Simulating the sound of one's own singing voice

Sook Young Won **Jonathan Berger** **Song Hui Chon**
CCRMA, Stanford University.
Stanford, CA, U.S.A.
sywon@ccrma.stanford.edu, brg@ccrma.stanford.edu, shchon@ccrma.stanford.edu

ABSTRACT
The often-discomforting percept of unfamiliarity of one's own voice when heard as a pre-recorded analog or digital signal has received limited attention in perception research. In this paper a digital filter that simulates internal hearing is studied in terms of self-perception of singing by a trained female singer. The processed sound is evaluated to determine the proximity of match to the sound of the imagined sung voice. We perform two experiments. In the first experiment we investigate the characteristics of skull vibration and determine the appropriate transfer function from air to bone conductance. Our results correspond to classical research that describes cranial related boost of low frequencies and attenation of high frequencies. With this observation we perform a second experiment in which a female singer adjusts the magnitude of the transfer function in the low-frequency region to match her perception of the sound of her singing voice.

Keywords
Internal hearing, Bone conduction, Self-perception, Singing voice

INTRODUCTION
There is often a sense of disappointment when a singer hears a recording of his or her own singing voice. The perceptual disparity between the live and recorded sound of one's own singing is due in large part to the effect of bone-conducted signal to the auditory mechanism.

A good deal of research has focused on cranial bone conduction, the characteristics of skull vibrations and the transmission paths of these vibrations to the ear described by von Békésy (1954) and Stenfelt(2003).

Understanding the mechanical functionality of the hearing system provides a basis for perceptual studies to further understand the self-perception of vocalization.

In the area of speech perception, the ratio of bone-conducted signal to signal processed through the ear has been discussed by von Békésy (1949) and studied by Maurer and Landis (1990). In addition to these experiments, Shuster and Durant (2003) have attempted to determine a transfer function to describe and simulate the bone conducted speech signal. The result supports earlier findings that the speaker's voice is heard as a substantially low-pass filtered signal.

The experiments described in this paper concentrate on the singing voice which has been the subject of only limited research in this regard. Aside from the purely musical interest one rationale for focusing on singing is that the temporally extended vowels characteristic of singing provide greater clarity for timbral and spectral based study of signal characteristics. The approach used here compares the perceived processed recorded sound with that of the self-produced sound.

EXPERIMENT 1
In the first experiment vocalizations were recorded both as signals propagated through air and as bone conducted signals. The two classes of sounds were analyzed and compared using a variety of sung phonemes and vocal production types. The experiment is described in three stages, first signal acquisition, second, analysis and estimation of the transfer function, and finally, implementation of a filter that simulates the bone conducted sound.

Experiment Design
The first step of the experiment is setting up a reliable recording tool for capturing skull vibrations. The bone-conducted sound is transduced by a ceramic piezo-electric

transducer. The air-conducted sound was recorded using a Sennheiser ME65/K6 microphone. Both signals were recorded into two channels of portable Marantz Professional CDR300 direct CD recorded.

To acquire two types of sound simultaneously, we attach the piezo contact microphone on the flat region close to the ears, and place the Sennheiser ME65/K6 in front of the mouth of the subject.[1] The subject then speaks a single vowel and sings various patterns such as a single note, a chirp and a scale.

As is the case with many perceptual experiments the use of unambiguous test signals (in this case a noise impulse, white noise or Golay code sequence for testing impulse response) is of limited use for testing perception in real-world situations. However, even subtle muscle movements of a singer can change the resonant field and cause significant variation in the resulting vocal timbre.

Therefore we attempt to investigate not the transfer function of the human skull itself, but rather a combined resonant model in which the vocalized air-conducted sound and the bone-conducted sound jointly effect the singer's perception of his/her own voice.

Estimating the Transfer Function

The recorded time domain signals are transformed into frequency domain signals, as computed by taking Discrete Fourier Transform. Then we extract the spectral envelopes of signals by Linear Predictive Coding and get the geometric means of the air-conducted sound and of the bone-conducted sound which are represented as $A(wk)$ and $B(wk)$ respectively. The transfer function $H(wk)$ from the air-conducted sound to the bone-conducted sound can be obtained by a simple division as the following formula.

$$H(\omega_k) \triangleq \frac{B(\omega_k)}{A(\omega_k)}$$

Result

Figure 1 shows the estimated transfer function as the result of the first experiment. As setting the air-conducted sound for reference, magnitude value of bone-conducted sound below 3200Hz is boosted, and above it the magnitude is cut. The peak goes up to 19dB and the trough goes down to -23dB.

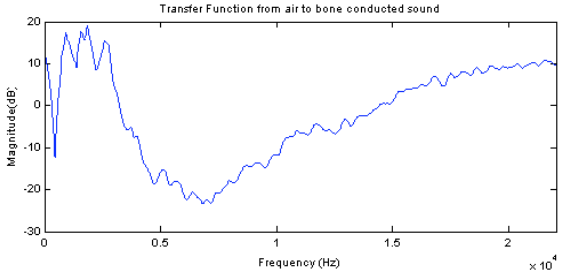

Figure 1. The transfer function from the first experiment

[1] The subject is a classical soprano singer with vocal control of head voice or chest voice.

As evident in the transfer function in figure 1, the magnitude of bone-conducted sound is boosted below 3200Hz, above which the magnitude is attenuated. These characteristics are in correspondence to the common view that solid materials cut high frequencies because of their material damping properties. They also explain the tendency for preference of the timbre of one's own singing voice to that of a recording of the same signal. This may be due in part to masking effects since the boost in low frequencies can conceal defects of singing techniques such as inconsistent pitch and irregular vibration. In addition, the low frequency boost adds a sense of timbral enrichment .

Implementation

The last stage of the experiment is implementing the filter based upon the transfer function. We then processed the original recording through air transmission and compared the filtered output to the original recordings.

Figures 2 (a) and (b) show the original recordings by the the piezo and microphone. The figure (c) shows the filtered sound simulating the bone-conducted sound.

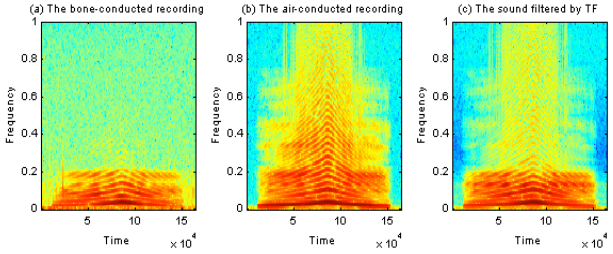

Figure 2. (a) the bone-conducted sound, singing a chirp pattern (b) the air-conducted sound, singing a chirp pattern (b) the filtered sound by the transfer function

EXPERIMENT 2

In the second experiment we allow a singer to adjust the magnitude of the transfer function with the objective of matching the output sound to the internal sound of her own singing voice.

Experiment Design

For recording air-conducted sound as input samples, a B&K 4133 microphone and a measurement amplifier B&K Nexus were used. The air-conducted sound go into the portable direct CD recorder, Marantz Professional CDR300, in an anechoic recording room at Plantronics, Inc. Each recording consists of approximately 3 seconds of the vowel 'Ah' sung at a single pitch, in the range between C4 and G5. Different voicing styles were used in different frequency range, including head voice in high frequencies, chest voice in mid-range, and throat voice in low frequencies. Other sung vowels, including 'Ee' were also recorded, but ultimately excluded from the experiment due to range limitations. However in self-perception tests these sounds seemed to lead to similar results as with the test set of recorded sounds used.

The First Self-perception Test

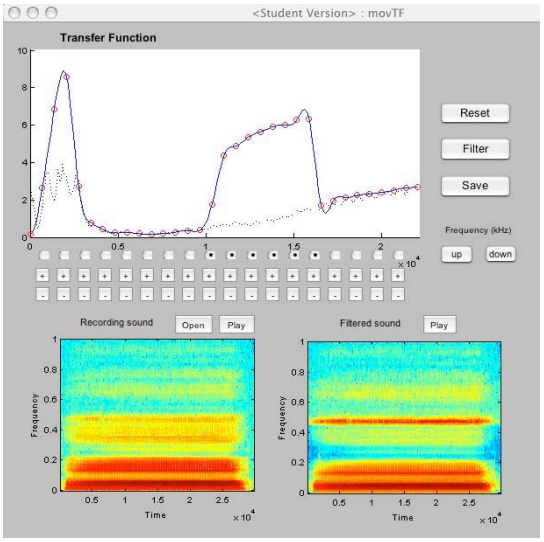

Figure 3. The program for the first self-perception test

This program was designed to explore various transfer functions with the purpose of delineating the range of timbral sensitivity of the resulting filter. The default magnitude values on frequency bins, dashed line in a upper axes, are based upon the experiment result in Figure 1. The circled points on the Transfer Function curve on same axes are movable by mouse clicking and dragging and the values between dots are interpolated using cubic spline. With the adjusted Transfer Function curve, the input sound recording (shown in below left axes) is filtered and then the timbre of filtered sound (shown in below right axes) is observed.

After numerous variations of magnitude settings were applied to this curve, the first three moving points were determined to be the most important in terms of their effect on the perceived timbre of the filtered output. Therefore, the next self-perception test was done with fixed points (magnitude) for most of the frequency bins, with the exception of the first three points.

The Second Self-perception Test

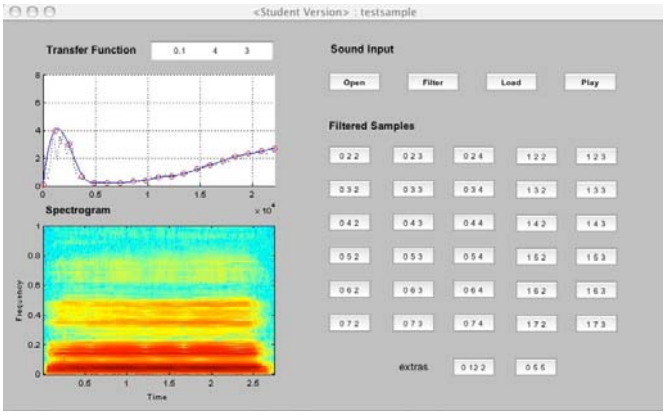

Figure 4. the program for the second self-perception test

On the right side of the program, there are four buttons with different functionalities. Those along the first row concern operations on the recorded sample – an *Open* button to select, open and load a specific recorded input file, a *Filter* button to convolve the input signal with the impulse response from the modeled transfer function whose magnitude are pre-specified. The *Load* button loads pre-filtered samples in a .mat file, avoiding the computational time needed to filter on the fly. The *Play* button plays the original recorded sample.

In the Filtered Samples part, each of the buttons play a sound sample pre-filtered by the transfer function having the magnitude of the first three circles corresponding to the three numbers on it. After the specific sample is played, those three magnitude values appear in the edit box next to the 'Transfer Function' caption. Note that 0 on these buttons actually mean 0.1 as shown in the edit box. Through trial and error experimentation the number of buttons included in the template was reduced to a more manageable size. The two buttons at the bottom of the template, next to the *extras* caption, show two extreme settings that were considered for the experiment. The first one [0 12 2] is almost symmetrical with a huge boost around the region of the second circle. This muffled sound was subjectively considered by the first author to be similar to the (primarily bone-conducted) sound heard when she listens to herself sing while plugging her own ears.

The other one [0 5 5] is another extreme where the second and the third coefficients are set to the same value, forming a sort of a plateau. The singer in the experiment complained that the output sound was too bright.

The perceptual test was mostly done by the first author who recorded the original sound samples. The author marked the closest match and several acceptable matches as comparing the original recording sound, filtered sound samples, and live singing voice.

RESULT

Table 1. Result of the second self-perception test. Acceptable matches are marked as '✓' and the closest matches are

marked as ‘’✔✔’ over difference frequencies with difference filters

	C 4	D 4	E 4	F 4	G 4	A 4	B 4	C 5	D 5	E 5	F 5	G 5
0 3 2	✔	✔	✔	✔		✔		✔	✔		✔	✔
0 4 2	✔	✔	✔	✔	✔	✔✔	✔	✔	✔	✔	✔	✔✔
0 5 2	✔	✔	✔	✔	✔	✔		✔	✔	✔		✔
0 6 2					✔							
0 3 3									✔			
0 4 3	✔✔	✔✔	✔	✔✔	✔		✔	✔	✔✔		✔✔	
0 5 3			✔✔	✔	✔✔	✔	✔✔	✔	✔	✔	✔	
0 6 3					✔							
1 3 2	✔	✔										
1 4 2	✔	✔	✔			✔	✔	✔	✔	✔✔		✔
1 5 2	✔	✔	✔		✔	✔	✔			✔		✔
1 4 3	✔	✔	✔	✔				✔	✔	✔	✔	
1 5 3			✔	✔	✔			✔✔			✔	

As can be seen in Table 1, the results in most experiments have common preferred settings for acceptable matches, which are with [0 3 2], [0 4 2], [0 5 2], [0 4 3], [0 5 3], [1 4 2], [1 5 2] and [1 4 3]. Perceptually closest matches exhibited the setting of [0 4 3] and [0 5 3]. And a few cases are appeared in different preferred pattern of [0 6 2], [0 6 3] and [0 3 3], but they were quite rare comparing with other cases.

It seems that with these exceptions most experiments from one subject have a common set of preferred filter settings.

As the first author compared her live singing voice to each of the filtered sounds, the other two authors verified that the recorded samples are close to what they hear in the first author's live singing. They also concurred that the sample that the first author picks to be the closest perceptually is much darker in timbre, with a big boost in low frequency. Even though the perceptually closest sample sounded different from what they usually hear, the two authors agreed that this is feasible judging from their personal experience of discrepancy between hearing their own voice normally and hearing it from a recording.

CONCLUSION

With experiments, various singing styles such as throat voice, chest voice and head voice were used at the time of recording, and still yielded pretty consistent result in terms of ‘acceptable range’ over the different pitches. It was observed that in most cases the preferred filter settings fall close in a region. Due to the exceptions, we could not come up with a model that satisfies all the cases. We hope to be able to develop this model using adaptive signal processing technique in the future.

ACKNOWLEDGMENTS

We would like to thank professor Sunil Puria and Chales Steels in OtoBiomechanics group at Stanford for providing the knowledge about biomechanical engineering. We also thank Ryan Cassidy for helping recording procedure.

REFERENCES

v. Békésy, G. (1949). The Structure of the Middle Ear and the Hearing of One’s Own Voice by Bone Conduction, *The Journal of the Acoustical Society of America, v21*(3), 217-232

v. Békésy, G. (1954). Note on the Definition of the Term: Hearing by Bone Conduction, *The Journal of the Acoustical Society of America, January, v26*(1), 106-107

Hakansson, B., Brandt, A., Carlsson, P., and Tjellstrom, A. (1994). Resonance Frequency of the human skull in vivo, *The Journal of the Acoustical Society of America, v95*(3), 1474-1481

Maurer, D. and Landis, T. (1990). Role of Bone Conduction in the Self-Perception of Speech. *Folia Phoniatrica, 42*, 226-229.

Purcell, D. W., Kunov, H., and Cleghorn, W. (2003). Estimating bone conduction transfer functions using otoacoustic emissions”, *The Journal of the Acoustical Society of America, 114*(2), 907-918

Shuster, L. and Durant, J. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders. v36*(1), 1-11.

Stenfelt, S., Hakansson, B., and Tjellstorm, A. (2000). Vibration characteristics of bone conduction sound in vitro”, *The Journal of the Acoustical Society of America, 107*(1), 422-431

Stenfelt, S., Hato, N., and Goode, R. L. (2002) Factors contributing to bone conduction: The middle ear, *The Journal of the Acoustical Society of America, 111*(2), 947-959

Stenfelt, S., Hato, N., and Goode, R. L. (2004) Fluid volume displacement at the oval and round windows with air and bone conduction stimulation, *The Journal of the Acoustical Society of America, 115*(2), 797-812

Stenfelt, S., and Goode, R. L. (2005) Transmission properties of bone conducted sound: Measurements in cadaver heads, *The Journal of the Acoustical Society of America, 118*(4), 2373-2391