SPATIALNOTES: BINAURAL MIDI MUSIC OF DIRECTIONAL NOTES

ZHENG Zhi

Tsinghua University

ABSTRACT

The existing binaural music mainly separates different instruments, which is reproduction of music listening in reality. We propose SpatialNotes, with notes appearing at different directions, providing listeners with novel listening experience. This also helps listeners identify the pitch of notes and feel the change of melody. We synthesize music from MIDI files and apply BRIR to it for binaural effects. The output music pieces are unique and fair-sounding.

***Index Terms*—** Binaural, MIDI, HRTF, BRIR, music

1. INTRODUCTION

When listening to music and enjoying the melody, people would sense the changes of pitch. However, the ability of such sensation varies among people. Those who have good sensation of pitch, usually having received long-term musical training, can easily capture the changes and picture the notes like a music score. Let alone the small portion of people who have sense of absolute pitch, can tell the exact pitch of notes and even replay the melody with an instrument. While for ordinary people like us authors, they only have a blurry feeling of the pitch and have to take a while to figure out where the melody is going.

Those ordinary people who are interested in music, although not good at sensing the pitch of melody, may want to get better understanding of some musical compositions. To this end, they can look at the scores or some visualizations. But this is not convenient because people don't want their eyes occupied when listening to music. What if we can tell the pitch and get a clearer picture of the music only by listening?

We propose SpatialNotes, a novel form of binaural music with notes appearing at different direction. It takes advantage of binaural audio and people's ability of sound localization. According to 12-tone equal temperament, we place notes into 12 direction, with C in the front and pitch going up clockwise every 30 degree. Even if people have little sense of pitch, they can recognise the direction of notes, thus tell the changes of pitch and even the absolute pitch, and then have a better understanding of the musical composition. Potentially, after long-term listening of such binaural music, people might naturally gain better sensation of pitch.

On the other hand, binaural songs nowadays (usually known as Apple Spatial Audio, Dolby Atmos etc.) mainly separate different instruments into various directions, for example, vocal in the front, drums in the left, piano in the right etc. Although this is a better presentation of music compared to traditional stereo music, it is still reproduction of the listening experience in reality, without exploiting the potential of binaural audio sufficiently. If we view that as listeners being surrounded by instruments, the SpatialNotes would be listeners being surrounded by notes and melody, and being immersed in the music itself. which would be a fresh and unique listening experience.

In the rest of this paper, we will explain some basic concepts and principles of binaural audio. Then we will discuss different approaches to our goal. Finally we will talk about the limitations and future work.

2. CONCEPTS AND PRINCIPLES

2.1. Sound Localization

Sound localization is a listener's ability to identify the location or origin of a detected sound in direction and distance. For human, the mechanisms have been well studied. There are several cues for human's binaural auditory system to localize a sound, including interaural time difference (ITD), interaural level difference (ILD) and spectral information. The spectral information, explained by pinna filtering effect [1], is mainly due to the shape of pinna and head, which can provide clues of front back up or down only with one ear.

2.2. Externalization and Internalization

Except for sound localization, there is another key issue in sound recording, playing and perceiving: externalization and internalization [2]. The natural sounds originating from the environment, are perceived to be outside of the head, which is so-called external. While the sounds played by headphones or earphones, are likely perceived to be inside the head, which is so-called internal. The internal sounds are unnatural and people in many cases want to make sounds external. However, this effect is not only about physical acoustics, but also

Codes at https://github.com/ZhengZhiGeorge/SpatialNotes
Examples at https://www.bilibili.com/video/BV1BD4y1L7uU

psycho-acoustics, which means the perception can be affected by other aspects.

There are several cues for externalization and internalization:

- **Reverberation**: The natural sounds in the environment such as a room would have reverberation and reflection except for the direct sounds. While the sounds including most music recorded in a studio covered by sound absorbent materials have no reverberation and therefore are more likely to be internal[2, 3].

- **Spectral information**: The natural sounds in the environment have to travel by head and pinna to arrive at ear drum, thus have their frequency spectrum affected correspondingly. While the sounds recorded by ordinary microphones (such as omnidirectional microphones) and played by headphones wouldn't be affected by pinna. In contrast, sounds recorded by binaural microphones or dummy heads would have the same spectral features as external sounds[2].

- **Azimuth**: Internal sounds are perceived to be in the middle. So the sounds which have big ILD and perceived to be lateral are more likely to be external [2, 3].

- **Head movement**: When people move or rotate their heads, the natural sounds from the environment would be perceived to change accordingly, while sounds played by headphones would remain the same. This contributes much to the extinction of externalization and internalization [4].

- **Vision**: The natural sound sources are some actual objects in the environment, and people would psychologically mount sound sources with the objects. When there are objects that can be viewed as sound sources, the sounds are more likely to be perceived as external[2, 3].

2.3. HRTF

Head related impulse response (HRIR) describes how a direct sound travels by head and pinna and arrives at ear drum in time domain. While head related transfer function (HRTF) is the same signal in frequency domain. HRTF can be applied to a sound by simply convolving the sound signal and HRIR (which is the same with all kinds of IR). In fact, a pair of HRTFs contains all ITD, ILD and spectral information needed by sound localization. There are three parameters to define HRTF: azimuth, elevation and radius. HRTF varies a little from person to person. Besides, HRTF is usually recorded in a studio with absorbent materials as a free field, by dummy head or real person. But it also can be simulated computationally [5].

2.4. RIR

Room impulse response (RIR) describes the acoustic features of a room with a speaker and listener in certain positions. Unlike HRTF, RIR also contains information of reflection and reverberation. As the reflection and reverberation come from various directions, RIR can't be fully represented by traditional one or two channel audio. As for this, Ambisonics format is often used to record RIR, with no less than four microphones pointing at different directions, from which we can calculate sounds coming from any directions.

2.5. BRIR

Binaural room impulse response (BRIR) describes how sounds travel to people's ear drum in a given room and can be seen as a combination of HRTF and RIR. It varies for different positions of the speaker, and positions and orientations of the listener. BRIR is usually recorded by a dummy head in a room. Because of the lack of affordability of deploying speaker array given positions of speaker and listener, and the big number of combinations of speaker and listener, BRIR datasets are usually less complete than HRTF datasets (less choices of elevation). BRIR can also be calculated by combining HRTF and RIR of all directions, or simulated in simplified conditions [6].

3. APPROACHES

3.1. Audio separation or MIDI

We want to remix existing musical compositions to binaural SpatialNotes form.

One solution is audio separation, i.e. to separate different notes from music audio, no matter it is of one instrument, multiple instruments or vocal. With this method, we would have easy access of all music resources. However, the complexity of music audio may cause difficulty, for example, multiple notes can be on simultaneously like chords, and acoustic features of different instruments are various. Thus even the tiniest uncontrollable modifications on audio caused by the deficits above can lead to terrible listening experience.

The other solution is to synthesize music from MIDI file. MIDI files have almost all information of musical composition playing, from which music can be synthesized with soundfonts files. MIDI has some disadvantages, for example, MIDI files are not so accessible as audio, vocal cannot be synthesized and MIDI lack details of playing especially for some instruments such as violin, but the synthesized music is clean. So we finally choose this approach.

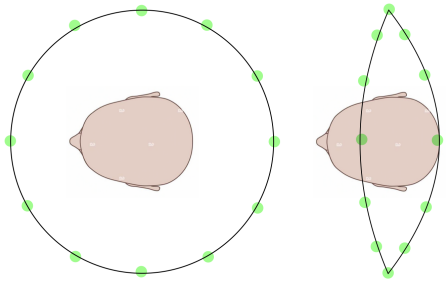
3.2. Selection of IR

In order to get binaural effect, we should apply one of the IRs (HRIR, BRIR) to the audio. The goals are clear localization

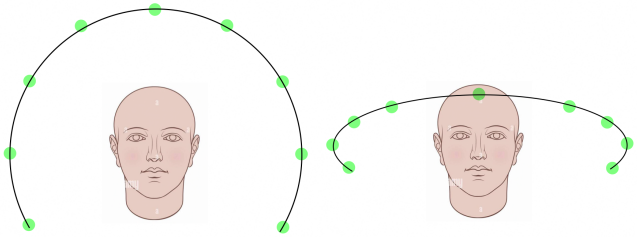
of notes and good musical quality. After careful experiments and comparing, we finally choose one BRIR dataset recorded in a concert hall [7], which has great binaural effect and reverberation. The following are the comparison of different approaches of IR.

3.2.1. HRTF

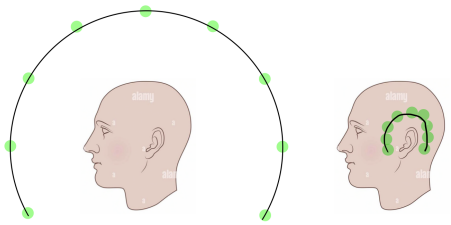
Many HRTF datasets have been tried, including datasets from different countries and of different races, such as MIT KEMAR[8], IRCAM LISTEN, Nagoya, KAIST[9] etc. The binaural effect is dissatisfying, that notes around the median plane tend to be internal. This is mainly because the soundfonts and HRTF are recorded too clean without reverberation etc., thus perceived to be unnatural and internal. The expected and actual binaural effects in three planes are showing in figure 1:



(a) horizontal



(b) frontal



(c) median

Fig. 1. Comparison of expected and actual binaural effects in three planes. Green dots stand for the positions of note

3.2.2. BRIR

We tried several BRIR datasets[7, 10], and the binaural effect is much better than using HRTF. The best of datasets is recorded in a concert hall and has great reverberation [7]. Multiple positions of listener are provided, which have different effects. The position closest to the stage has clearer hearing, while further positions have stronger reverberation. The overall hearing performance of this BRIR set is brilliant and we can choose different position for different musical composition according to its speed, instruments and emotion etc.

3.2.3. RIR + HRTF

Combining RIR and HRTF, an all direction BRIR set can be calculated. The biggest advantage is, with a complete BRIR set, the binaural audio can dynamically change when people rotating their heads, as if the sounds originate from the environment and are strongly external, very alike the experience of VR. However, this dynamic effect has high requirements, that the audio should have more channels other than 2, the music player is able to render the music real-time, and the headphone is capable of head tracking, which is not so affordable for most consumers. Besides, we didn't find an RIR set good enough as the concert hall, and we don't need audio to change when head rotating (e.g. we want to keep C in the front). So finally we didn't choose this solution.

4. LIMITATIONS AND FUTURE WORK

The overall listening experience with SpatialNotes is good and unique, but there are several aspects short of expectations. The localization of notes is not as precise and easy as expected, that listeners might fail to concentrate on the identification of notes when tempo gets fast. Besides, due to the incompleteness of the BRIR dataset, notes are restricted on the horizontal plane without variance of elevation, thus leaving room for improvement of artistic performance. In addition, using MIDI to synthesize music influences the quality of music expressing and restricts the music resources that we can utilize.

As for other projects or researches, binaural audio is a promising topic, as it is closely related to VR. How to realize virtual listening experience and embed it with vision are critical for creating immersive metaverse. Therefore it is worth attention of HCI researchers.

5. REFERENCES

[1] Dwight W Batteau, "The role of the pinna in human localization," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 168, no. 1011, pp. 158–180, 1967.

[2] Virginia Best, Robert Baumgartner, Mathieu Lavandier, Piotr Majdak, and Norbert Kopčo, "Sound externalization: A review of recent research," *Trends in Hearing*, vol. 24, pp. 2331216520948390, 2020.

[3] Thibaud Leclère, Mathieu Lavandier, and Fabien Perrin, "On the externalization of sound sources with headphones without reference to a real source," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2309–2320, 2019.

[4] W Owen Brimijoin, Alan W Boyd, and Michael A Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PloS one*, vol. 8, no. 12, pp. e83068, 2013.

[5] Parham Mokhtari, Hironori Takemoto, Ryouichi Nishimura, and Hiroaki Kato, "Comparison of simulated and measured hrtfs: Fdtd simulation using mri head data," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[6] Sanjeev Mehrotra, Wei-ge Chen, and Zhengyou Zhang, "Interpolation of combined head and room impulse response for audio spatialization," in *2011 IEEE 13th International Workshop on Multimedia Signal Processing*. IEEE, 2011, pp. 1–6.

[7] Bogdan Ioan Bacila and Hyunkook Lee, "360° Binaural Room Impulse Response (BRIR) Database for 6DOF spatial perception research," Mar. 2019.

[8] William G Gardner and Keith D Martin, "Hrtf measurements of a kemar," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[9] Gyeong-Tae Lee, Sang-Min Choi, Byeong-Yun Ko, and Yong-Hwa Park, "Hrtf measurement for accurate sound localization cues," 2022.

[10] Jon Francombe, "Iosr listening room multichannel brir dataset," 2020.