Sonicstrument: A Musical Interface with Stereotypical Acoustic Transducers

Jeong-seob Lee and Woon Seung Yeo
Audio & Interactive Media Lab
Graduate School of Culture Technology, KAIST
335 Gwahangno, Yuseong-gu, Daejeon, Korea
jslee85@kaist.ac.kr, woony@kaist.edu

ABSTRACT
This paper introduces *Sonicstrument*, a sound-based interface that traces the user's hand motions. Sonicstrument utilizes stereotypical acoustic transducers (i.e., a pair of earphones and a microphone) for transmission and reception of acoustic signals whose frequencies are within the highest area of human hearing range that can rarely be perceived by most people. Being simpler in structure and easier to implement than typical ultrasonic motion detectors with special transducers, this system is robust and offers precise results without introducing any undesired sonic disturbance to users. We describe the design and implementation of Sonicstrument, evaluate its performance, and present two practical applications of the system in music and interactive performance.

Keywords
Stereotypical transducers, audible sound, Doppler effect, hand-free interface, musical instrument, interactive performance

1. INTRODUCTION
Sonicstrument is a simple but powerful interface that detects the user's hand motions with sound. The system does not utilize any ultrasonic sound and related devices; instead, it consists of a pair of stereotypical earphones and a microphone which transmit and receive signals whose frequencies range within the transducers' bandwidths (mostly covering human's theoretical audible range) but are barely perceptible for most people. To assure its reliability in an acoustically uncontrolled environment (e.g., loud ambient noise), the system performs Doppler analysis in signal processing to detect the transducers' motions in one dimension.

Sonicstrument aims to provide the simplest hand gesture interface for general users including interactive media artists. The system utilizes commonly used bud earphones and a microphone as transmitter and receiver, and does not incorporate any extra hardware such as special ultrasonic transducer and/or wireless system that may require technical expertise to use. Also, the small and handy nature of the earphones controlled by the user makes the system highly practical and suitable for interactive performance.

As a musical interface, sonicstrument has been featured in two different performances with contrasting scenarios: 1) a smaller-size, near-field environment for computer-synthesized virtual instrument performance, and 2) an interactive dance performance at a larger scale. In both cases, the system successfully traced the motion of the user.

This paper is organized as follows: we first discuss the detection mechanism of the Sonicstrument based on the review of previous studies, and describe the design and implementation of the system. Finally we present two application examples mentioned above.

2. RELATED PRIOR WORK

2.1 Motion Detection
Motion detection has been widely adapted to numerous studies in a variety of fields ranging from pure scientific research to practical fields (i.e., sports and security) and from music to media art.

Numerous motion detection systems have been developed with a variety of detection mechanisms and sensors. Examples include sound (acoustic transducers), optics (cameras, infrared sensors), electromagnetic field (compass), and motion/vibration (accelerometers) [7]. For short-range or indoor motion detection, ultrasonic signals – sounds with frequencies greater than the highest limit of human perception – are frequently selected due to the following:

- Although systems with radio-based location techniques such as the Global Positioning System (GPS) perform well in widely open areas, they are prone to severe multipath effects when used inside buildings.
- Electromagnetic sensors may be interfered with by unexpected magnetic fields as well as metal structures.
- Moreover, optical systems generally require expensive imaging detectors and suffer from line-of-sight problems [4].

While ultrasonic motion detection systems are free from these problems and are deemed suitable for moderate-scale indoor applications, they require special transducers, which can be expensive, and/or require technical expertise for use and implementation, thereby imposing practical limitations in terms of cost and technology.

2.2 Sound-based Motion Detection
Sound can be considered as a mechanical phenomenon that contains information about a physical "event." Many attempts have been made to detect, analyze, and classify certain events only from their sounds [3, 11 13].

In this paper, we focus on sonar-style examples that use sound "propagation" characteristics for detection, most of which are based on Time-of-flight (TOF) observation [4, 6, 9, 10, 12]. In other words, these sonar-style examples consist of mobile beacons and static receivers that communicate Radio Frequency (RF) and ultrasound signals. The RF channel triggers both beacons and receivers to transmit/detect a pulse, and the time delay between the RF trigger and detected sound – assuming constant speed of sound and negligible RF propagation time – becomes the TOF and is used to calculate the distance between a beacon and a receiver. This one-dimensional (1D) distance detection can be expanded to three-dimensional (3D) localization by deploying multiple receivers and using trilateration method.

TOF method suffers from limited precision due to the irregular time delay of the system process. To solve this problem, Lopes et al. suggested a localization method that compensates for the time delay by adopting a new variable d in addition to the 3D position coordinates x, y, and z [6]. Still, this method assumes an equal processing delay *d* for all receivers, which rarely happens in reality.

Also, to enable the indoor localization of mobile devices without any special equipment, the system uses an audible sound instead of an ultrasound (a 4.01 [kHz] tone with 0.2 [s] of duration is emitted as the pulse signal). This sound is not only prone to interferences from ambient noise, but can also be perceived by (and irritating to) most people.

3. Features of Sonicstrument

As mentioned above, the Sonicstrument addresses these issues with the following key features:

3.1 Doppler Effect

The Sonicstrument measures the Doppler shift of beacon signals. Compared to the aforementioned TOF approach, this method is more robust to ambient noise. Also, it is independent of the irregular temporal delay of the platform, while the TOF method is vulnerable to the delay. Furthermore, when the temporal resolution and audio bandwidth is limited, the Doppler shift analysis is expected to provide a better "resolution" for detection than pulse reflection analysis [1, 8]. In addition, since this method does not require any interval between the acoustic signals, the system is suitable for continuous detection, whereas the TOF method should wait for reverbs to decay.

3.2 Frequency Bandwidth

Similar to [6], the Sonicstrument also utilizes a non-ultrasonic, audible sound, but it also focuses on a different frequency range that is rarely perceived by most users. Typical acoustic transducers (e.g., everyday earphones and microphones) cover a frequency response range from around 20 [Hz] to above 22 [kHz] [5], thereby spanning the human audible range. Still, even within this range, there are frequency bands (both high (above 18 [kHz]) and low (below 60 [kHz])) that are practically inaudible for most people at usual loudness levels. These "marginal" areas can be utilized for motion detection with virtually no disturbance or overlap against the sonic "contents." At the same time, using these frequency ranges allows for easy implementation of the system, as described below.

4. SYSTEM DESIGN

4.1 Overview

This system uses a laptop (Dell Inspiron 1420) as its platform; it is equipped with an internal stereo microphone and outputs a maximum of 5.1 channel audio (a common feature with most personal computers these days), among which we use two

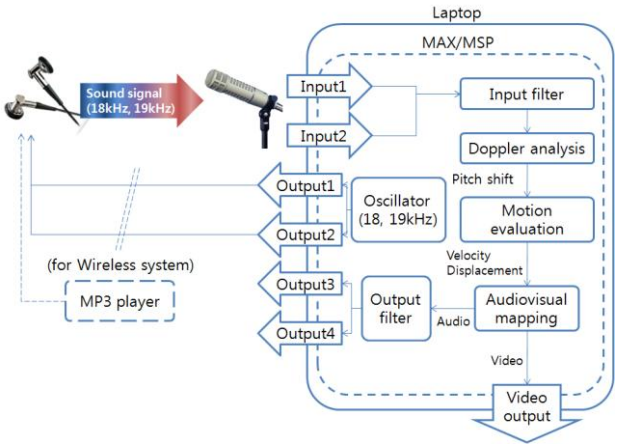

Figure 1. System block diagram of Sonicstrument system

channels for control signal output and other two for final sound output. The earphones that act as a controller are connected to the first two channels that transmit high frequency sound waves. Therefore, the system requires no extra device except for one computer. (For use on stages, the sound can be emitted from a commercial MP3 player, which can help in keeping the performer wire-free. This will be discussed later.) As the user moves his hands with the earphone, the Doppler shift occurs on the signal sound wave. The internal stereo microphone receives this distorted sound wave, while the MAX/MSP software analyzes the Doppler shift of the sound and calculates the hand motion in a normal direction for the microphone. Finally, the motion data are mapped to show visual and audio output.

This methodology is basically identical to the aforementioned ultrasonic sound tracking in mechanism (which is also used for human motion detection). However, this system uses stereotypical microphones that can handle the whole bandwidth of "theoretical" human hearing, which makes it necessary to use audio filters for noise elimination in most of the audible range. Furthermore, the final sound output from the system should not overlap the control signal bandwidth in order to have no interruptions.

4.2 Frequency of Control Signal

First of all, there are two choices of frequency. The frequency can be lower than the practical audible range or higher. A higher frequency range was chosen for two reasons. First, common room noise is distributed in a lower frequency range than a higher one. By using a higher frequency range, the system gets less influence from these noises. This higher noise-immunity is desirable for this system because it is more reliable when common earphones and microphones with lower precision are used. Second, a higher frequency yields a higher Doppler shift resolution, as the frequency changes more for the same hand velocity.

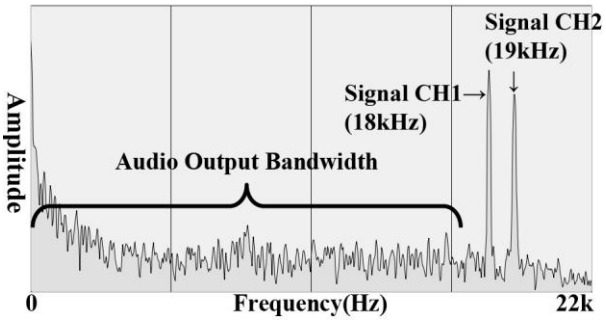

Figure 2. Spectrum distribution of control signal and audio output

Again, most of the commercial earphones cover a frequency range from around 20 [Hz] to around 22 [kHz]. Some kinds of earphones cover a narrower range around 20 [kHz]. This is the upper limit of frequency that can be used for this system. Also, the frequency range should not get too low to avoid being perceived by the user.
For these reasons, the sinusoids of 18 [kHz] and 19 [kHz] are used as the control signals for the left and right earphone respectively.

4.3 Signal Processing

4.3.1 Input Filter

The first step for the control signal that the microphones received is the band-pass filter. Because the Doppler shift is detected by a peak-tracking module, all of the noise peaks outside of the controller frequency need to be eliminated. To pass the two control signal ranges through while cutting off other ranges as much as possible, two narrow band-pass filters are connected in parallel.

4.3.2 Doppler Analysis Module

The signal that is able to pass through the filter is then sent to a Doppler analysis module where the 'fiddle~' object takes the biggest role. 'fiddle~' is a Max/MSP external object by Miller Puckette that tracks down multiple peaks and returns their frequency and amplitude in real-time. We already know the control signal is 18 [kHz] and 19 [kHz]. By comparing these reference frequencies to the detected peak frequencies, the motion velocity, which is a function of the frequency ratio, is evaluated. And the displacement is also numerically evaluated with this velocity data.

4.3.3 Output Filter

Finally, the evaluated motion data are connected to a proper visual or audio reaction. Before doing that, we have to consider that the audio output and control frequency are both in the audible range. To prevent the risk of any interruptions to the control signal, audio output should pass through a low-pass filter.

5. PERFORMANCE TEST

First, a qualitative test was carried out to see if the system can stably trace the motions of the handheld earphones (a demonstration video – titled as 'Video 1' – is available at [14]). The system successfully traced the motions of the earphone very smoothly and with no interference from ambient noises and between two control signals.

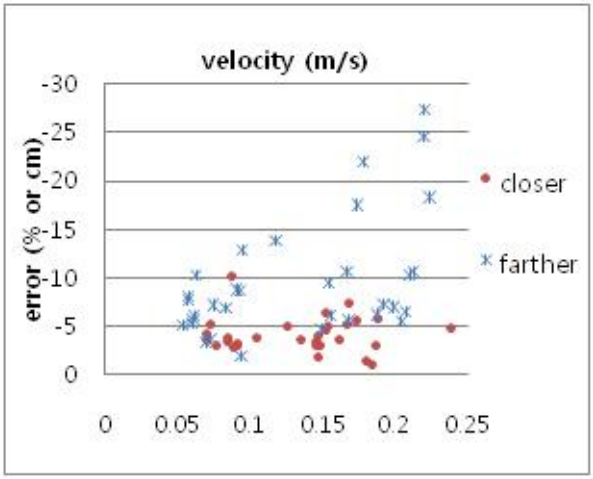

Figure 3. Error distribution of Sonicstrument from a performance test

In addition, a quantitative experiment was conducted to see whether the system can properly evaluate the displacement of the earphones. This was done by moving an earphone for 1 [m], and comparing the system-measured distance with real displacement. There were two criteria: direction (farther or closer) and the velocity of the motion. Datasets were collected thirty times for each direction, while the velocity values were randomly distributed.
Figure 3 shows the results; the mean and standard deviation of the measured values for each direction are -4.16 [cm] / 1.81 [cm] (closer) and -9.58 [cm] / 6.37 [cm] (farther). We can also see that the error distribution tends to increase when the direction is farther away and the velocity is faster. In our applications, the system could detect the motions of the user with reasonable precision and robustness. However, depending on the application, this error may not be neglegible and, since this aspect may be related to the frequency resolution, increasing the FFT window size may solve this problem and enhance the accuracy. Also, increasing the sampling rate would help reducing the latency due to the FFT and detecting more slight frequency shift in high frequency signal.

6. DEMONSTRATIONS

6.1 Musical Instrument

The first application of the system was on a computational musical instrument, which was exhibited at Anthracite, Seoul, Korea in 2010 (a demonstration video – Video 2 – is available at [14]). We took the violin as the metaphor for this motion-sound mapping; for a right-handed violinist, the left hand presses the strings to determine the pitch and the right hand bowing action excites the strings to generate the sound and control the volume. In our case, displacement of the earphone on the left hand from the microphone was mapped to the pitch (1 scale per 0.1 [m]) and the velocity of the earphone on the right hand corresponded to the audio gain.
As its platform, the system used the same laptop PC that we used for basic system implementation. We used two channels for control signal output, and two other channels for sound output.
The output sound was generated in real-time using subtractive synthesis. Multiple filters sculpted a pink noise to make a violin-like sound, while filter coefficients were manipulated to control pitch. Also, as mentioned in *4.3.3*, the output is filtered to prevent the interference with the control signal.

In this performance, the system successfully functioned as a musical instrument; pitch and gain values were controlled as the user intended. One problem, however, was the error accumulation of the estimated position of the left hand. To compensate for this, a reset function was implemented to be triggered when the gain from the right hand became large enough.

Figure 4. The first demonstration of Sonicstrument (musical interface & visualization)

Figure 5. The second demonstration of Sonicstrument (Interactive performance '4nm')

6.2 Interactive Performance

The second application of the system was at an interactive dance performance. As sonicstrument can work at a distance of about 10 [m] maximum, it can be utilized for most small-sized theater performances. In order to make the system wireless (which is critical for devices used in active dance performances), a portable music player was used to generate a control signal instead of PC; the music player was attached to the performer's body, and the earphone transmitters from the music player were held by the performer. Also, instead the internal microphone of the laptop in previous case, an external microphone was placed behind the curtain on the side of the stage to detect the control signals.

This system enabled us to detect the performer's hand motions or changes in body position. Measured control inputs were mapped to appropriate audiovisual stage effects; in a piece called *4nm*, the velocity data triggered a water flow sound and visual distortion effects.

Through this setup, sonicstrument showed its potential as an easy-to-use and highly effective interactive device for larger-scale performance. A video footage from this performance (Video 3) is also available at [14].

7. CONCLUSION

Sonicstrument is aimed at providing a virtual motion-detection interface without extra sensor devices. By using a sinusoid signal at an audible range, the commonly used earphones are utilized as a signal transmitter. The signal frequency is technically in the humans' audible range, but it is an extremely marginal area that most humans cannot perceive, thereby enabling continuous detection. As the system does not require any extra sensor device, individual users can easily own their tangible interface, and it can be simply applied to the performing arts.

This system uses a Doppler analysis instead of dominant TOF method. This is advantageous for reliability in a noisy environment. There is also a higher resolution that is not constrained directly by the system's time resolution and irregular time delay, while the TOF method has these restrictions. This is another good trait for popular interface.

In contrast, we were able to reconfirm the inherent limitation of the Doppler analysis in displacement estimation: error accumulation. Future work to compensate for this limitation can use the combination of the TOF and Doppler analysis method.

Also, for an environment with multiple microphones, like the laptop that we used that has 2 microphones, the direction-of-arrival (DOA) estimation technique for audible sound [2] can be adapted and it is expected to increase the degree of freedom in the system.

8. REFERENCES

[1] Amundson, I., Koutsoukos, X. and Sallai, J. Mobile Sensor Localization and Navigation using RF Doppler Shifts. In *Proc. the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, ACM Press (2008), 97-102.

[2] Chandran, S. 2006, Direction Estimation of Broadband Sources for Auditory Localization and Spatially Selective Listening, *Advances in Direction-of-Arrival Estimation*, Artech House, Boston, pp. 305-326.

[3] Dogaru, T., Le, C. and Kirose, G. Analysis of the Radar Doppler Signature of a Moving Human.

[4] Harter, A., Hopper, A., Steggles, P., Ward, A. and Webster, P. The Anatomy of a Context-Aware Application. *Wireless Networks 8, 2* (2002), 187-197.

[5] Headphone – Wikipedia http://en.wikipedia.org/wiki/Headphones

[6] Lopes, C.V., Haghighat, A., Mandal, A., Givargis, T. and Baldi, P. Localization of Off-the-Shelf Mobile Devices Using Audible Sound: Architectures, Protocols and Performance Assessment. In *Proc. SIGMOBILE 2006*, ACM Press (2006), 38-50.

[7] Motion Detection – Wikipedia. http://en.wikipedia.org/wiki/Motion_detection.

[8] Paradiso, J., Abler, C., Hsiao, K. and Reynolds, M. The Magic Carpet: Physical Sensing for Immersive Environments. *Ext. Abstracts CHI 1997*, ACM Press (1997), 277-278.

[9] Priyantha, N.B., Chakaborty, A. and Balakrishnan, H. The Cricket Location-Support System. In *Proc. MobiCom 2000*, ACM Press (2000), 32-43.

[10] Reynolds, M., Schoner, B., Richards, J., Dobson, K. and Gershenfeld, N. An Immersive, Multi-user, Musical Stage Environment. In *Proc. SIGGRAPH 2001*, ACM Press (2001), 553-560.

[11] Seniuk, A. and Blostein, D. Pen Acoustic Emissions for Text and Gesture Recognition. In *Proc. 10th International Conference on Document Analysis and Recognition*, IEEE (2009), 872-876.

[12] Vlasic, D., Adelsberger, R., Vannucci, G., Barnwell, J. and Markus G. Practical Motion Capture in Everyday Surroundings. In *Proc. SIGGRAPH 2007*, ACM Press (2007).

[13] Zhang, Z., Pouliquen, P.O., Waxman, A. and Andreou, A.G. Acoustic micro-Doppler radar for human gait imaging. In *Journal of the Acoustical Society of America Express Letters*, Vol. 121, No. 3(2007), pp. 110-113.

[14] Video clips of Sonicstrument demonstrations, http://aimlab.kaist.ac.kr/~badclown/Sonicstrument