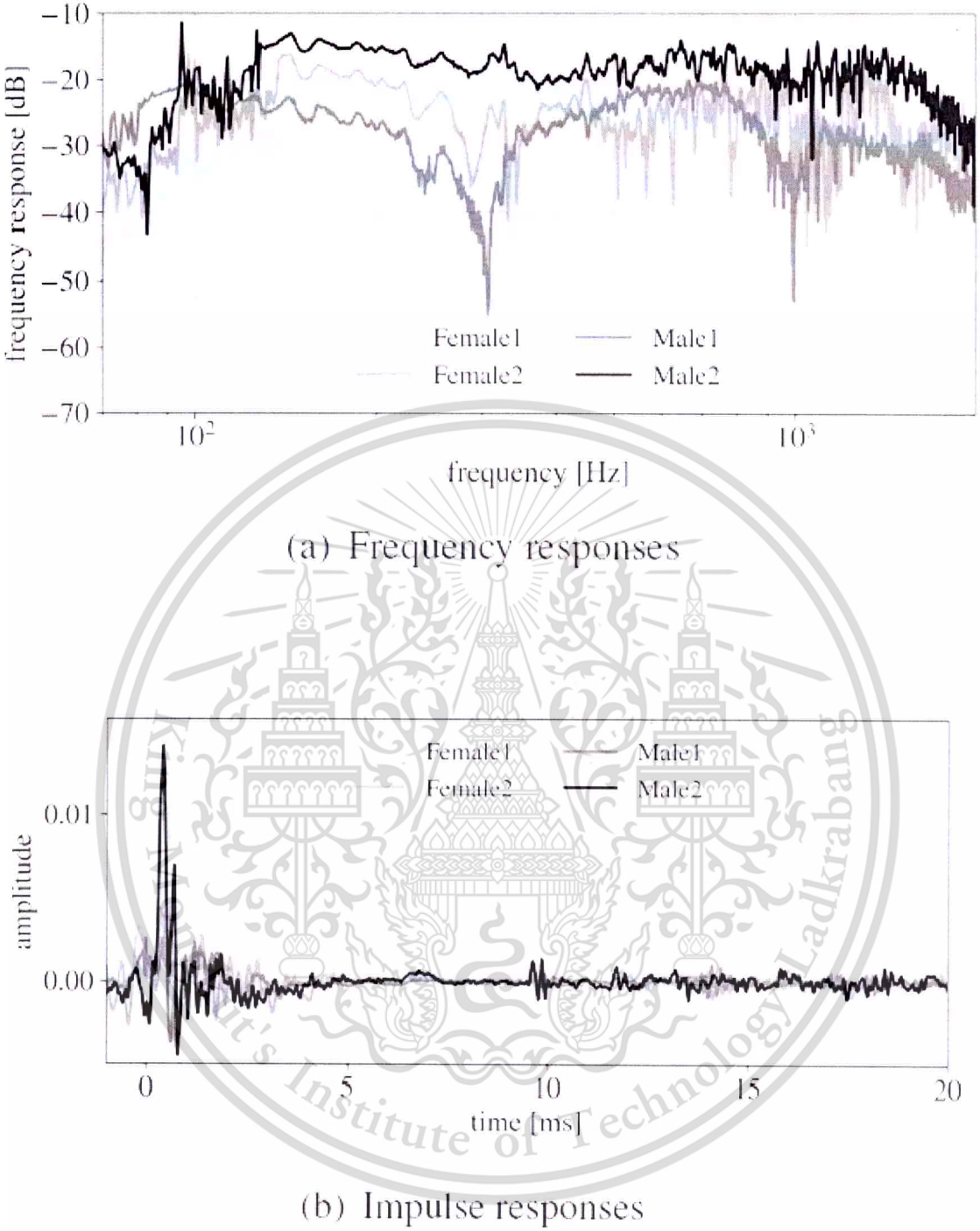


Fig. 6: With Process 3 applied at a=45 dB, the OBRFRs and OBRIRs of all singers are compared (left ear only).

4.3 Measurement protocol for practicality

4.3.1 Results

OBRIR was estimated from each set of 16 recordings, where Process 3 was applied with a=45 dB. The results are shown in Fig. 7 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). 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 with a=45 dB, and therefore, a lower value, a=35 dB was used. In Fig. 8, 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 a might have to be adjusted.

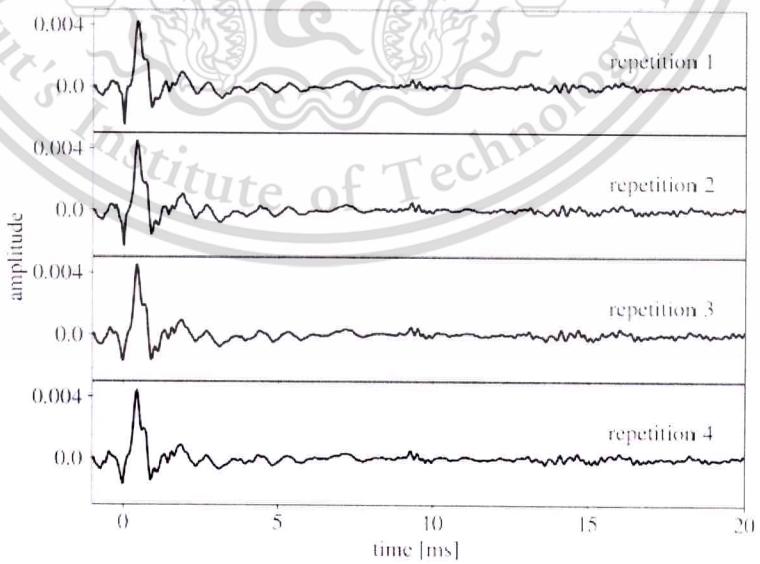


Fig. 7: The results of the OBRIR measurement repeated four times (left ear; Male 3). Each OBRIR was estimated from the recording of 16 scales.

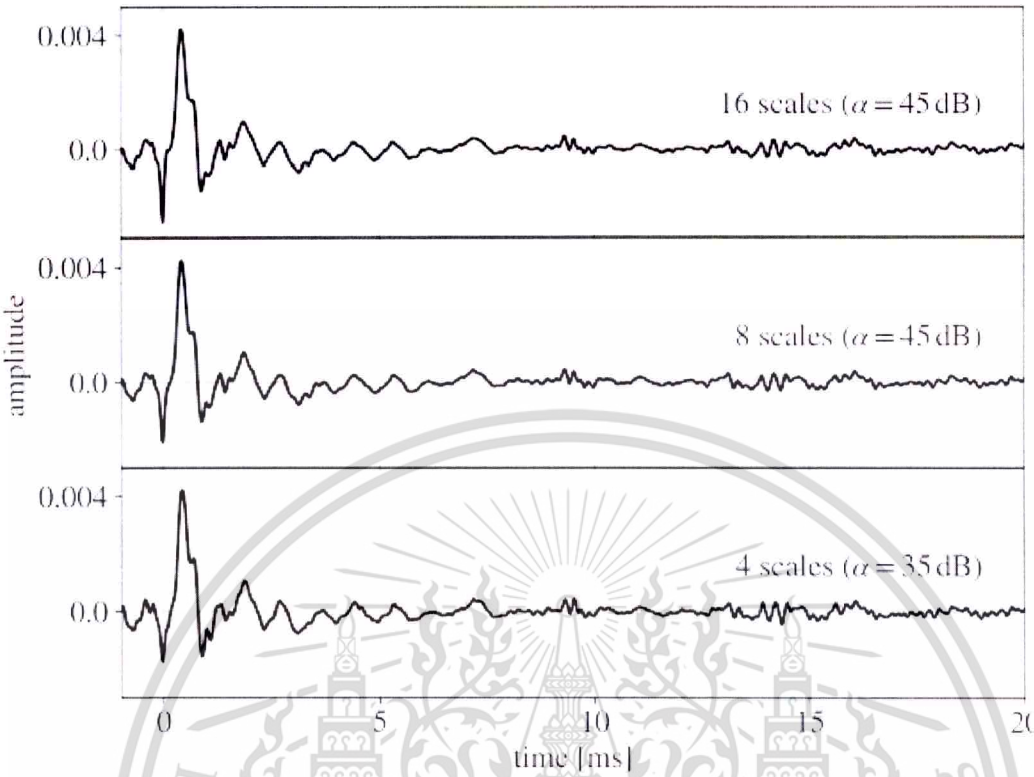


Fig. 8: The OBRIRs estimated from the recordings of 16, 8 and 4 scales (left ear; Male 3).

As a matter of fact, the value of a 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 mouthmic 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 a for individual singers.

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 nusing 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 postprocessing 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. A further investigation is to be carried out to use the estimated OBRIRs for the purpose of auralization in rooms of various size.

Reference

[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," in *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 oralbinaural room impulse responses," in *Proceedings of Meetings on Acoustics* ICA2013, ASA, 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 *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.

Appendix

Researcher's Curriculum Vitae

ประวัติส่วนตัว

ชื่อ-สกุล	**ผศ.ดร. มูนฮีม บัค (Asst. Prof. Munhum Park)**
เพศ	ชาย
วันเดือนปีเกิด	15 ธันวาคม พ.ศ. 2517
อายุ	44 ปี
ตำแหน่ง	Lecturer
เริ่มทำงาน ณ สจล.	ตั้งแต่เดือน กันยายน พ.ศ. 2559

ประวัติการศึกษา

ชื่อปริญญา

BSc Department of Physics Seoul National University, South Korea 2543

MSc Institute of Sound and Vibration University of Southampton, UK 2547

MPhil/PhD Institute of Sound and Vibration University of Southampton, UK 2550

ประสบการณ์วิจัยหรือสาขาที่ชำนาญ

Spatial audio, audio signal processing, subjective evaluation of audio systems and soundscape, computational models of human and animal hearing, effects of noise on work performance

ผลงานวิจัย/งานสร้างสรรค์ที่ตีพิมพ์เผยแพร่ (ระดับชาติและระดับนานาชาติ)

• M. Park, P. Vos, B. Vlaskamp, A. Kohlrausch, and A. W. Oldenbeuving, "The influence of PACHE II score on the average noise level in an intensive care unit: an observational study," BMC Anesthesiology, vol. 15, no. 1, p. 42, 2015.

• K. S. Simons, M. Park, A. Kohlrausch, M. van den Boogaard, P. Pickkers, W. de Bruijn, and C. de Jager, "Noise pollution in the ICU: time to look into the mirror," Critical Care, vol. 18, no. 4, p. 493, 2014.

• M. Park, A. Kohlrausch, W. de Bruijn, P. de Jager, and K. Simons, "Analysis of the soundscape in an intensive care unit based on the annotation of an audio recording," The Journal of the Acoustical Society of America, vol. 135, no. 4, pp. 1875–1886, 2014.

• M. Park, A. Kohlrausch, and A. van Leest, "Irrelevant speech effect under stationary and adaptive masking conditions," The Journal of the Acoustical Society of America, vol. 134, no. 3, pp. 1970–1981, Sep. 2013.

• M. Park, "Comment on 'Short-term acoustic forecasting via artificial neural networks for neonatal intensive care units' [J. Acoust. Soc. Am. 132, 3234–3239 (2012)]," The Journal of the Acoustical Society of America, vol. 134, no. 1, pp. 5–8, 2013.

• M. Park and R. Allen, "Pattern-matching analysis of fine echo delays by the spectrogram correlation and transformation receiver," The Journal of the Acoustical Society of America, vol. 128, no. 3, pp. 1490-1500, 2010.

• M. Park, P. A. Nelson, and K. Kang, "A model of sound localisation applied to the evaluation of systems for stereophony," Acta Acustica united with Acustica, vol. 94, no. 6, pp. 825–839, Nov. 2008.

• M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," The Journal of the Acoustical Society of America, vol. 118, no. 5, pp. 3094–3103, 2005.

การเสนอผลงานวิชาการ

• M. Park, R. Vermeulen, S. Laroche, and M. Gillies, "A preliminary analysis of the sources of noise in an open-plan neonatal intensive care unit," in Proceedings of INTER-NOISE and NOISE-CON Congress and Conference, 2072–2079, 2017

• P. Chuprasert, N. Phupantachsee, and M. Park, "Analysis of warning sounds used in public transportation system," in Proceedings of INTER-NOISE and NOISE-CON Congress and Conference, 2520–2526, 2017

• M. Park, R. Vermeulen, S. Laroche and D. van Minde, "A preliminary analysis of the soundscape in an open-plan neonatal intensive care units," in Proceedings of WESPAC 2015, Singapore, 2015.

• T. U. Senan, R. Navarro, M. Park, S. Jelfs, and A. Kohlrausch, "Spectral and temporal features as estimators of the irrelevant speech effect," in Proceedings of EuroNoise2015, Maastricht, 2015.

• A. Kohlrausch, M. Park, W. de Bruijn, P. de Jager, and K. Simons, "Neue Ergebnisse zur Messung und Analyse der Schallfelder in Intensivstationen," L.rmbek.mpfung, vol. 10, no. 2, 2015.

• M. Park, P. Vos, A. Kohlrausch, and A. W. Oldenbeuving, "Trends of the acoustic condition in an intensive care unit based on a long-term measurement," The Journal of the Acoustical Society of America, vol. 135, no. 4, p. 2403, 2014.

• M. Park, P. Vos, A. Kohlrausch, B. Vlaskamp, and A. W. Oldenbeuving, "Analysis of the acoustic environment in an ICU using patient information as a covariate," Critical Care, vol. 18, no. Suppl 1, P16, 2014.

• A. H.rm., A. Kohlrausch, and M. Park, "Predicting the subjective evaluation of spatial audio systems," in Proceedings of International Conference on Spatial Audio, Erlangen, 2014.

• A. H.rm., A. Kohlrausch, and M. Park, "Data-driven modeling of the spatial sound experience," in Proceedings of the 136th Convention of the Audio Engineering Society, Berlin, 2014.

• K. Simons, M. Park, W. de Bruijn, M. van den Boogaard, A. Kohlrausch, and P. de Jager, "A comparative analysis of acoustic conditions in an old and a new ICU room," in Proceedings of the 26th Annual Congress of the European Society of Intensive Care Medicine, Paris, 2013.

• M. Park, A. Kohlrausch, W. de Bruijn, P. de Jager, and K. Simons, "Source-specific analysis of the noise in an intensive care unit," in Proceedings of the 42nd International Congress and Exposition on Noise Control Engineering, Innsbruck, 2013.

• A. H.rm., R. van Dinther, T. Svedstr.m, M. Park, and J. Koppens, "Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface," in Proceedings of the 132nd Audio Engineering Society Convention, Budapest, 2012.

• P. A. Nelson, M. Shin, F. M. Fazi, T. Takeuchi, M. Park, J. Seo, and K. Kang, "Systems for virtual sound imaging," in Proceedings of the 5th International Universal Communication Symposium, Gumi, Republic of Korea, 2011.

• A. H.rm. and M. Park, "Extraction of voice from the center of the stereo image," in Proceedings of the 130th Audio Engineering Society Convention, London, 2011.

• M. Park, A. H.rm., S. van de Par, and G. Tryfou, "Comparison of the width of sound sources in 2-channel and 3-channel sound reproduction," in Proceedings of the 128th Audio Engineering Society Convention, London, 2010.

• M. Park, P. A. Nelson, F. Fazi, and K. Kang, "Application of an auditory process model for the evaluation of stereophonic images," in Proceedings of the Institute of Acoustics, Reading, vol. 30, 2008.

• M. Park, P. Nelson, and K. O. Kang, "Evaluation of stereophonic images with listening tests and model simulations," in Proceedings of the 124th Audio Engineering Society Convention, Amsterdam, 2008.

• P. A. Nelson, M. Park, T. Takeuchi, and F. Fazi, "Binaural hearing and systems for sound reproduction," in Proceedings of Acoustics 08, Paris, pp. 3531–3536, 2008.

• M. Park, P. A. Nelson, and Y. Kim, "An auditory process model for the evaluation of virtual acoustic imaging systems," in Proceedings of the Institute of Acoustics, Southampton, vol. 28, 2006.

• M. Park, P. A. Nelson, and Y. Kim, "An auditory process model for the evaluation of virtual acoustic imaging systems," in Proceedings of the 120th Audio Engineering Society Convention, Paris, 2006.

• M. Park, P. A. Nelson, and Y. Kim, "An auditory process model for sound localization," in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, pp. 122–125, 2005.

• B. Rafaely and M. Park, "Plane-wave decomposition by spherical-convolution microphone array," The Journal of the Acoustical Society of America, vol. 115, p. 2578, 2004.

• B. Rafaely and M. Park, "Super-resolution spherical microphone arrays," in Proceedings of the 23rd IEEE Convention of Electrical and Electronics Engineers, Israel, p. 424, 2004.

ผลงานสิทธิบัตร/สิ่งประดิษฐ์/งานสร้างสรรค์ (ศิลปะหรืออื่นๆ)

• S. de Waele, M. Park, A. Kohlrausch, and A. den Brinker, "Method and device for effective audible alarm settings," US20170367663 A1, 2015.

• K. Leuschner, M. Park, A. Barroso, and F. Mueller, "Monitoring the exposure of a patient to an environmental factor," US20170364648 A1, 2015.

• T. Falck, M. Park, A. Kohlrausch, and R. van Dinther, "Apparatus and method for alarm detection and validation," US20160283681 A1, 2014.

• M. Park, A. Kohlrausch, and A. van Leest, "Directional sound masking," WO2014016723 A3, 2014.

• A. Kohlrausch, M. Park, S. Martin Jelfs, and T. Falck, "Apparatus and method for improving the audibility of specific sounds to a user," WO2014140053 A1, 2014.

• A. Kohlrausch, T. Falck, M. Park, S. Martin Jelfs, and K. Leuschner, "Systems and methods for reducing the impact of alarm sounds on patients," WO2014136069 A2, 2014.

• J. Bernd Roland, A. H.rm., and M. Park, "Audio rendering system and method therefor," WO2013111034 A3, 2014.

• M. Park, A. Kohlrausch, T. Falck, and A. den Brinker, "A medical monitoring system based on sound analysis in a medical environment," WO2013057608 A1, 2013.

• M. Park, A. Kohlrausch, A. den Brinker, and T. Falck, "A medical feedback system based on sound analysis in a medical environment," WO2013057652 A2, 2013.

• A. H.rm., M. Park, and G. Tryfou, "An audio system and method therefor,"

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 each note of four chromatic scales within the range, of which the pitch differed by 25 cents from each other. ('ah' was chosen since it is the phoneme that can be pronounced and sung easily, if not most easily.) 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).

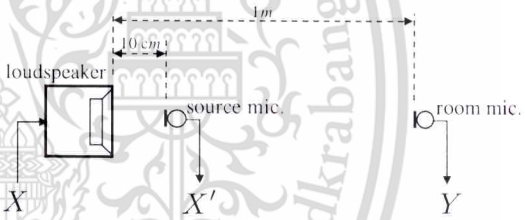


Fig. 1: Configuration for the impulse response measurement.

For the impulse response measurement, two microphones (M30, Earthworks) and a single-driver loudspeaker were placed in one of the empty rooms in the recording studio ($\sim 4.5\,m \times 5.5\,m \times 2.5\,m$) as shown in Fig. 1. 'Source mic' and 'room mic' were positioned at the distance of $10\,cm$ and $1\,m$ from the loudspeaker, respectively, where $1\,m$ was considered to be reasonably far from the loudspeaker given the room size, and $10\,cm$ sufficiently close to the sound source. First, a sine sweep from 20 Hz to 20 kHz (with spectrum X) was played back over the speaker, and the sounds recorded at the source- and room-mics were Fourier-transformed (labeled as X, X' and Y in Fig. 1), and the two frequency responses $H = Y/X$ and $H' = Y/X'$ were estimated, which served as the reference responses. Then, the voice recorded for each note was played back one

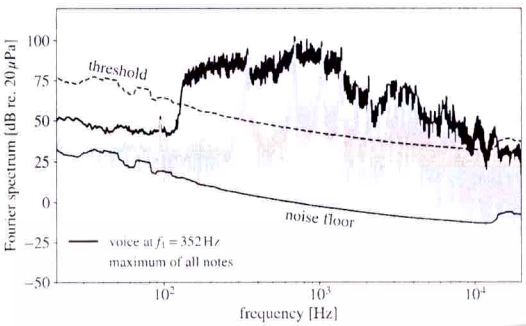


Fig. 2: An example spectrum of a singer's voice is shown in grey with the maximum obtained across all notes indicated in black. The threshold spectrum is determined at α dB above the noise floor.

by one over the speaker, and, similar to the sine-sweep case, $H' = Y/X'$ was estimated for each note. These 'raw' responses were further refined by using one of the following three methods:

- 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. 2). (The noise floor was estimated from the quite intervals of the voice recording at the source mic.) 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 usable frequency bins (rather than range) in the source-mic spectrum for each note, which usually corresponded to those close to the harmonic frequencies of the note (see Fig. 2). 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. All post-processings and analyses described in the preceding text were performed on Python.

2.2 Results

Figure 3 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. 3). 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).

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.

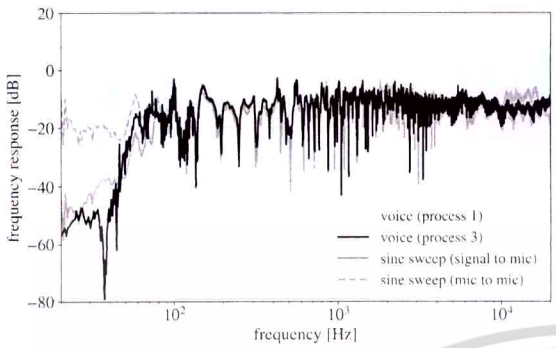


Fig. 3: Frequency responses compared between four cases.

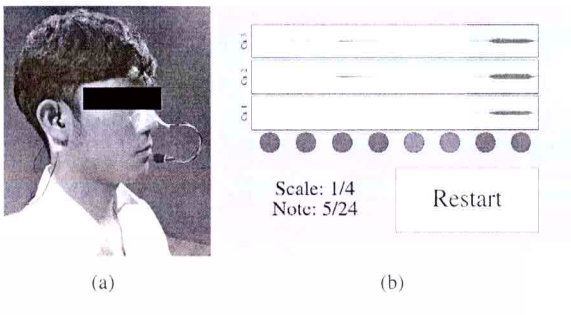


Fig. 4: (a) Microphones worn by singer: Two at the entrance of the ear canal and another at the center of the mouth opening. (b) A graphical user interface which the singers used to listen to the guide tone and to control the progress.

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

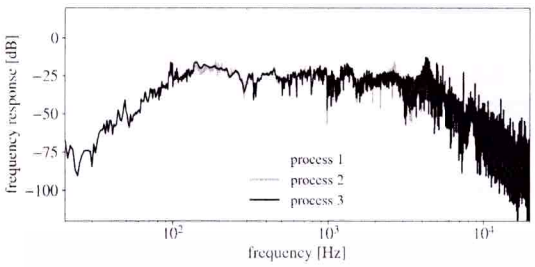
The same four volunteer singers were invited to the recording studio. In each recording session, a singer sat on a chair with backrest (without headrest) and wore two in-ear microphones (AT9905, Audio-Technica), and a lavalier microphone (AT9904, Audio-Technica) was also positioned just in front of ($< 1\ cm$) and at the center of the mouth opening when he/she sang 'ah' [see Fig. 4(a)]. 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. 4(b)]. 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.

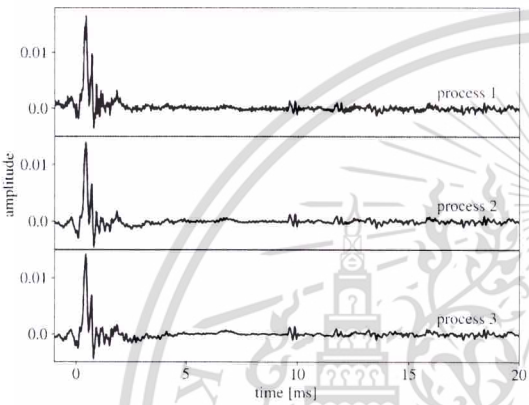
3.2 Results

Figure 5 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 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 6 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

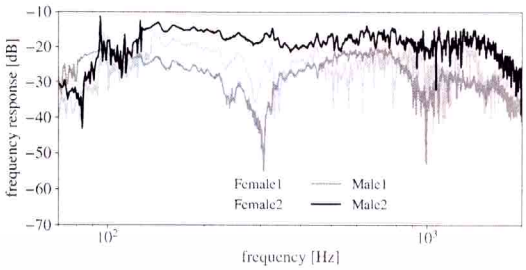


(a) Frequency responses

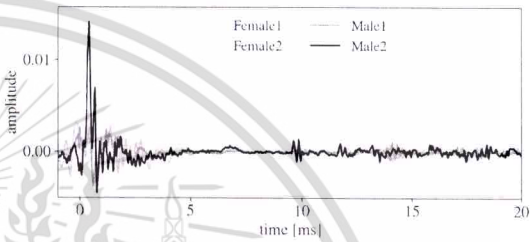


(b) Impulse responses

Fig. 5: OBRFRs and OBRIRs compared between three post-processing methods. (Left ear; Male 2)



(a) Frequency responses



(b) Impulse responses

Fig. 6: With Process 3 applied at α=45 dB, the OBRFRs and OBRIRs of all singers are compared (left ear only).

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. 6(a) 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. 6(b). 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 ~7ms, 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 1.2s (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 Dos 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 α=45 dB. The results are shown in Fig. 7 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).

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 with α=45 dB, and therefore, a lower value, α=35 dB was used. In Fig. 8, 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

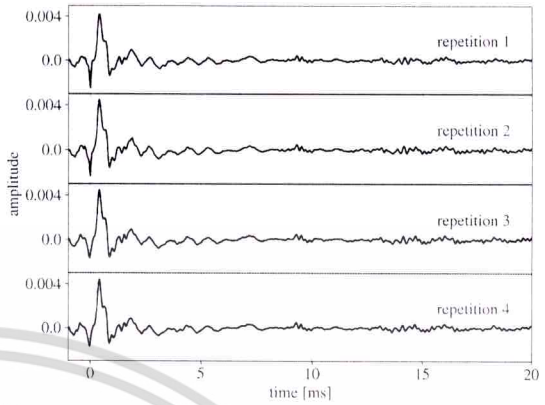


Fig. 7: The results of the OBRIR measurement repeated four times (left ear; Male 3). Each OBRIR was estimated from the recording of 16 scales.

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.

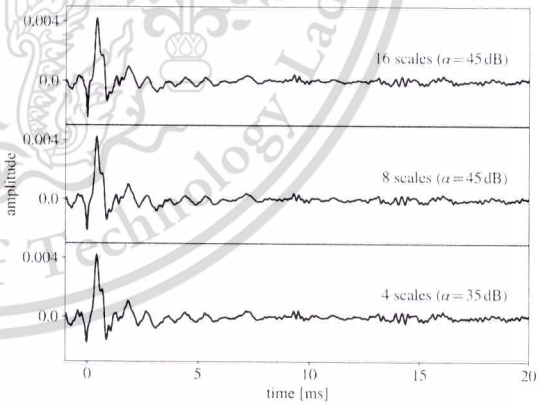


Fig. 8: The OBRIRs estimated from the recordings of 16, 8 and 4 scales (left ear; Male 3).

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. A further investigation is to be carried out to use the estimated OBRIRs for the purpose of auralization in rooms of various size.

6 Acknowledgement

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," in *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," in *Proceedings of Meetings on Acoustics ICA2013*, ASA, 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 *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.

KMITL
พระจอมเกล้าลาดกระบัง

AES 147: notification for paper 61

AES 147 <aes147@easychair.org> Tue, Jul 16, 2019 at 5:13 AM
To: Munhum Park <munhum.pa@kmitl.ac.th>

Dear Munhum Park,

It is our pleasure to inform you that your proposed paper 61 entitled "Measurement of oral-binaural room impulse response by singing scales" has been accepted for presentation the AES 147th Convention in New York, 2019 October 16 - 19. We are appending the feedback provided by the reviewers below and expect that you will include the comments and corrections suggested before you upload your final paper.

Mode of presentation: Lecture.
You will receive further notification regarding your final session assignment, date and time once the schedule has been finalized.

Please make note of the following important points

1) Papers (electronic manuscripts) must be submitted for both lecture and poster presentations to the AES 147th paper-collection site (https://easychair.org/conferences/?conf=aes147papers) by the 2019 July 26 deadline. Please consider the comments of the reviewers (see below) when you prepare your revised manuscript. Failure to submit a complete electronic manuscript by this date will force us to withdraw your paper and presentation from the convention.

2) The paper-submission site will be reopened soon. You may visit the submission site to make changes as many times as you wish through the July 26 deadline.Questions? Please contact Bill McQuaide at wtm@aes.org (+1 212 661 8528, ext. 22).

3) If you make changes on your final manuscript to your title, abstract, or coauthors please be sure to update this information on the paper-collection site. The presenting author should be listed on the site as the corresponding author, as we will be sending information later about registration and the presentation schedule.

4) Lecture presentations will be assigned at 30-minute intervals: session chair intro, 20 minutes for the lecture, and a 5-minute question-and-answer period. Digital projectors for PowerPoint presentations will be available in each lecture presentation room at the convention. We will send lecture presenters a facilities form on which you can request additional audio/visual equipment. You should also periodically check for updates about the convention on the AES website at: www.aes.org/events/147/

5) As announced in the Call for Papers, the presenting authors for each paper in categories 1 through 3 will be required to pay a Presenting Author Fee at the following applicable rate:

- AES members (Life member* / Fellow / Member / Associate categories): $470
- AES student members: $130
- Non AES members: $595

These amounts are below advance registration fees for the Convention and are required payment for all presenting authors in categories 1 through 3. No additional registration fee will be required; all Presenting Authors will receive a free Four-day All Access Badge on payment of their Presenting Author Fee.

Payment must be made by September 6th, 2019.

Presenting Authors SHOULD NOT REGISTER for the Convention on the main registration website, your registration will be done automatically for you on payment of your Presenting Author Fee. A presenting author invoice will be sent as a PDF by email to the presenting author indicated on the submission site, if there is any change in presenting author name, please inform Bill McQuaide at: bill.mcquaide@aes.org Authors presenting more than one paper in categories 1 through 3 are only required to pay one Presenting Author Fee and will only

receive one invoice. Co-authors of a paper are welcome, and encouraged, to attend the presentations but will require an All Access Badge (available for four, two or one days) to access the sessions.

*Note: While Life Members have the privilege of no-cost All Access registration, Life Members who are Presenting Authors in categories 1 through 3 are required to pay the Presenting Author Fee.

6) Printed copies of convention papers will not be provided or sold at the convention.

If you have any questions regarding your paper, please contact Areti Andreopoulou and Braxton Boren, the 147th papers co-chairs (147th_papers@aes.org).

Regards,
Areti Andreopoulou & Braxton Boren
147th AES Convention papers co-chairs

Reviewer Comments:

SUBMISSION: 61
TITLE: Measurement of oral-binaural room impulse response by singing scales

---------- REVIEW 1 ----------
SUBMISSION: 61
TITLE: Measurement of oral-binaural room impulse response by singing scales
AUTHORS: Munhum Park

---------- Scientific quality ----------
SCORE: 3 (fair)
---------- Overall evaluation ----------
SCORE: 2 (accept)
---------- Review ----------
An interesting study, in an area definitely related with the main topics of the conference.

In general, the novelty of the approach is clear, and the scientific quality of the study is ok, but certain sections of the paper need a thorough revision.

-I am not aware of studies looking at the individual features of OBRIRs - this is definitely something interesting to explore further.
-Some of the choices the author made are not properly backed, and often relevant details are not given. Was there a reason for choosing 'ah' rather than other sounds for the voice recordings? Why were the two microphones placed at 10cm and 1m from the loudspeaker (i.e. why these two specific values)? "in one of the empty rooms in the recording studio" - what are the details of the room, e.g. large or small, reverberation time, etc?
-First paragraph Page 3 - the explanation of the various impulse responses that have been measured/calculated is rather confused. Also the whole explanation regarding the use of the loudspeaker with reference and room mics is not particularly clear. I would strongly recommend revising Section 2 in its structure and language - most of the information is already there, but it is not presented in a clear and more "standard" manner.
-Sections 3 and 4 are definitely clearer.
-The comparison between the three "post-processing" methods is interesting, but it would have been good to establish a proper metric to quantify the differences rather than just looking at diagrams and graphically compare between them. Similarly for the comparisons in Section 4.

Finally, I am surprised not to see any mention of bone-conducted sound that could be picked up by the in-ear microphones. I don't know the extent of the phenomenon, but it is definitely worth mentioning it and the extent in which this could alter the recordings done using the proposed technique.

I am happy to recommend this paper to be accepted for publication, but i strongly encourage the author to revise the text according to my comments above, in particular Section 2.

---------- REVIEW 2 ----------
SUBMISSION: 61
TITLE: Measurement of oral-binaural room impulse response by singing scales
AUTHORS: Munhum Park

---------- Scientific quality ----------
SCORE: 4 (good)
---------- Overall evaluation ----------
SCORE: 2 (accept)
---------- Review ----------
The abstract suggests that this is an interesting paper concerned with the measurement of OBRIRs. They have attempted to overcome some issues from difficulties encountered by previous researchers in making measurements. They suggest that the results they obtained means that their technique is reliable and repeatable for measuring OBRIRs, making it useful.

Going through the full paper in detail it is reasonably well put together. The structure is pretty clear through the paper and there is only one typographic error where the they say 'on python' on pages 2 and 4 which should be 'using python'. I think the diagrams are appropriate throughout the the paper though it is hard to discern the various signals in the spectral plots. The most difficult aspect of the paper for me was the explanation of the procedures for Processes 1 to 3. Ideally, though it might be some work, a diagram for each of these could be included. The OBRIR measurements read well and the final version of the algorithm appears to perform according to the plots. However, it would be useful to add some future work to the final section to at least set out what could be done next to improve the results and to extend the work.

Scopus

Search Sources Alerts Lists Help SciVal Register Login

3,922 document results

View secondary documents View 8342 Mendeley Data

SRCTITLE (audio AND engineering AND society AND convention)

Edit Save Set alert Set feed

Search within results...

Analyze search results

Show all abstracts Sort on: Date (newest)

All

Refine results

Access type

Other (3,922)

Year

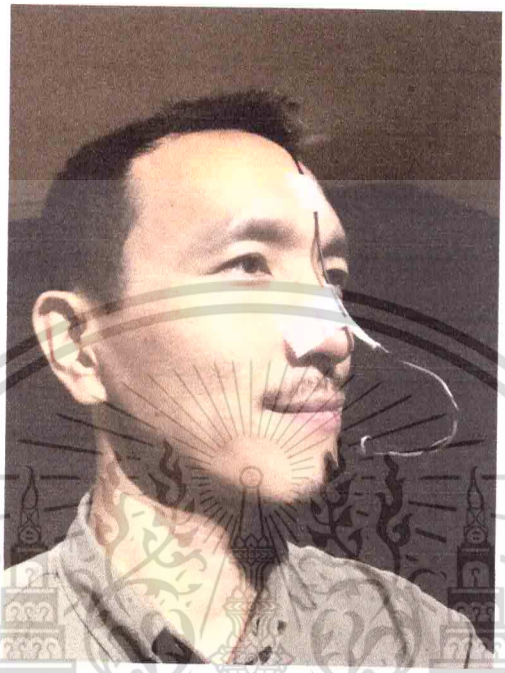
2018 (265)
2017 (265)
2016 (197)
2015 (265)
2014 (202)

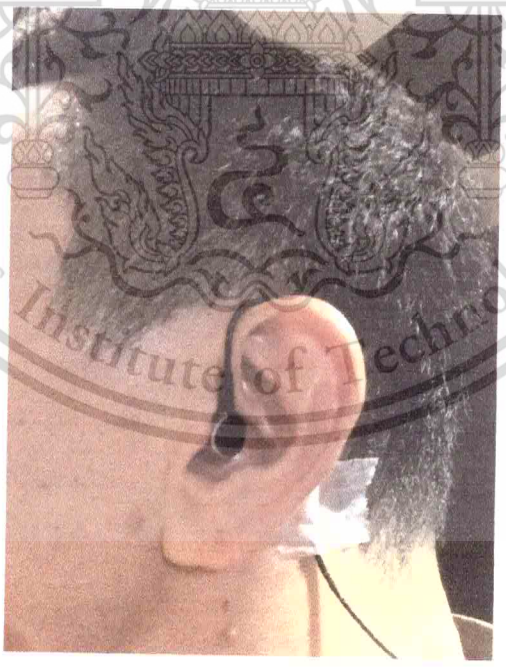
View more

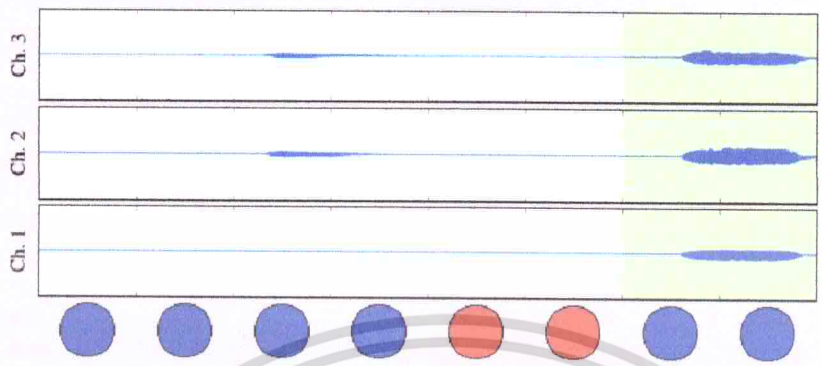
Author name

Faller, C. (43)
Pulkki, V. (43)

	Document title	Authors	Year	Source	Cited by
1	Evaluation of binaural renderers: Externalization, front/back and up/down confusions	Reardon, G., Zalles, G., Genovese, A., Flanagan, P., Roginska, A.	2018	144th Audio Engineering Society Convention 2018	2
	View abstract Related documents				
2	Characteristics of vertical sound image with two parametric loudspeakers	Aoki, S., Shimizu, K., Itou, K.	2018	144th Audio Engineering Society Convention 2018	0
	View abstract Related documents				
3	An allpass chirp for constant signal-to-noise ratio impulse response measurement	Canfield-Dafilou, E.K., Abel, J.S.	2018	144th Audio Engineering Society Convention	1

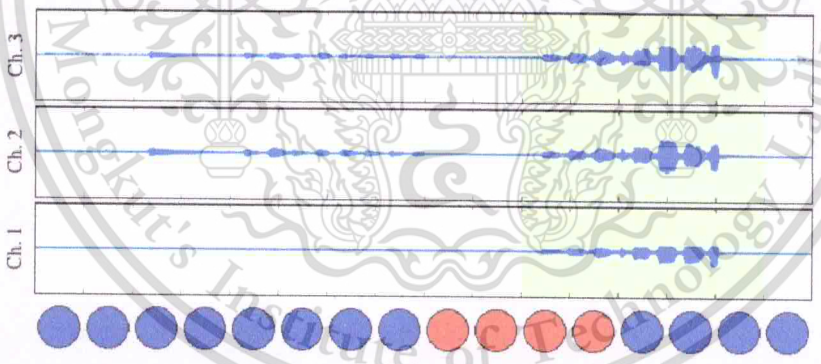






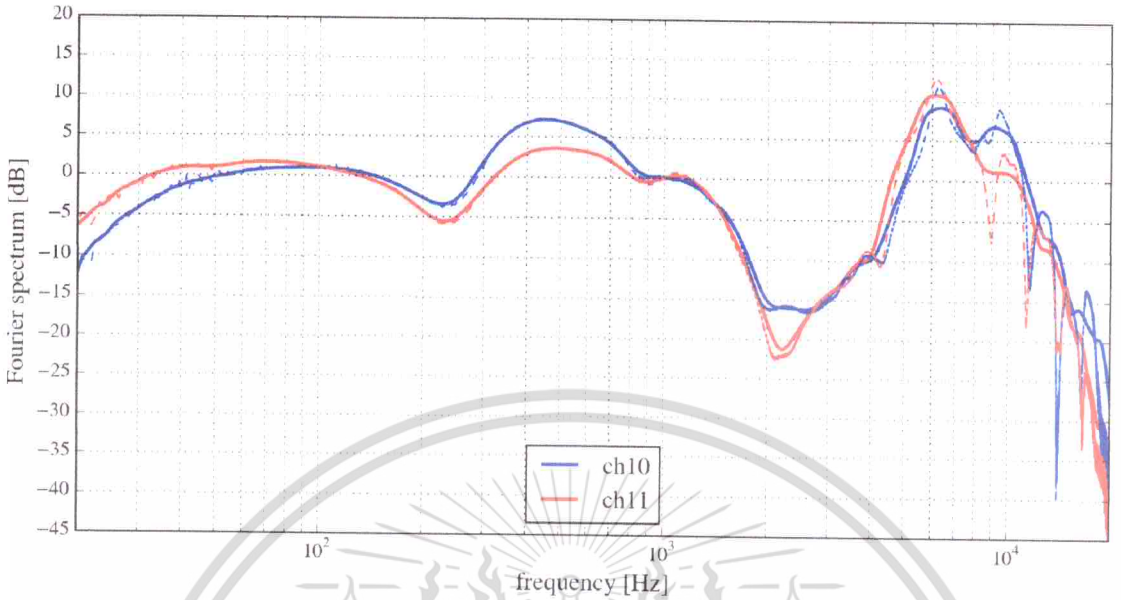
Scale: 1/4
Note: 5/24

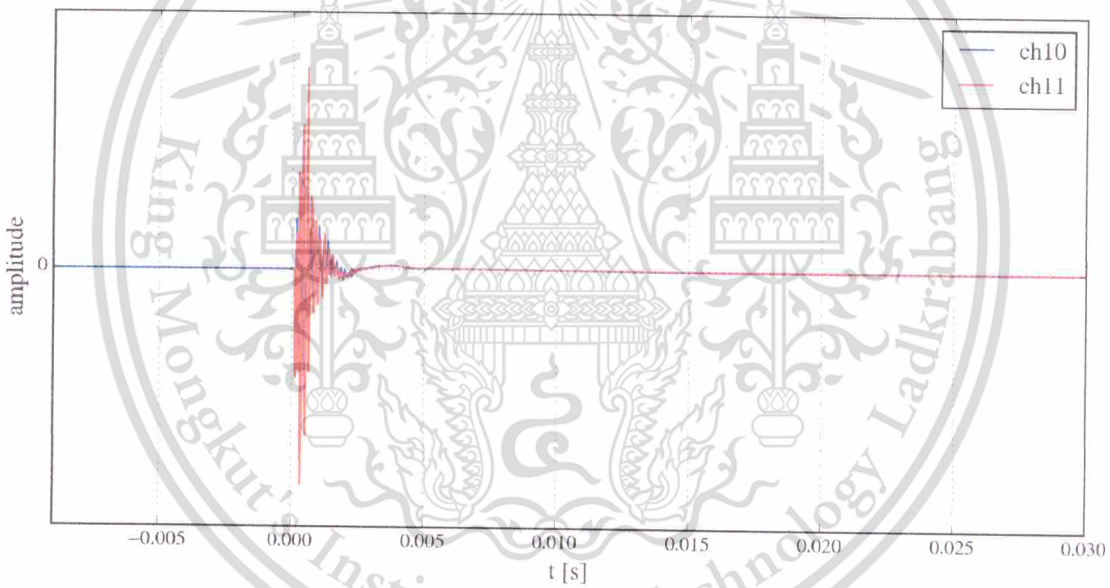
Restart

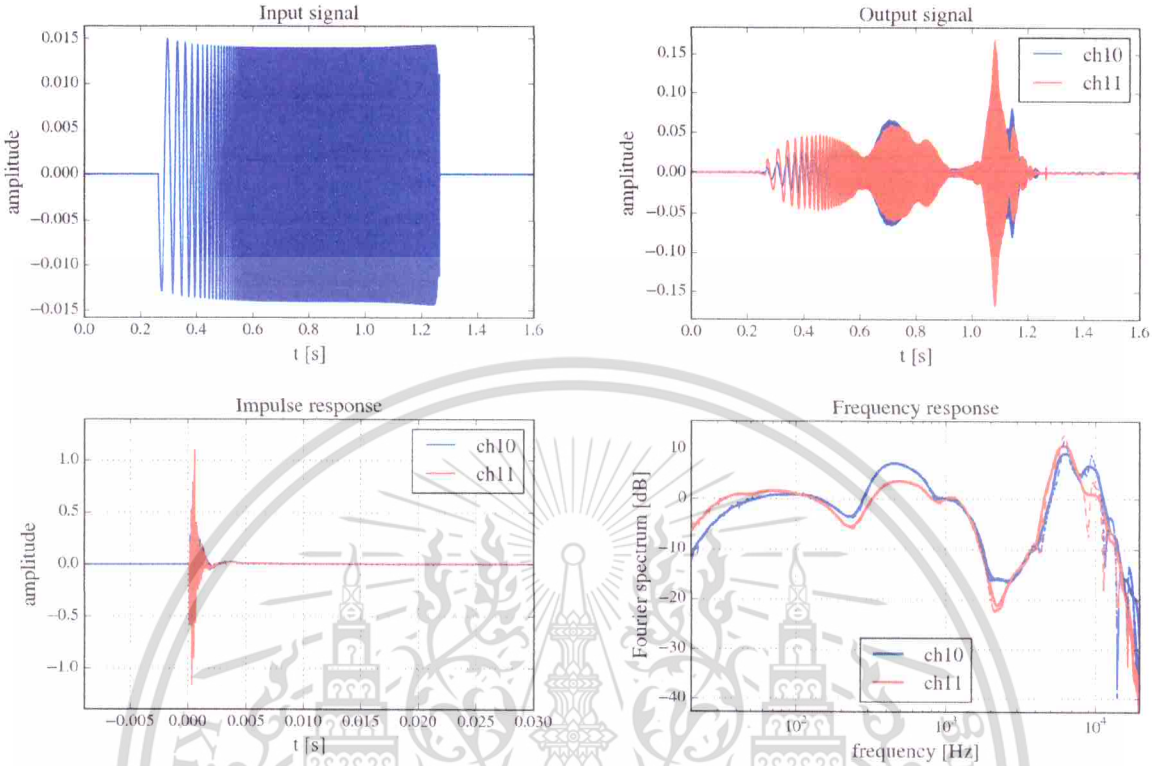


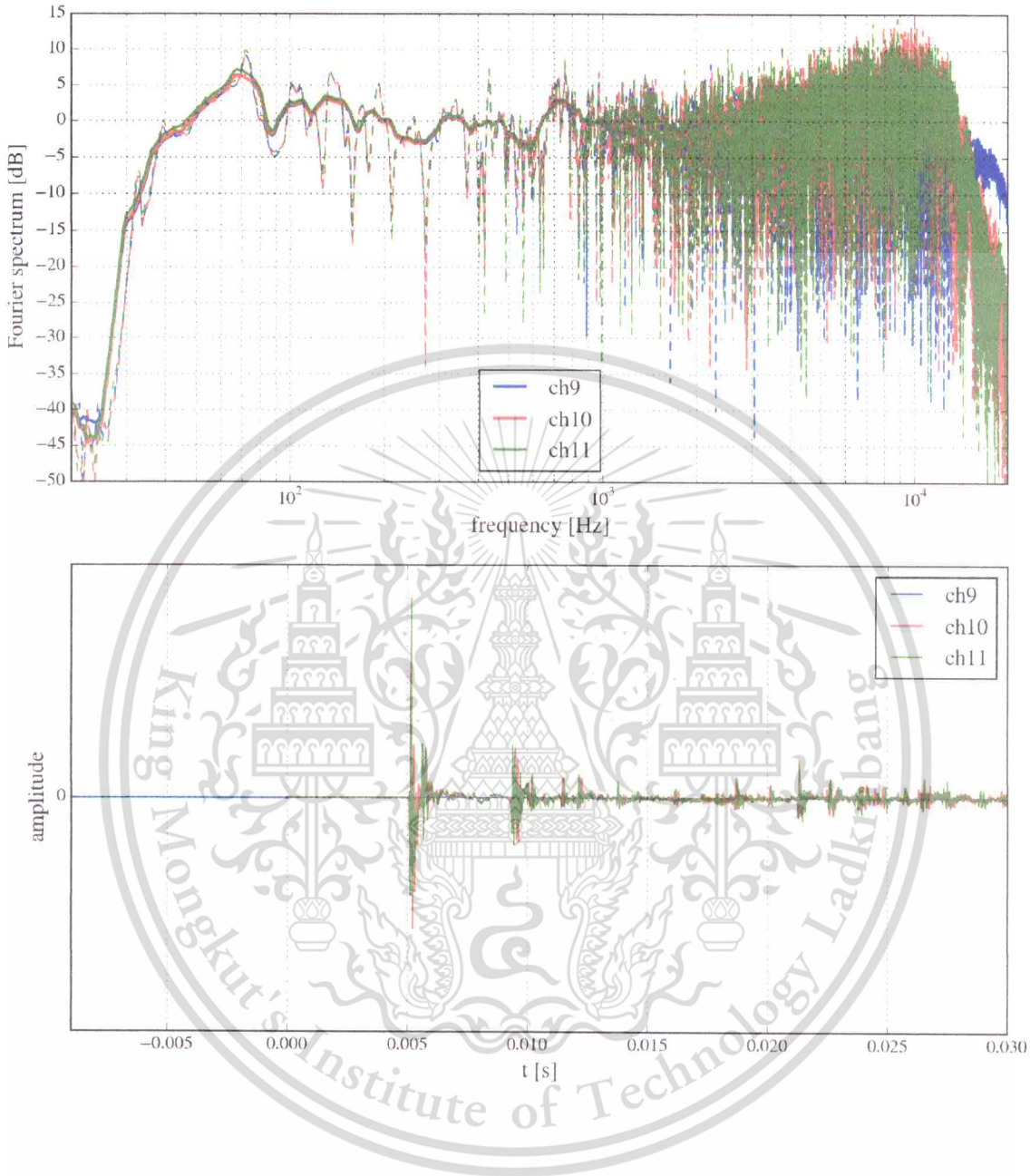
Scale: 3/16
Repetitions: 1/4

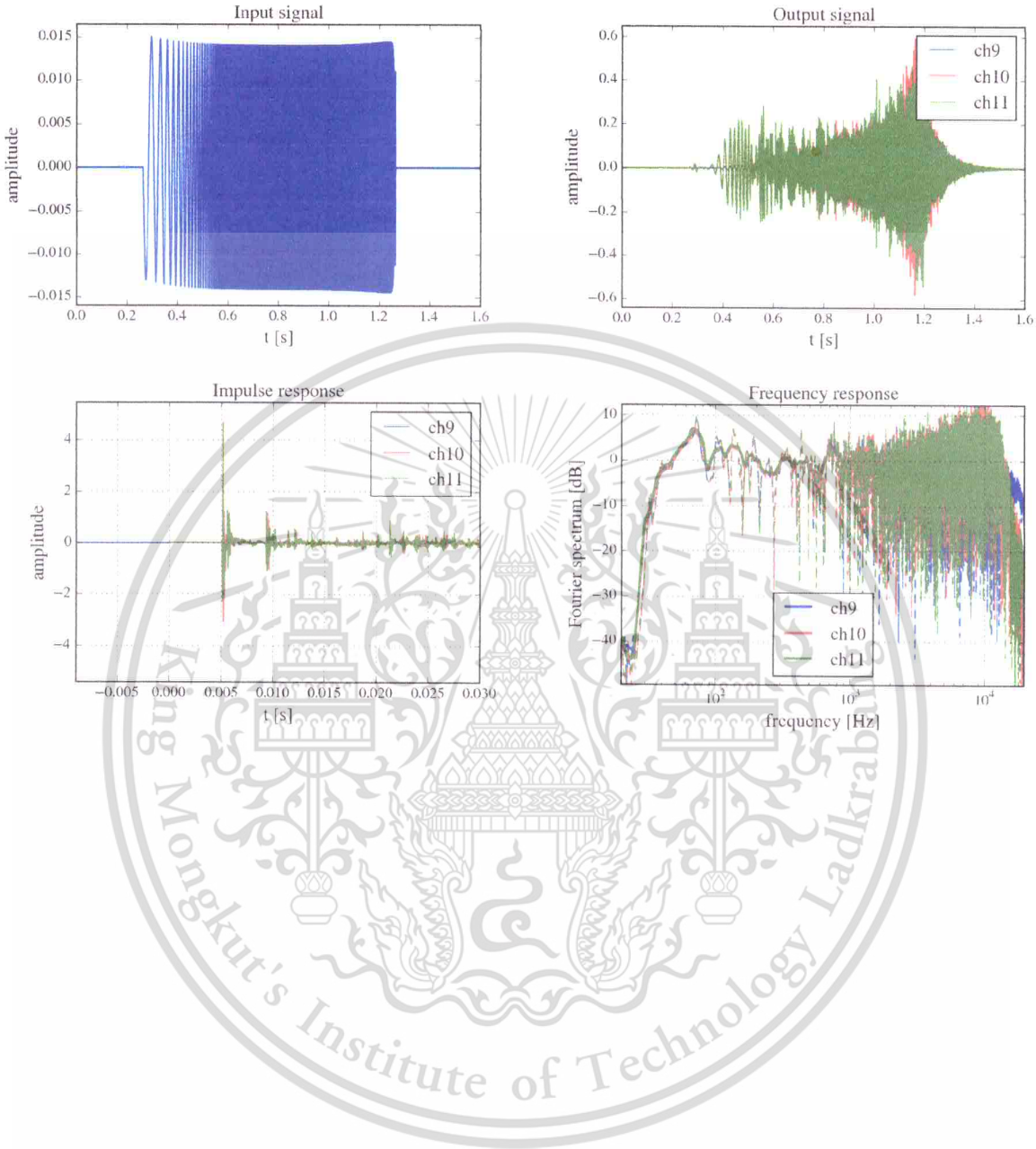
Restart

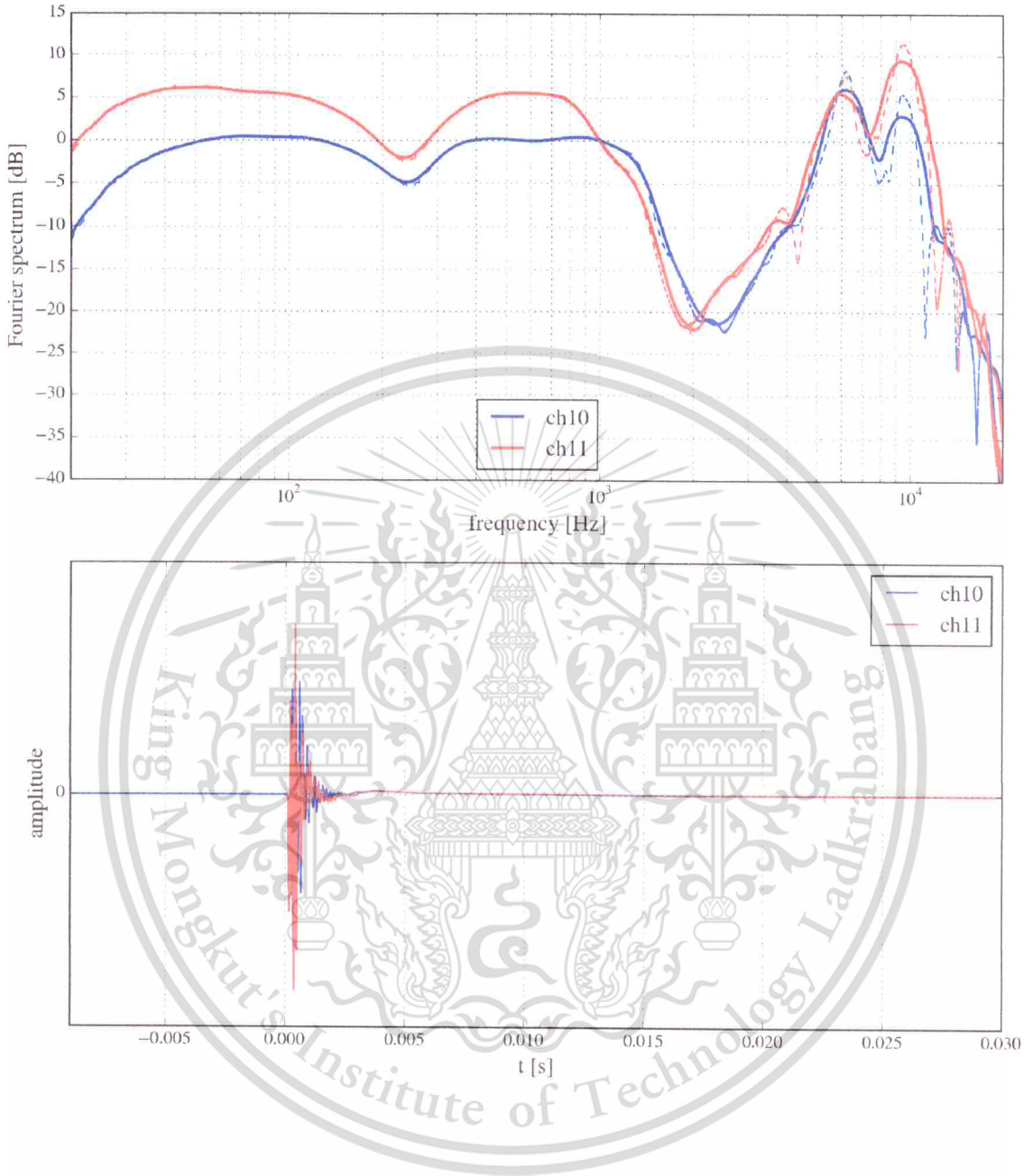


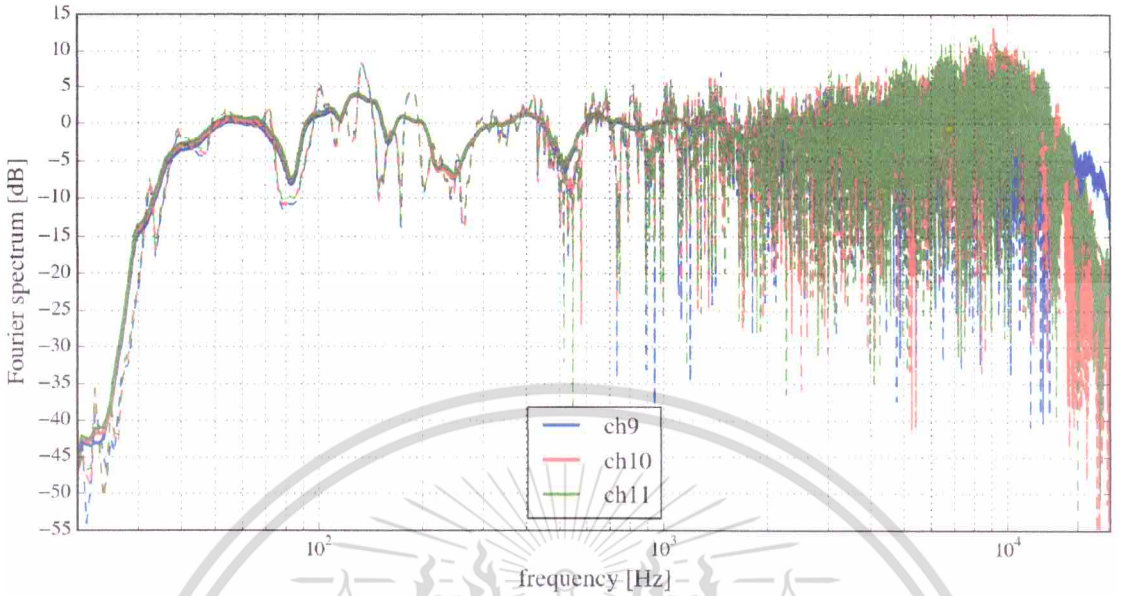


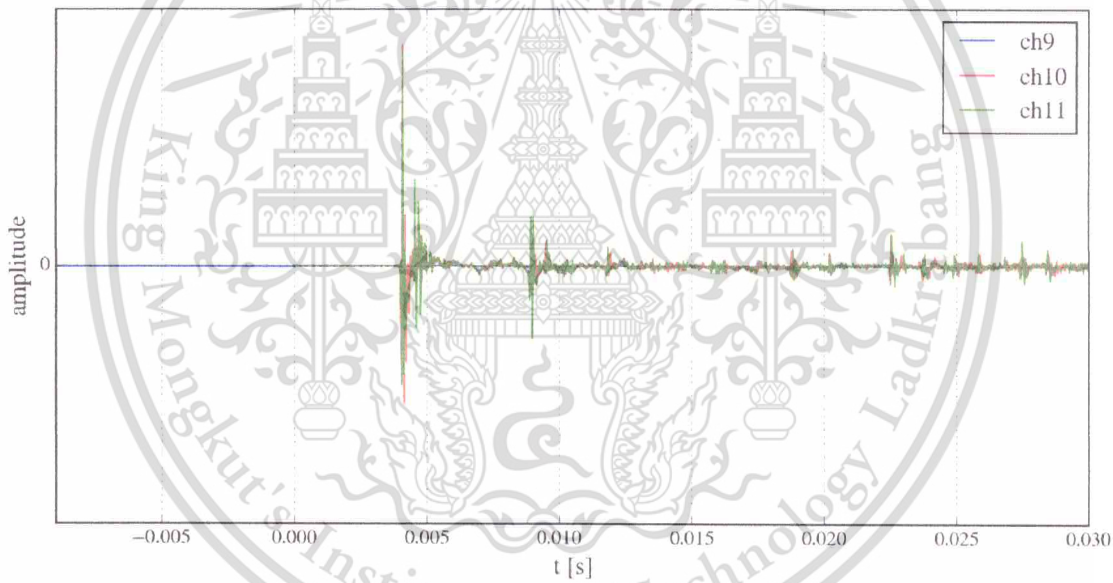


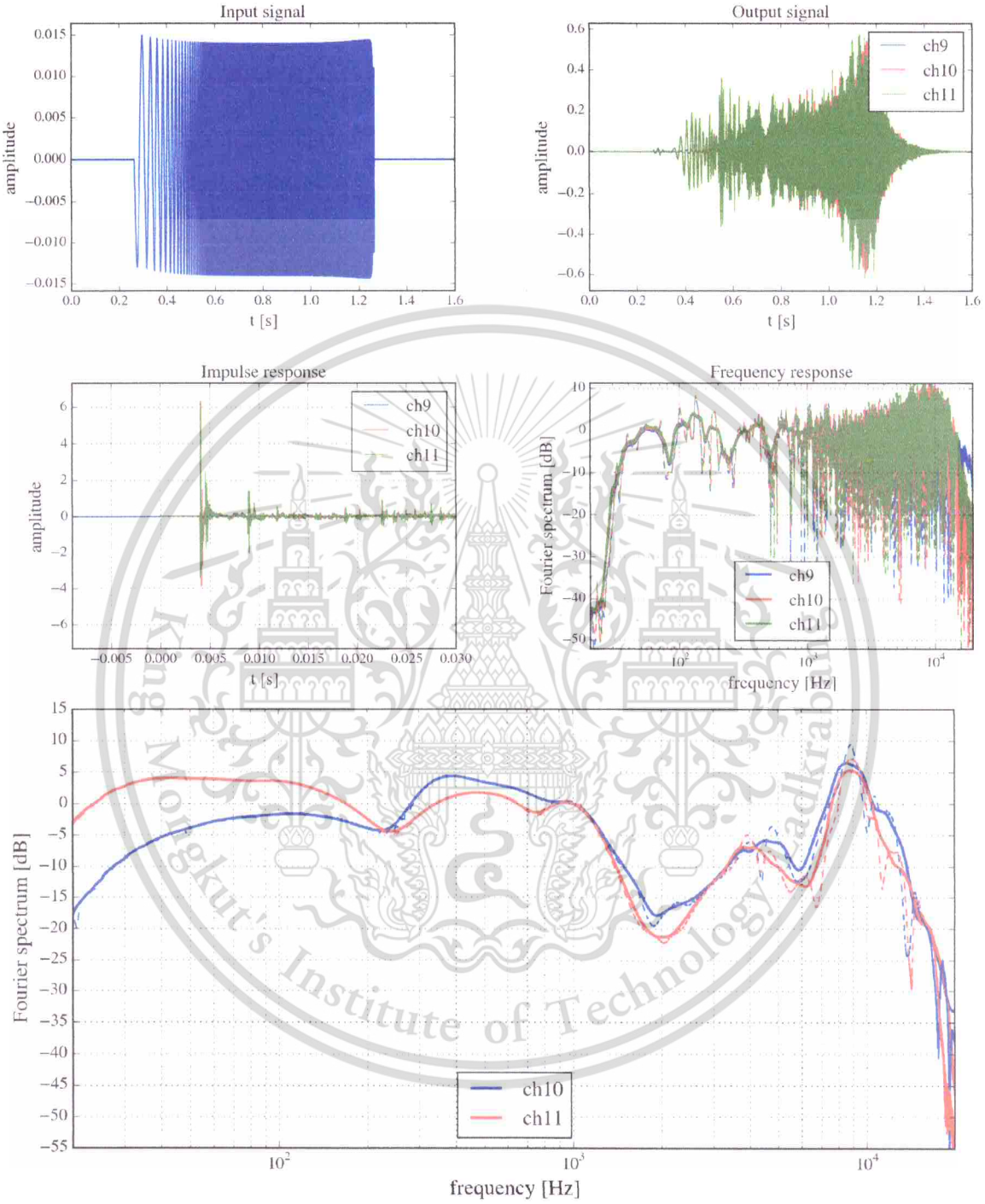


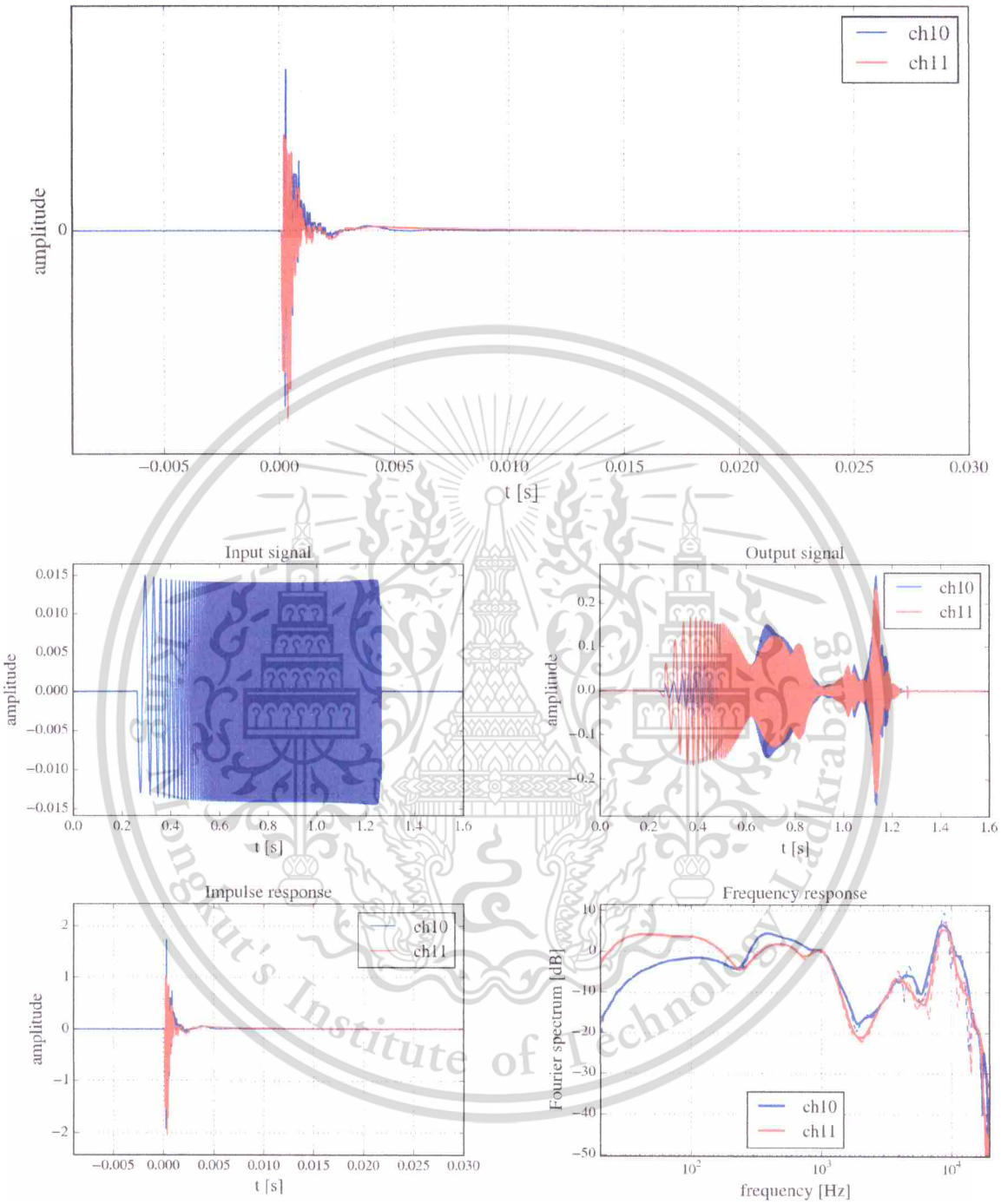


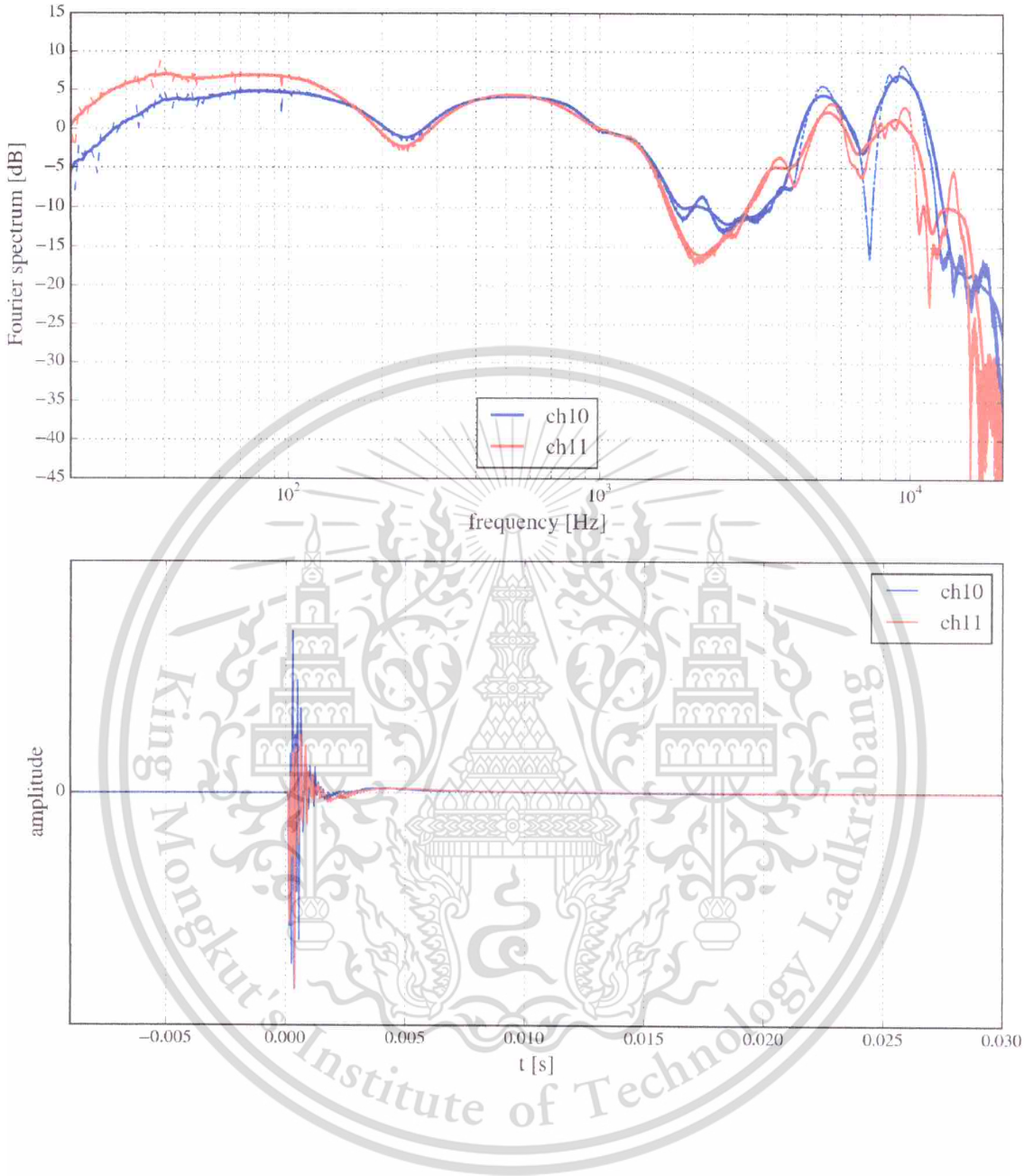


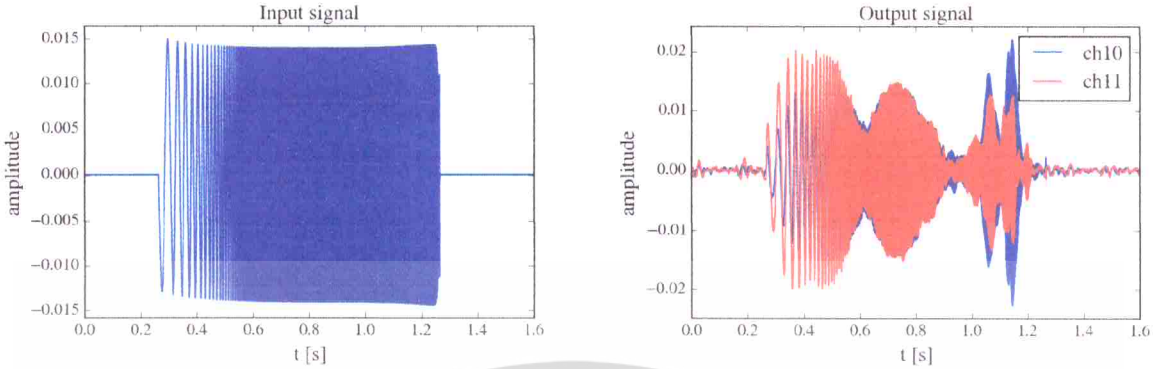


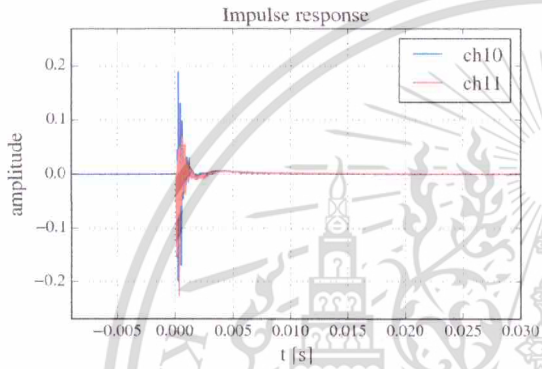


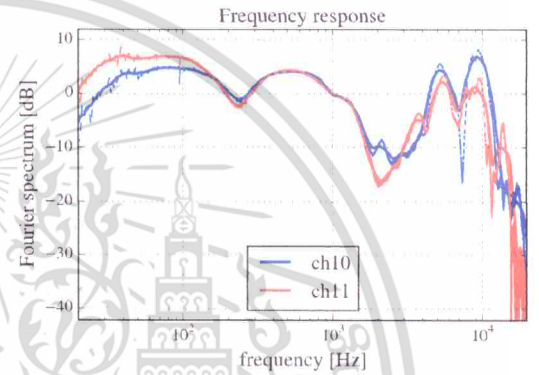


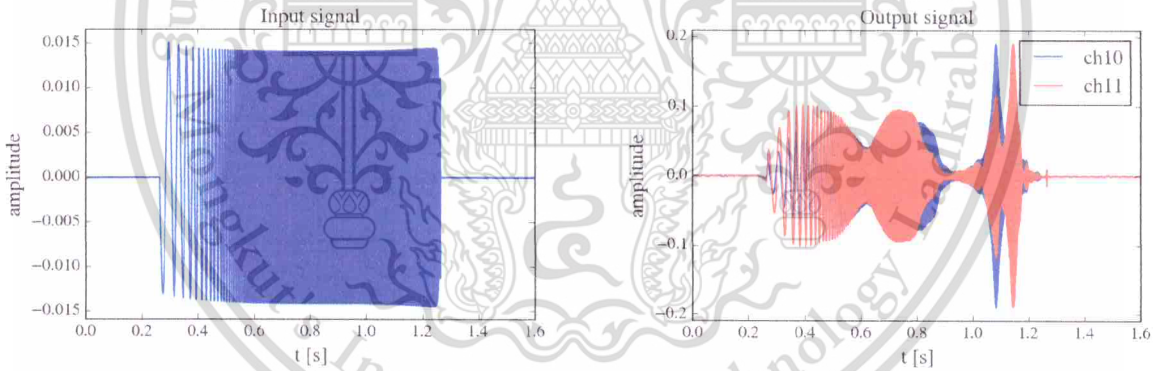


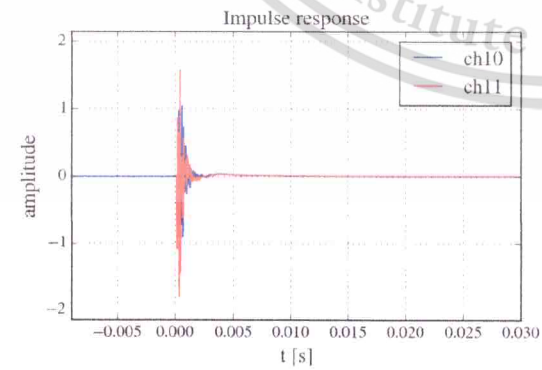


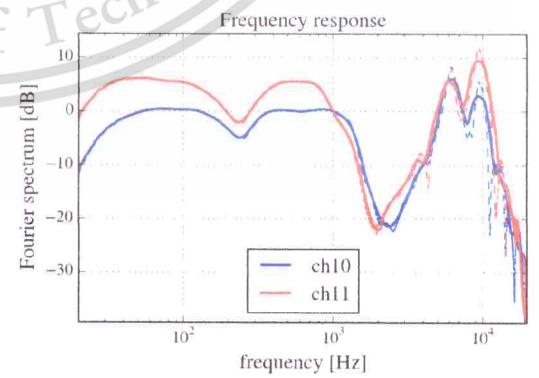












Measurement of oral-binaural room impulse response

1) Three microphones will be attached to your ears and nose.
   a. A small frame made of thin aluminium wire will be used to fix the microphone to your nose.
   b. The ear buds on the in-ear microphones are cleaned with methyl alcohol after every use.
   c. Clinical tape will be used to fix the microphone cables to your skin.
   d. This preparation step will take about 5~10 minutes.

2) First, the lowest and the highest notes you can sing with comfort will be recorded.
   a. Do not sing lower or higher than the comfortable range of your voice.
   b. This step will take about 1 minute.

3) Then, you will sing a scale with 'ah' or 'oh' sound from the lowest to the highest note, supported by the guide tone.
   a. While singing,
      i. Completely lean back to the chair.
      ii. Keep the same posture and maintain the same shape on your lips.
      iii. When the red dots appear, sing at the pitch of the guide tone, and stop before the next blue dots appear.
   b. You can press "Pause" button any time to take a break.
   c. You will sing a total of eight scales.
   d. This step will take about 30~40 minutes.

4) Finally, you will wear headphones and the frequency response from the headphones to the in-ear microphones will be measured.
   a. Sine sweep will be used.
   b. The volume of the sine sweep will be carefully adjusted to prevent any hearing damage.
   c. This step will take about 1 minute.

- This measurement is completely voluntary, and during the measurement, you can stop and walk away at any time!
- No monetary compensation will be made for your participation.
- Any question?

If you would like to take part in this measurement, please tick the following, date and sign.

☐ I have read and fully understood the information given above.
☐ The experimenter explained the measurement procedure in detail and answered my questions.
☐ I agree to take part in this measurement.

Date: _____

Name: _____

Signature: _____

Measurement of oral-binaural room impulse response

1) Three microphones will be attached to your ears and nose.
    a. A small frame made of thin aluminium wire will be used to fix the microphone to your nose.
    b. The ear buds on the in-ear microphones are cleaned with methyl alcohol after every use.
    c. Clinical tape will be used to fix the microphone cables to your skin.
    d. This preparation step will take about 5~10 minutes.

2) First, the lowest and the highest notes you can sing with comfort will be recorded.
    a. Do not sing lower or higher than the comfortable range of your voice.
    b. This step will take about 1 minute.

3) Then, you will sing a scale with 'ah' sound from the lowest to the highest note, supported by the guide tone.
    a. While singing,
        i. Completely lean back to the chair.
        ii. Keep the same posture and maintain the same shape on your lips.
        iii. When the red dots appear, sing at the pitch of the guide tone, and stop before the next blue dots appear.
    b. You can press "Pause" button any time to take a break.
    c. You will sing a total of 16 scales repeated once.
    d. This step will take about ~20 minutes.

- This measurement is completely voluntary, and during the measurement, you can stop and walk away at any time!
- No monetary compensation will be made for your participation.
- Any question?

If you would like to take part in this measurement, please tick the following, date and sign.

☐ I have read and fully understood the information given above.
☐ The experimenter explained the measurement procedure in detail and answered my questions.
☐ I agree to take part in this measurement.

Date: _____

Name: _____

Signature: _____