# Simulation of One's Own Voice in a Two-parameter Model

*Sook Young Won,*[1] *Jonathan Berger,*[2] *Malcolm Slaney*[2]

[1]Stanford University, [2]Microsoft Research

## ABSTRACT

It is well known that people often are uncomfortable while hearing their recorded singing and speaking voice. This unfamiliarity with the recorded voice, compared to normal hearing, is due to a different transmission mechanism; listening to one's recorded voice only involves a single air-conduction pathway, whereas the voice we hear when we sing and speak is largely due to a bone-conduction pathway. Despite the well-known phenomenon, one's own hearing has received less attention among researchers since it is a very complex process involving multiple paths from vocal cords to hearing sensation. Furthermore, we are studying the perception of living humans, thus adding more difficulty to proceed mechanical studies because of an ethical reason.

In this study, we aim to measure one's own hearing through a perceptual experiment using a graphical equalizer. We assume that if a subject matches a self-hearing and a hearing of recorded voice by altering slider levels on the equalizer, we can determine spectral characteristics of bone-conduction sound.

First, we design an equalizer consisting of a set of peak and shelf filters for eight frequency bands. Then, we conduct two experiments with different groups as asking participants to find the best fit to their own singing and speech voices by processing their recorded voice on the equalizer.

We estimate transfer functions from air conduction to one's own hearing for both singing and speaking voices based on the chosen equalizer settings. We observe that the transfer functions intra subject are relatively consistent and features mostly band-pass filters, broadly amplifying around 300 Hz to 1200 Hz. Moreover, the averaged transfer functions among subjects also present relatively high degree of similarity regardless of gender and experience level of singing. Finally, we successfully derive a two-parameter model of self-hearing as proceeding experimental data simplification and a validation experiment.

## 1. INTRODUCTION

When you listen to your own voice, the sound produced by your lungs and vocal folds is delivered to hearing organs through multiple pathways, including the air and bones. However, you are the only person who hears the bone-conducted sound since others only can hear the air-conducted part of the voice. The missing bone-conducted sound is unexpected when you hear a recording of your voice. Figure 1 shows the hearing pathway; a dashed line represents air-conduction (AC) pathway and a solid line indicates bone-conduction (BC) pathway.

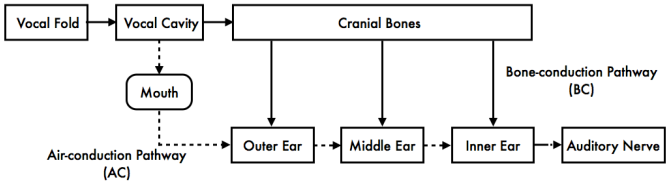


**Figure 1:** The overview of one's own hearing (von Békésy, 1954; Tonndorf, 1968).

Although the transmission pathways of AC and BC sound are different, both sounds linearly combine at inner ear and excite the basilar membrane similarly (von Békésy, 1932; Stenfelt, 2007). However, since it is impossible to capture the BC sound during vocalization, we cannot just add the BC measurement to AC sound. Therefore, to access the bone-conducted sound, we estimate a transfer function $H$ that converts the air-conducted sound to what would be heard as one's own hearing. We do this by extending the equalizer method of Shuster and Durant (2003)

$$\textit{One's own hearing} = H \cdot \textit{AC hearing}. \quad (1)$$

In this paper, we present details of the equalizer method, analysis of the results, and a simplified model of one's own hearing with the following contributions:

- We describe the implementation of a graphical equalizer software having eight frequency bands optimized for a voice processing. The software use the Cocoa GUI and STK(The Synthesis ToolKit).
- We report results of perceptual experiments with two different groups; amateur singers and professional singers. Then, we explain how to estimate the transfer function $H$ based upon the equalizer settings chosen by each subject. The shape of these transfer functions show high degree of consistency intra subject, and are even similar among inter subjects. Overall, the transfer functions feature band-pass filters mostly emphasizing the region from 300 Hz to 1200 Hz.
- Our principle contribution is deriving a model of one's own hearing by decomposing the transfer functions using a Singular Value Decomposition

(SVD) and further simplification processes. As a result, the original eight-parameter model, corresponding to the eight frequency bands used in the equalizer experiment, is simplified to a two-parameter model.

- In addition, we describe a validation process and then confirm the model with positive feedback from subjects. Since the model is relatively easy to manipulate and independent of subjects' gender and level of singing experience, we conclude that applying this model on the recorded voice is a feasible way to simulate one's own hearing and also has potential to be a practical application.

## 2. EQUALIZER EXPERIMENT

We conducted self-perception tests with two different subject groups: amateurs and professional singers. For the purpose of experimental optimization, we designed our own equalizer software in which a recorded voice was altered by peak and shelf filters in real time.

### 2.1 Experiment Software Design

A graphical parametric equalizer use peaking and shelving filters in order to amplify or attenuate frequencies in the vicinity of a specific center frequency and smoothly connected given gains by interpolation. We adopted second-order peak and shelf filters for the following two reasons. First, these filters possess self-similar shape on a log magnitude scale which agree with psychoacoustic measurement as shown in Figure 2 (a) and (b). Second, their self-similarity enables a linear least-squares optimizer to match a desired dB magnitude in cascaded filters of the equalizer (Abel & Berners, 2004).

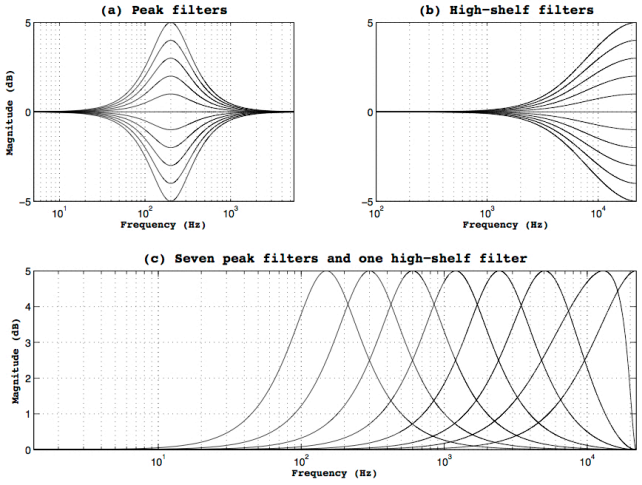


**Figure 2:** (a) Eleven peak filters and (b) eleven high-shelf filters with a peak gain from -5dB to 5dB with steps of 1dB . (c) Seven peak filters and one high-shelf filter over experimental frequency range with a 5 dB peak.

Since we manipulated speech and singing voice, the experimental equalizer should effectively encompass the vocal range. According to Fastl and Zwicker(2007), the spectrum of speech sounds extends from near 100 Hz to near 7 kHz. Thus, we placed seven second-order peaking filters and one high shelving filter equally over the log-scale frequency axis as shown in Figure 2 (c). The center frequencies of the peak filters were at 150 Hz, 300 Hz, 600 Hz, 1200 Hz, 2400 Hz, 4800 Hz, and 9600 Hz, and the high shelf filter ended at the Nyquist rate – 22050 Hz in our recording setting. Consequently, we had eight frequency bands corresponding to eight sliders in the equalizer. Through several pilot tests, the necessary range of those filters was set from -5dB to 20dB with steps of 1 dB.

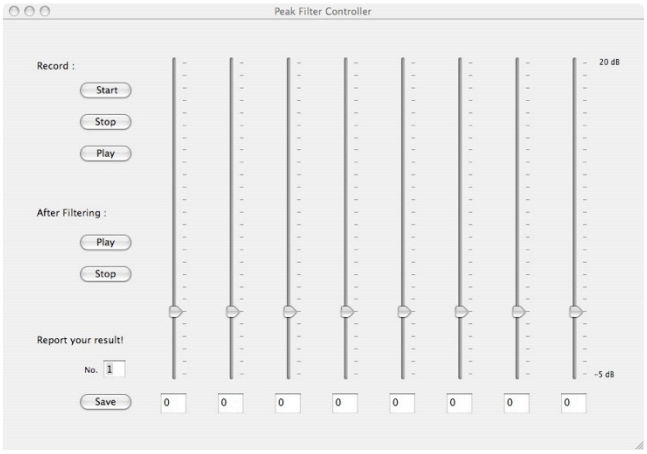


**Figure 3:** Screenshot of the equalizer experiment software

We designed a graphical user interface to allow a user to easily choose parameters and control the pace of the experiment. The overall design consists of four control parts: (1) recording the subject's voice, (2) interactively selecting a set of different filters, (3) comparing the processed sound resulting from the filter setting selections to one's own voice and (4) finally, saving the selected settings and sound files for the later analyze.

The usage of the experimental software is as follows. A subject first records his/her voice using **Start** and **Stop** buttons in a '**Record**' panel on top-left of the interface. A third button in the record panel labeled **Play** allows for playback of the last recording. Eight sliders in the center of the GUI control the peak and high-shelf filters shown in Figure 2. A numeric field below each slider displays a dB peak level of each filter, which the subject can choose based on her/his preference. Instantaneous changes of the sound are heard either by moving the sliders or by typing a number in the value field. In addition, the latest filtered state can be heard at any time by selecting the **Play** button in the panel labeled '**After Filtering**'. Once the subject encounters the filtered sound bearing the closest timberal similarity to the subject's own voice, the subject submits the filter setting by

pressing the **Save** button on bottom-left of the software display, saving the subject's choices in a text file for analysis.

## 2.2 Experiment Setup

There were two subject pools in our study – a group of 8 amateurs and 13 singers with professional training, and the experiments were carried out at two different locations with each group.

We recruited eight participants who studied at Stanford University for the first experiment. There were 4 males and 4 females in their 20's to early 30's. All participants had non-professional musical training with various instruments such as piano and flute, but not professional vocal training. Thus we defined this group as amateur singers. We carried out the first experiment at the listening room located at CCRMA (Center for Computer Research in Music and Acoustics), Stanford University.

After the first experiment, we expanded the experiment with participants having professional vocal training. We expected that the professional singers might be more sensitive to change of their vocal timbre by the filtering process. Thus we recruited 13 undergraduate students in early 20's who majored in classical vocal music at Seoul National University in Republic of Korea. The subject group consists of six sopranos, one mezzo-soprano, four tenors, one baritone, and one bass. All participants received monetary compensation. The second experiment was carried out in an anechoic room located at Applied Acoustics Laboratory in Seoul National University.

We used exactly the same hardware equipment for both experiments. We recorded vocalizations using a Schoeps CMC microphone placed approximately two inches from the lips of subjects. The recorded signal was digitized using a MOTU Traveler and fed to the experimental software running on a MacBook Pro. As described above, subjects applied filters to the recorded signal and confirmed the changes heard through an AKG K240 headphone connected to the audio output of the MOTU Traveler used for D/A conversion.

## 2.3 Procedure

Participants were asked to sing eight notes over one octave with a vowel 'Ah' and to speak four short sentences including their name, age, living city and a random sentence. The vowel 'Ah' was chosen to maximize clarity of singing voice timber and its spectral change. The range of singing samples was from C3 to C4 for males and from C4 to C5 for females.

As the first step, participants sung and recorded 3 to 4 seconds with a vowel 'Ah' at C3 (for male) or C4 (for female). Right after recording a sample voice, they used the experimental software to find the closest filter setting to their own voice by altering slider levels. To find the best fit, they sung the same note several times to compare the filtered sound and their own hearing. Usually, it took ten to fifteen minutes to find the best filter set for the first sample, and then it got much shorter. The reference singing note were provided in a given order C3, G3, D3, F3, A3, E3, B3, C4 for male subjects and speech samples were placed between sung notes. The references notes were one octave higher for female subjects.

An entire experiment took from forty minutes to an hour, including time for listening to introduction about the experiment, conducting the equalizer experiment, and giving feedback. Participants marked a total of twelve preferred equalizer settings corresponding to eight singing and four speech samples.

## 2.4 Result and Analysis

**Participants' choices on equalizer**

As a result of the two experiments, we obtained 21 matrices arranging the result of each subject's preferences. Each matrix consists of 12 rows (number of voice samples) and 8 columns (number of sliders). Table 1 shows an example of the experimental result matrix from a male subject in the first group. We listed the results by singing pitch from low to high and then speech. Each number is between -5 and 20 and represents the peak gain of filters in dB according to the participants' choice during the equalizer experiment.

| Sample Type | Slider Order | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| D | 13 | 14 | 12 | 10 | 9 | 8 | 6 | 2 |
| D3 | 14 | 13 | 12 | 10 | 8 | 7 | 6 | 2 |
| E3 | 11 | 13 | 13 | 12 | 10 | 8 | 6 | 3 |
| F3 | 10 | 11 | 13 | 13 | 11 | 9 | 7 | 4 |
| G3 | 11 | 12 | 13 | 13 | 11 | 9 | 7 | 4 |
| A3 | 11 | 11 | 12 | 13 | 12 | 10 | 8 | 4 |
| B3 | 10 | 11 | 12 | 13 | 12 | 10 | 8 | 6 |
| C4 | 7 | 9 | 11 | 12 | 12 | 12 | 11 | 8 |
| Speech 1 | 11 | 12 | 12 | 10 | 8 | 6 | 4 | 2 |
| Speech 2 | 15 | 12 | 11 | 11 | 10 | 9 | 8 | 8 |
| Speech 3 | 12 | 13 | 13 | 12 | 11 | 9 | 7 | 5 |
| Speech 4 | 11 | 12 | 13 | 13 | 11 | 9 | 6 | 5 |

**Table 1:** Peak gains of sliders. An example test result of a male subject 1 from the first equalizer experiment.

**Estimating a transfer function of one' own hearing**

We estimated the overall transfer function via the following process. We calculated the coefficients of eight filters by applying chosen peak gain and then filtered a 3 second-long impulse though a cascade of peak and shelf filters. Finally, the overall filtered impulse was transformed to frequency response by FFT. This gave us a 2048-point estimate of the desired transfer function.

After applying this estimation process for each sample (each row of Table 1), subject 1 had twelve preferred transfer functions. In general, the loudness of a subject's voice varies over the singing range; individuals commonly tend to sing louder as the pitch increases and this affects the absolute level of the chosen filter

settings. Therefore, we normalized the transfer functions by subtracting the maximum magnitudes so that all normalized transfer functions had same peak level, 0 dB. Then, we calculated an average of normalized transfer functions for comparison with other subjects' results. As displayed in Figure 4, the normalized transfer functions of subject 1 in group 1 show a high degree of consistency; most transfer functions have peak between 300 Hz to 1200 Hz (Slider no. 2 to no. 4) except one for singing note C4 which has a peak at 2400 Hz, the fifth peak filter. The standard deviation of these transfer functions is 3 dB. In addition, we could not see noticeable influence by pitch of singing or speech.

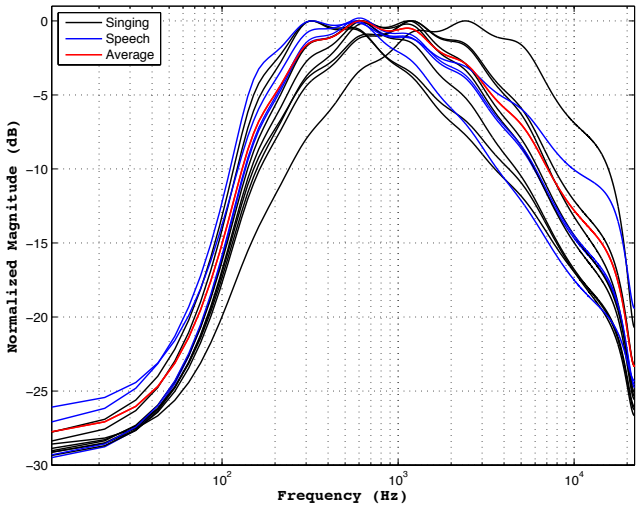


**Figure 4:** Normalized frequency responses of twelve singing and speech samples and the average of those from a subject 1.

### Analysis

We applied this process to all result matrices, and then observed that all participants had self-similar transfer functions of which standard deviations are less than 6 dB intra-subject. Figure 5 displays the eight averaged transfer functions for group 1 and thirteen averaged transfer functions for group 2 separately. In group 1, most transfer functions have peaks at 1200 Hz (the 4th slider) while many of the peaks are at 600 Hz (the 3rd slider) in group 2. Except the two outliers in group 1, all transfer functions in both groups feature a broad band-pass filters with a strong emphasis from 300 Hz to 1200 Hz – some transfer functions had even broader bandwidth extended as high around 2400 Hz. Furthermore, all averaged transfer functions strongly resemble each other. In order to characterize the experimental results, we plot transfer functions by gender per experimental group. However, there is no clear difference caused by gender and level of singing experience.

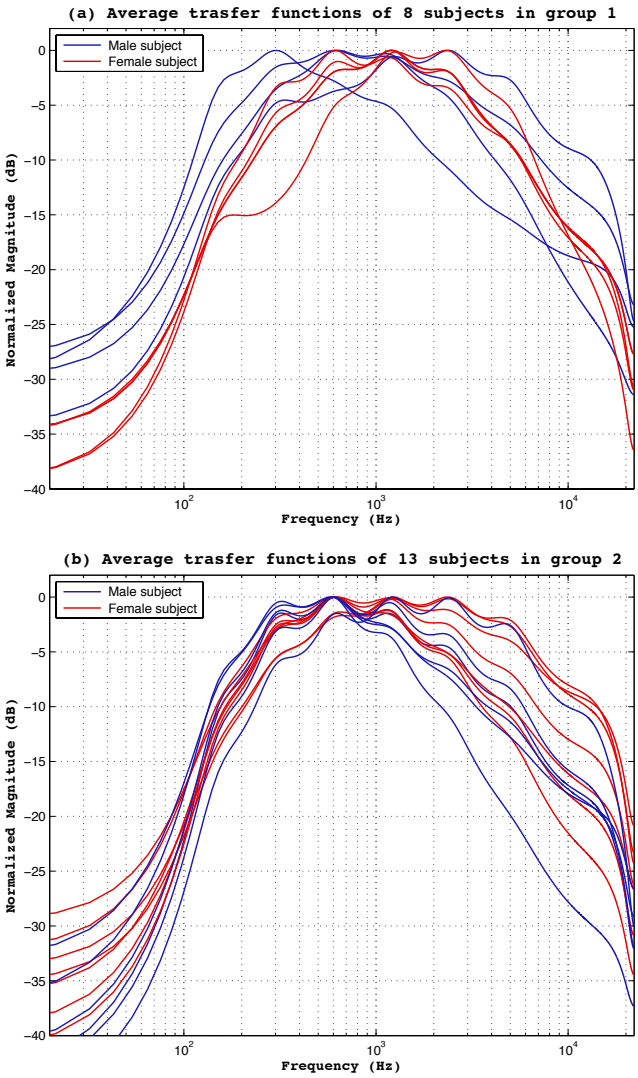


**Figure 5:** Averaged transfer functions of equalizer experiments.

# 3. MODELING

## 3.1 Singular Value Decomposition

We used a singular value decomposition (SVD) to find a low dimensional approximation of the measured transfer function data. The SVD takes a matrix $A$ and decomposes the matrix into three matrices; two unitary matrices ($U$ and $V$, where $V^*$ is conjugate of $V$) and one diagonal matrix ($S$)

$$A = USV^*. \quad (2)$$

The total subjects from two experiments were 21. However, we excluded two outliers' results in group 1 for better modeling. Consequently, we considered 19 averaged transfer functions, which transformed air conduction recordings to one's own hearing, into a matrix $A$ (2048×19). After applying the SVD in

Matlab, we obtained three matrices representing *filter shapes, filter weights* and *individual differences* corresponding to $U$ (2048×19), $S$ (19×19), and $V$ (19×19) respectively.

## 3.2 Advanced Modeling

The first two diagonal values of the filter weight matrix $S$ (blue and green circles in Figure 6. (b)) are significantly higher than the other 17 filter weights. By approximating the full transfer functions with the first two singular vectors, we produced a matrix $A'$ having high accuracy and reduced dimensionality. Since the differences between matrix $A$ and $A'$ are relatively small – mostly within -2 dB and 2 dB, we wished to develop a model to simulate one's own hearing with these simplified matrices. We performed several pilot tests with the two SVD filters (the first and second eigenvectors in green and blue lines in Figure 6 (a)). However, some participants complained that it was confusing to use the two filters affecting on the whole frequency range.

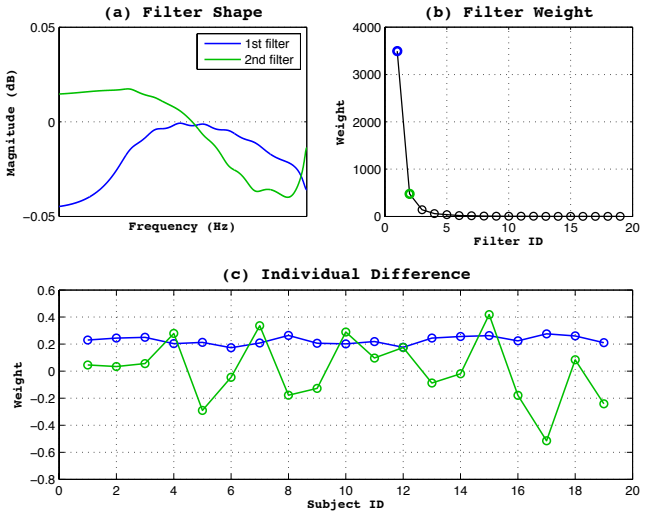


**Figure 6:** Decomposed and simplified three matrices by SVD. (a) U′ - frequency responses of the first two filters. (b) S - the original diagonal matrix representing filter weight. (c) V′ - individual weight differences of the first two filters.

In order to avoid this side effect of the model with the two SVD filters, we designed a one-parameter model. We set a fixed weight for the first SVD filter since its individual difference (Figure 6 (c)) was relatively small compared to its weight. Then, the model had only one parameter that controlled the weight of the second SVD filter. Another pilot test proved that this one parameter model was simple and easy to control. However, it lost the ability to alter vocal timbre as much as subjects wanted. Therefore, we considered a third model that aimed to have balance between easy controls and flexible simulation.

Our subjects in pilot tests told us that altering the transfer functions by frequency was easiest. Thus, we separated the second filter into two filters at 864 Hz where the filter crosses zero dB. Consequently, the model $M$ has one constant filter and two variable filters as shown in equation 3

$$M = Filter\ 1 + \alpha \cdot Filter\ 2 + \beta \cdot Filter\ 3. \quad (3)$$

During another trials with the model $M$, we finally had positive feedbacks from subjects. Therefore, we decided to perform a validation experiment with this two-parameter model.

# 4. VALIDATION EXPERIMENT

We recruited another 14 participants, consisting of 7 males and 7 females who studied at Stanford University. Among them, 9 participants had musical singing experience, either musical theatre singing or opera singing. The experiment was carried out in the listening room at CCRMA, as in the first equalizer experiment. We also used the same hardware equipment for recording and listening to the voice.

## 4.1 Experiment Software Design

We modified the previous experimental software by placing two sliders, corresponding to filter 2 and filter 3, instead of eight sliders manipulating eight frequency equalizer bands. The left panel of the validation experiment software – the six buttons for recording voice, listening unfiltered and filtered voices, and saving the choices – was exactly same as the original equalizer experiment software. The filter weight α for filter 2 are from -2 to 2 with steps of 0.2, and β for filter 3 are from -0.5 to 0.5 with steps of 0.1. The experimental range and step size for filter 2 and filter 3 were tuned through several pilot tests. Figure 7 displays the three filters used in the software to simulate one's own voice.

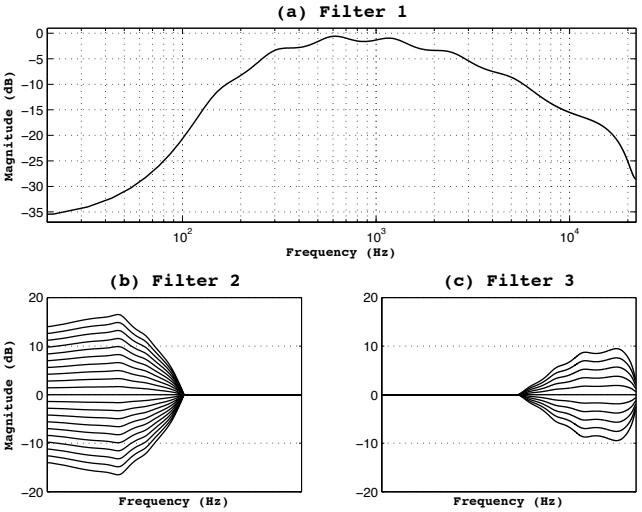


**Figure 7:** Three filters used in a validation experiment.

## 4.2 Procedure

Participants recorded three singing and three speech samples of their own choice. Similar to the previous experiments, they attempted to find the best match of their imagined self hearing by adjusting two sliders that controlled the filters applied to their

recorded voices. Additionally, they were asked to provide proximity score about their selected filtered voice compared to their own hearing, where the air conduction hearing was zero and their own hearing (AC + BC) was 100.

## 4.3 Result and Analysis

We obtained filter weights α and β for each recorded sample, thus there were six measurements per participant. The averaged transfer functions from each participant are presented in Figure 8. As expected, the transfer functions are close to each other, although subjects do not know the details of filters processed by the software. Compared to the estimated transfer functions from the original experiments, the peaks of average transfer functions from the validation experiment are lower than the original experimental results perhaps due to the limitation of the controls and predefined filter shapes. More importantly, most participants gave proximity score above 80, up to 95 – one subject answered that his filter selections had 70. Again, we could not find any factors differentiating the transfer functions among subjects such as gender and singing experience.

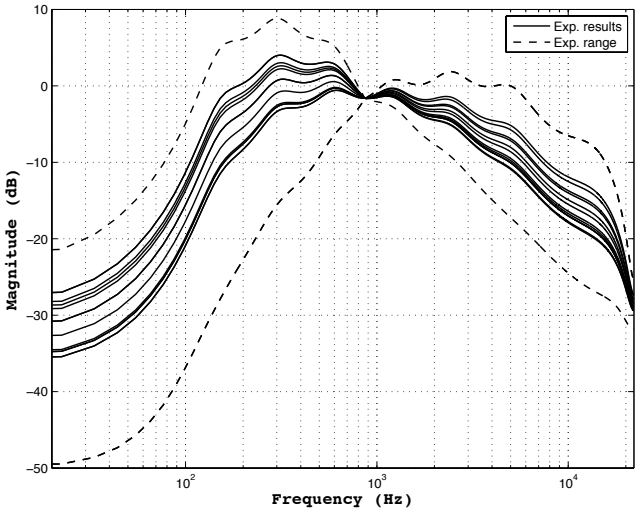


**Figure 8:** Results of the validation experiment – 14 averaged transfer functions in solid line and the experimental range represented in dashed line.

# 5. CONCLUSIONS

In our two equalizer experiments, we derived transfer functions *H* from twenty-one subjects and observed that most transfer functions had strong emphasis from 300 Hz to 1200 Hz. This result matches previous studies investigating BC characteristics by measuring the resonant frequencies of a skull. Franke(1956) applied a vibrating piston to a dry skull and found the first resonance to be at 800 Hz. In the same experiment with a skull filled with gelatin, the resonance was reduced to 500 Hz. Later, Pörschmann(2000) estimated the frequency response of bone conduction with a masking experiment and found the amplified region from 700 Hz to 1200 Hz and rapid attenuation above 5 kHz, and these partially accorded with our estimated transfer functions from air conduction to one's own hearing.

Since we obtained a high degree of similarity among the estimated transfer functions inter-subject, we were able to derive a model simulating one's own voice by altering the air-conducted voice with two variable filters. Furthermore, we confirmed the model with a validation experiment. The validation experiment software provided simpler means to simulate one's own voice, however, the accuracy of the transfer functions seems to fell slightly off compared to those processed by the equalizer software.

As a next step, we plan to implement a practical application to allow a user to reproduce his/her own voice in real time. This can be applied in hearing aids for naturally amplifying user's own voice as well as other's voices.

# 6. REFERENCES


Abel, S. J., & Berners, D. P. (2004). Filter design using second-order peaking and shelving sections. *In proceedings of the International Computer Music Conferences,* Miami, FL.

Fastl, H., & Zwicker, E. (2007). Psychoacoustics: Facts and Models, chapter 9, 259–286. Springer, London, 2nd edition,

Franke, E. K. (1956). Response of the human skull to mechanical vibrations. *The Journal of the Acoustical Society of America, 28*, 1277–1284.

Pörschmann, C. (2000). Influences of Bone Conduction and Air Conduction on the Sound of One's Own Voice. *Acta Acustica united with Acustica, 86(6),* 1038–1045.

Shuster, L., & Durrant, J. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders, 36(1),* 1–11.

Stenfelt, S. (2007). Simultaneous cancellation of air and bone conduction tones at two frequencies: Extension of the famous experiment by von Békésy. *Hearing Research, 225(1-2)*, 105–116.

Tonndorf, J. (1968). A new concept of bone conduction. *Arch Otolaryngol, 87(6),* 595–600.

von Békésy, G. (1932). Zur theorie des hörens bei der schallaufnahme durch knochenleitung. *Annalen der Physik, 13,* 111–136.

von Békésy, G. (1954). Note on the Definition of the Term: Hearing by Bone Conduction. *The Journal of the Acoustical Society of America, 26(1)*, 106–107.