Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Enhancing headphone reproduction of an electronic piano: Control of dynamic interaural level differences coupled with a player's active head movement

Kim, Sungyoung (1), Anandhi Ramesh (1), Masahiro Ikeda (1), and William L. Martens (2)

(1) Sound & IT Development Division, Yamaha Corporation, Hamamatsu, Japan
(2) Faculty of Architecture, University of Sydney, Sydney, Australia

PACS: 43.75.Tv

ABSTRACT

Modern electronic instruments not only provide improved tonalities but also allow players to select a monitoring method between loudspeakers (public) and headphones (private). Therefore, an ideal electronic instrument, such as an electronic piano, would require reproducing a perceptually similar sound for both reproductions. While the tonal quality remains relatively the same for both reproductions, a headphone-reproduced sound is distinctively different from that reproduced by a loudspeaker, primarily because the spatial coordinates of a headphone sound tends to synchronously follow a player's head movement. A system utilizing a motion track sensor might enable the headphone sound to remain steady. However, such a system faces several challenges, including the latency of processing and the timbre change. Here, we present the results and provide the details of a new method developed for reproducing piano sound via headphones; this method primarily adjusts the level of difference between the left and right headphone signals according to a player's horizontal head movement, i.e., yaw. For the level adjustment, the authors measured the interaural level differences (ILDs) of each key of a grand piano varying with the yaw angle. These ILDs enabled a headphone piano sound to rotate toward the opposite direction of head movement. Coupled with the motion tracking sensor attached to headphones, the proposed method could stabilize the headphone sound of a piano, regardless of the player's active movement during performance. A subsequent analysis revealed that the eighty eight sets of ILDs could be equivalently reduced to six subsets, by grouping adjacent keys that have similar ILDs. Further, the six sets of ILDs were fitted into six equations that parametrically represented the measured ILDs. A subsequent informal listening test on the proposed method showed that players could perceive steady, natural, and present piano imagery.

INTRODUCTION

Over the last couple of decades, signal processing technologies for electronic instruments have progressed significantly to allow the reproduction of authentic impressions of real instruments. For example, John Chowning proposed a method that recreates a musical instrument's idiosyncratic spectra using Frequency Modulation (FM) [Chowning (1973)] synthesis, and then a commercial synthesizer based on FM synthesis (Yamaha DX-7) was introduced. Later, a new method was introduced, which stores a real instrument's time domain waveform (sample) and manipulates it so as to reproduce the stored sample for the required duration with a post filtering process (Roland D-50 and Korg M-1). In principle, this technique is identical to digital recording, leading this sample-based synthesis to be called PCM (Pulse Code Modulation) synthesis or wavetable synthesis[1]. More recently, not only static snapshots of frequency or time domain characteristics of sound wave propagation of a real instrument, but also the dynamic behavior of excitation and radiation of an instrument and associated change have been modeled. This attempt is called *Physical Modeling Synthesis*. As Dodge and Jerse point out [(Dodge and Jerse 1997, p. 277)], this method is "considerably more intuitive" to musicians since it allows them to control the "bowing pressure" rather than the "index of modulation."

In addition to the enhanced tonality of the recent synthesis methods, a modern electronic instrument has various advantages such as the functionality to combine multiple layers of instrumental sounds for a new tone. Another advantage of an electronic instrument is enhanced control of its reproduced sound level. Players can reproduce their performance either via loudspeakers (internal or external) or via headphones. Typically, a player would practice his/her private performance while monitoring it via headphones, while a public performance is reproduced through a large sound reinforcement system. It is natural for a player to expect that a private monitor via headphones would let him/her anticipate the sound field as it would be heard in a live or public performance through loudspeakers. Therefore, an ideal electronic instrument would be able to reproduce a perceptually similar sound for both reproductions.

While the tonal quality remains relatively the same for both reproductions, a headphone-reproduced sound has its own distinct characteristics in terms of spatial attributes compared to a loudspeaker reproduction.

First, it is rare for an electronic instrument to equip a dedicated sound for headphone reproduction. In other words, the sound samples stored in an electronic instrument are meant to be reproduced via a pair of loudspeakers. Typically, this two-channel stereo signal is directly reproduced via headphones, resulting in a so-called *biphonic stereo* image (More detailed information on *biphonic stereo* is well summarized in [Marui and Martens (2006)]). One common drawback of such *biphonic*

[1] Even before the digital era, musicians of *Music concrète* created compositions consisting of recorded samples reproduced by a specially designed tape deck designed to replay tape loops, such as Phonogene[Manning (1985)].

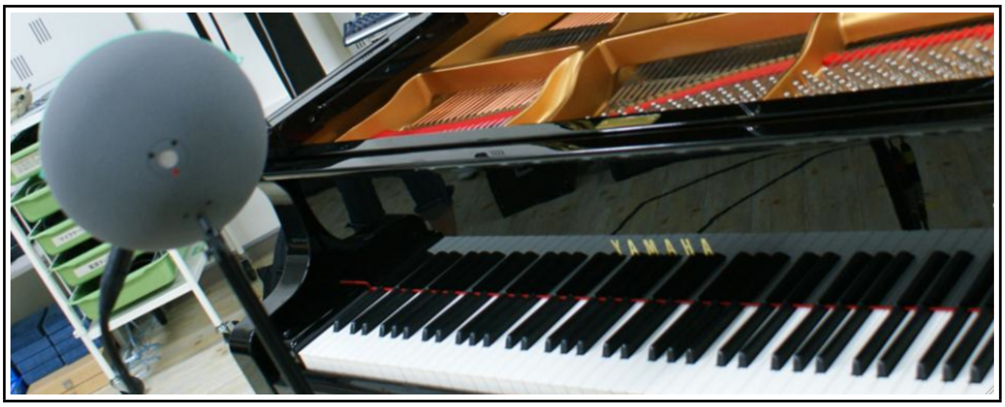

Figure 1: The sphere microphone, Schoeps KFM360, used to record pseudo-binaural signals for the given azimuth variation and each piano note automatically played by a MIDI controlled apparatus, DISKLAVIER. As pictured, the sphere microphone faces to +30° from the center position. In addition, three cardioid microphones were placed near the hammers to be used for multichannel reproduction of piano sound. For the current study, the center cardioid microphone was used as a reference signal between the two channels of the sphere microphone.

stereo is that the reproduced sound field tends to be heard as if it is reproduced inside the listener's head, referred to as IHL (inside-the-head localization). In order to prevent this IHL, many researchers have investigated methods to locate sound images outside from the listener's head, OHL (outside-the-head localization), for example, by synthesizing transaural crosstalk and reproducing it at the contralateral ear. It is true that such an externally located sound image gives a natural sound field. On the other hand, researchers have debated whether listeners care about IHL or OHL and want sound existing outside their heads when they are listening with a headphone[2].

Secondly, a spatial coordinate of headphone sound tends to follow a player's head movement. In other words, when the listener turns his or her head, the reproduced sound field turns to where the listener's head faces as well. This is especially problematic for the representation of a relatively large instrument - such as a grand piano. It is still true that visual cues dominate auditory cues; when a player watches a keyboard that is not moving, he/she believes that auditory imagery is also static even though it is not. However, many trained players easily notice asynchrony between visual and auditory cues and experience "unnaturalness" and "discomfort." In order to overcome this artifact, a system utilizing a motion track sensor is often used, which readjusts the spatial coordinates so that the sound field remains static regardless of listener's head movement. This type of system controls the perceived spatial coordination by imposing directional cues on the original sound using convolution with head related impulse response (HRIR) or binaural room impulse response (BRIR). While such a system creates a convincingly stable sound field, it faces other challenges, such as the high processing power which prevents real-time control and the timbre alteration. Recent works revealed that the complicated nature of HRIR could be parametrically represented [Iida et al. (2007)][Breebaart et al. (2010)]. With these methods, however, it is hard to obtain "acceptable" timbral quality and spatial stability simultaneously. In particular, maintaining authentic timbral character is much more important for any electronic instrument than enhancement of spatial attributes. Thus, a relatively simpler cue, such as Interaural Level Difference (ILD), Interaural Time Difference (ITD), or both, might be more adequate to dynamically control a spatial coordinate, corresponding to the player's head movement. This idea of applying relatively simple cues has been researched previously. The results showed that when head movement is coupled with simple ILD and ITD, "front-back confusions are easily disambiguated" and "these cues tend to dominate over" HRIR-based modifications [(AES Staff Writer 2007, p. 303)][Martens and Kim (2009)].

Based on these results, the authors hypothesized that appropriate control of ILD and ITD coupled with a listener's head motion would generate a convincing representation of a piano sound for headphone reproduction. This hypothesis was divided into two research questions:

1. What is the angular variation in azimuth associated with piano players' head movements?
2. Would it be possible to parametrically represent the ILD and ITD values associated with two variables, a player's horizontal head movement, i.e. yaw, and each piano note? Further, using such parametric representations of ILD and ITD, would it be possible to create a static piano sound field over a player's horizontal head movement?

METHOD

Measuring azimuthal variation in piano players' head movement during actual performance

In order to recreate a static electronic piano sound field over headphones, the authors first measured how much piano players rotated their heads, especially variations in yaw, during actual performance. For the measurement, we attached a motion tracking device, Polhemus ISOTRAK II, to a pair headphones and asked piano players to wear it during the performance of their preferred repertories. A total of five experienced piano players participated in this measurement. The analysis result showed that most variation in yaw occurred at the center and extended to the 40° limit on either side.

Measuring pseudo-ILD and ITD varying with head rotation

In order to extract parametric representations of ILD and ITD that vary according to yaw, we captured a pseudo-binaural signal of each piano note using a sphere microphone (SCHOEPS KFM360[3]) as shown in Fig. 1. The two pressure transducers built into each side (with a distance of 18 cm) of the spherical body of the microphone are meant to recreate interaural differences in terms of level and delay. In addition, a cardioid

[2] This issue was discussed in depth during the workshop titled *Binaural Technologies for Mobile Applications* at the 121st International Convention of AES, which is later summarized in [AES Staff Writer (2007)]

[3] Please refer *http://www.schoeps.de/en/products/kfm360/specs* for further information of the microphone.

Table 1: Parameters representing the whole gain structure of eighty-eight notes within the reduced six subsets. The first column indicates the subset number; the following two columns show the two corresponding polynomial functions that parametrically represent the gain variation of each subset according to yaw variation, y (from -40° to +40°). The last column shows the equivalent frequency bandwidth of each subset.

Subset	Note Number	Equation 1 (+yaw)	Equation 2 (-yaw)	Equivalent Fundamental Frequency Region (in Hz)
1	1 to 48	$0.001 \cdot y^2 + 0.1213 \cdot y$	$-0.0008 \cdot y^2 + 0.1298 \cdot y$	< 440
2	49 to 56	$0.0025 \cdot y^2 + 0.2367 \cdot y$	$-0.0014 \cdot y^2 + 0.2144 \cdot y$	440 to 700
3	57 to 62	$0.0059 \cdot y^2 + 0.4222 \cdot y$	$-0.0021 \cdot y^2 + 0.2924 \cdot y$	700 to 1000
4	63 to 72	$-0.0044 \cdot y^2 + 0.1502 \cdot y$	$-0.0014 \cdot y^2 + 0.2348 \cdot y$	1000 to 1760
5	73 to 78	$0.005 \cdot y^2 + 0.5041 \cdot y$	$-0.0046 \cdot y^2 + 0.3513 \cdot y$	1760 to 2500
6	79 to 88	$0.0027 \cdot y^2 + 0.4284 \cdot y$	$-0.0026 \cdot y^2 + 0.2882 \cdot y$	> 2500

microphone was used as a reference microphone to the sphere microphone. The sphere microphone was placed at the approximate location of the player's head in front of the piano and aligned such that the center of the microphone faced the piano's center where the reference microphone was. In the first position of the microphone, all eighty-eight piano notes were played, and then the microphone was rotated according to preset yaw values which were -70, -50, -40, -30, -25, -15, -13, -8, -5, -3, 0, +3, +5, +8, +13, +15, +25, +30, +40, +50, and +70° from the center. Here, the symbol + indicates the player's clockwise rotation, while - represents counter-clockwise rotation. These values were heuristically determined so as to give higher resolution nearer the center than at the fringes. While it has been experimentally shown that players do not move their heads over ±40°, angles of ±50° and ±70° were included in order to investigate whether or not these areas cause any significant difference in ILD and ITD.

In the measurement, it was important to capture "constant" playing of piano notes as much as possible. While a trained piano player could play a relatively constant performance, this experiment adopted a controlled method utilizing a MIDI controlled acoustic piano, YAMAHA DISKLAVIER. Using a sequencer, the piano was programmed to play all of its notes in sequence, with constant MIDI velocity (100) and duration (three seconds).

The recording was made at the studio B of Epicurus Studio located in Tokyo, Japan. The acoustical condition of the room is not anechoic, but rather is typical of what a normal piano player would experience in his or her practice or performance.

RESULTS

ILD Parameterization

The interaural level difference (ILD) was estimated in the following manner. The Root Mean Square (RMS) values of the three signals, the center reference microphone (denoted as C), the left side pressure capsule of the sphere microphone (L), and the right capsule (R), were first calculated. Then the level differences between C and L were calculated and used to represent the relative level change in the player's left ear. In the same way, the difference between C and R was used for the right ear's level change. In the top panel of Fig. 2, the calculated level change in dB varying with yaw and piano note at player's left ear is represented. The X axis is the note number, which is equivalent to a log frequency scale. The Y axis contains the yaw values from -90 to 90. For each combination of note number and yaw, the relative gain difference in dB is presented on the Z axis.

In order to reduce the number of gain values from all eighty-eight notes to a smaller but equivalent set, we attempted to group the adjacent keys that had similar ILDs. After several heuristic trials, a total of six subsets were determined. The subsets were created as follows - Subset 1 : Note 1 to 48, Subset 2 : Note 49 to 56, Subset 3 : Note 57 to 62, Subset 4 : Note 63 to 72, Subset 5 : Note 73 to 78, and Subset 6 : Note 79 to 88.

Subsequently, a representative gain relation for each subset was calculated using a quadratic polynomial fit. However, an attempt to fit the gain variation using a single polynomial function over entire ±40° failed (Based on the previously measured results for piano player's head movement, yaw variations outside the limit of ±40 were excluded for the parameterization). Therefore, another attempt was made to use two polynomial functions, one for yaw variation from 0 to +40°, and another for 0 to -40°, which was successful to parametrically represent the measured gain variation. Consequently, we devised a total of twelve polynomial equations that represented the gain variations of the associated six subsets. Table 1 shows the subset number, associated note numbers, two polynomial equations (for + yaw and - yaw, respectively) that parametrically represent the gain variation of each subset, and the frequency bandwidth of each subset, in a player's left ear. The gain relation of the opposite ear was simply calculated by multiplying -1 of the current equations due to the symmetric variation in two ear positions. The bottom of Fig. 2 shows the gain relations calculated by six representative equations (combination of two polynomial functions, as shown in the second and third columns of Table 1) for each subset.

ITD Parameterization

The interaural time difference (ITD) was estimated by the lag value between C and L calculated via cross-correlation between two, which indicated relative delay at a player's left ear. The same method was used to calculate the delay at the left ear associated with note and yaw variation. The measured result showed that time difference was strongly related to yaw variation, not to piano note. Therefore, it was decided to extract a global delay value which changes over yaw but remains constant over note. Equation 1 shows the equation used to represent the global time difference in milliseconds varying with yaw on the player's left ear. The variable y of this equation refers to the yaw variation of ±40° from the center.

$$ITD_{left} = -y \cdot 0.2564 \tag{1}$$

SUBJECT EVALUATION

Seven listeners who were either trained piano players or researchers working on the development of a new electronic piano participated in a pilot listening session. For the session, a prototype system dynamically controlled the ILD and ITD corresponding to yaw. The system used conventional stereo electronic piano samples for the headphone reproduction. In addition, the system measured the yaw variation using the previous motion tracker, Polhemus ISOTRAK II, which was attached to a pair of headphones. In the session, each listener compared

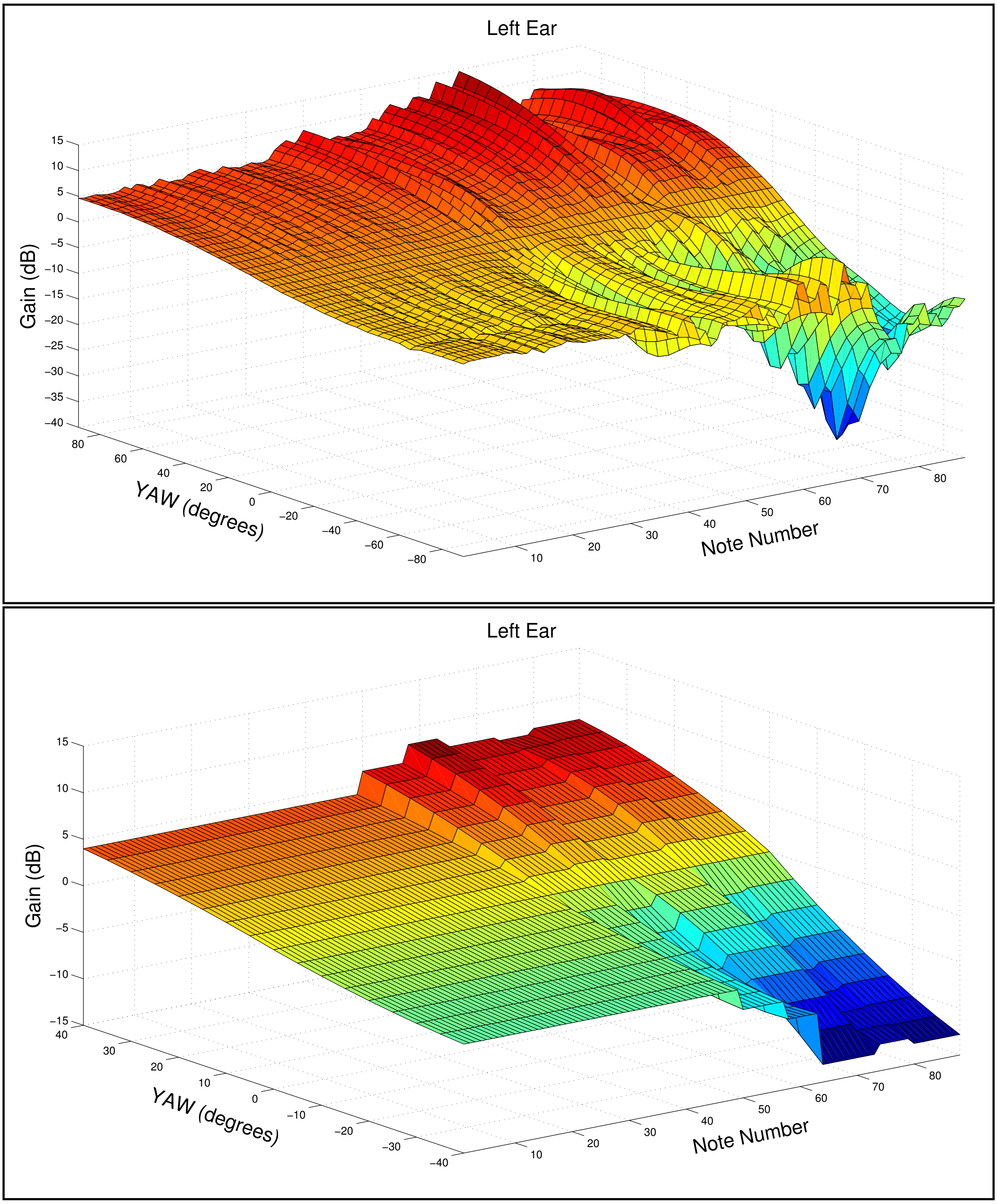

Figure 2: [Upper panel] The relative level difference between yaw angles and the reference position, 0°, for each piano note, measured at player's left ear position using a sphere microphone (SCHOEPS KFM360). The X axis represents the note number, which is equivalent to a log frequency scale. The Y axis contains the yaw values from -90 to 90°. For each combination of note and yaw, the relative level difference (gain) in dB is presented in the Z axis. [Lower panel] The level difference represented with six subsets of the adjacent keys that had similar ILDs. The slope of gain variation of each subset was represented with a combination of two second dimensional polynomial functions as shown in the second and third columns of Table 1. In this figure, the yaw variation (Y axis) was limited to ±40°.

a normal condition (a piano sound without head tracking) with the currently proposed method. At their first comparisons, some listeners did not notice the difference (because the two conditions would create the same sound fields without head rotation). After continuous playing for two or three minutes, however, most players found that their perceived image was different and reported that the new method created steady and natural piano sound over a headphone. Another observation was that applying ITD for entire notes had created an audible tonal change when a player moved his or her head relatively quickly. In contrast, when the ITD cue was excluded, with only ILD-based manipulation, the reproduced sound field caused low notes to be less steady on a player's head movement (following a player's head movement). Therefore, it was decided to apply the ITD control only for low notes (the subset 1, from note 1 to 48). As a result, this modified method, using control of ILD for all notes, with additional control of ITD for low notes, delivered enhanced sense of spatial perception to participating listeners.

DISCUSSION AND FUTURE WORKS

While the proposed method allowed players to experience a steady piano sound field regardless of their active head movement and, consequently, externalized piano sound over headphones, it is still true that players would experience IHL of piano sound when they do not move their heads. Thus, additional efforts should be made to devise a method to convert a biphonic representation (direct playback of conventional stereo piano sound via a headphone) into an externalized piano sound. With such an externalized piano sound, the proposed ILD manipulation coupled with players' head movement would create a higher degree of realism and feeling of presence. However, as previously stated, the externalization process should not affect the perceived timbral quality, which is challenging. The authors are currently investigating possible methods that would create an externalized spatial imagery with perceptually transparent timbre of headphone piano sound.

In the meantime, a formal and controlled subjective evaluation should be followed. The authors are currently preparing this evaluation using piano players to reveal their preferences of headphone sound reproduction as well as to identify salient attributes characterizing the proposed method compared to the normal biphonic reproduction.

CONCLUSION

This paper proposed a method that re-orients a headphone reproduced electronic piano sound coupled with a player's head movement, resulting in a spatially steady sound field. In order to provide relevant cues that control perceived direction of reproduced piano sound, the authors measured a set of pseudo interaural level differences (ILD) and time differences (ITD) created by each piano note at various angles in yaw, i.e., horizontal rotation. Subsequent analysis of those measured ILD and ITD data showed that ILD could be parametrically represented by six equations, each of which consists of a pair of two second order polynomial functions. On the other hand, ITD variation was modeled with a single equation across all piano notes. Coupled with the motion tracking device, the extracted parameters delivered an impression of naturalness and feeling of presence to players without degrading the timbre of reproduced piano sound.

REFERENCES

AES Staff Writer. Audio for Mobile and Handheld Devices. *J. Audio Eng. Soc.*, 55(4):301 – 305, April 2007.

Jeroen Breebaart, Fabian Nater, and Armin Kohlrausch. Spectral and Spatial Parameter Resolution Requirements for Parametric, Filter-Bank-Based HRTF Processing. *J. Audio Eng. Soc.*, 58(3):126 – 140, March 2010.

John Chowning. The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. *J. Audio Eng. Soc.*, 21(7): 526 – 534, September 1973.

Charles Dodge and Thosmas A. Jerse. *Computer Music: Synthesis, Composition, and Performance*. Schirmer, Thomson Learning, 2nd edition, 1997.

Kazuhiro Iida, Motokuni Itoh, Atsue Itagaki, and Masayuki Morimoto. Median plane localization using parametric model of the head-related transfer function based on spectral cues. *Applied Acoustics*, 68:835 – 850, 2007.

Peter Manning. *Electronic & Computer Music*. Oxford University Press, New York, 2^{nd} edition, 1985.

William L. Martens and Sungyoung Kim. Dominance of head-motion-coupled directional cues over other cues during active localization using a binaural hearing instrument. In *The* 10^{th} *Western Pacific Acoustic Conference*. WESPAC10, 2009.

Atsushi Marui and William L. Martens. Spatial Character and Quality Assessment of Selected Stereophonic Image Enhancements for Headphone Playback of Popular Music. In *Proc. Audio Engineering Society 120*th *Int. Conv.* AES, May 2006.