MUSIC STAGING AI

Kenta Niwa[1], *Kento Ohtani*[2], *and Kazuya Takeda*[2]

[1]: NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]: Graduate School of Information Sceience, Nagoya University, Japan

ABSTRACT

Through smartphones, user enables to download/listen music anytime and anywhere. As a concept of a future audio player, we propose a framework of ”music staging artificial intelligence (AI)”. In that framework, audio object signals, e.g. vocal, guitar, bass, drums and keyboards, are assumed to be extracted from stereo music signals. To visualize music as if live performance is virtually conducted, playing motion sequence is estimated by using separated signals. After adjusting the spatial arrangement of audio objects so as to each user prefers it, audio/visual rendering is conducted. We constructed two types of demonstration systems for music staging AI. In the smartphone-based implementation, each user enables to change the spatial arrangement through sliderbar dragging. Since information of user preferable spatial arrangement can be sent from each smartphone to server, it would enable to predict/recommend the user preferable spatial arrangement. In another implementation, head mount display (HMD) was utilized to dive into virtual music live performance. Each user enables to walk/teleport anywhere and audio is then changing corresponding to the user view.

1. CONCEPT OF OUR PROPOSAL

Source separation technology would be a main research topic in ICASSP. Many researchers including us have studied on source separation technologies. However, the issues on how to use separated signals may have not been discussed as a main topic in many studies. In our previous studies [1, 2], we have tried to generate binaural sounds corresponding to the user view. In this demonstration, we introduce our recent project named ”music staging AI”, as a new proposal of how to use separated signals.

In the framework of music staging AI, the stereo music $\mathbf{x}_t$ is decomposed into separated signals $\mathbf{y}_t$ [3], which are composed of five kinds of audio objects (vocal, guitar, bass, drums and keyboards). By using separated signals, playing motion sequences $\mathbf{d}_t$ are estimated for each audio object. This is conducted to generate visuals so that each avatar performs the musical instrument corresponding to the separated signal. Each user listens down-mixed stereo signals $\mathbf{z}_t$ where audio/visual rendering is conducted given the rendering parameter $\mathbf{p}$, which is composed of the position of a listener, viewer and audio objects (avatars).

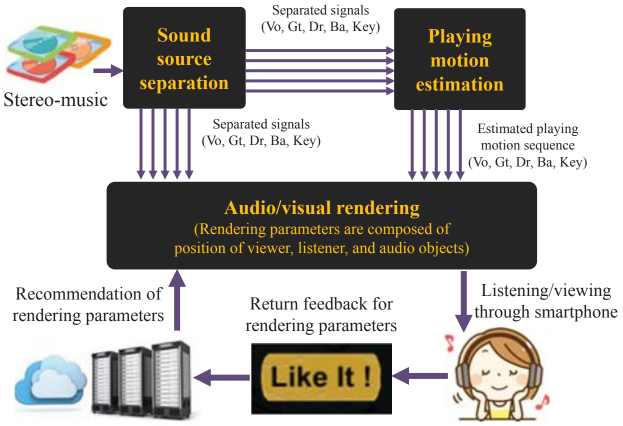
Stereo-music
Sound source separation
Separated signals (Vo, Gt, Dr, Ba, Key)
Playing motion estimation
Separated signals (Vo, Gt, Dr, Ba, Key)
Estimated playing motion sequence (Vo, Gt, Dr, Ba, Key)
Audio/visual rendering
(Rendering parameters are composed of position of viewer, listener, and audio objects)
Recommendation of rendering parameters
Return feedback for rendering parameters
Listening/viewing through smartphone
Like It !

Fig. 1. Concept of demo 1

2. DEMONSTRATION SYSTEMS

We have implemented two kinds of demonstration systems. In the demonstration 1 (demo 1), we assumed that each user uses music staging AI through a smartphone. In the demonstration 2 (demo 2), user enables to dive into virtually visualized live performance through an HMD.

2.1. Demo1: smartphone-based implementation

Figure 1 shows the concept of demo 1. From stereo music, we estimated five kinds of separated signals and corresponding playing motion sequences. Given the rendering parameters $\mathbf{p}$, audio/visual rendering is conducted in real-time. As an audio/visual rendering tool, game-engine (Unity 5) is utilized. How to select $\mathbf{p}$ so that user prefers it may be a serious problem in this system because the number of choices is huge. In our previous work [4], we proposed a method for selecting $J(\approx 10)$ kinds of rendering parameters whose auditory localizations are different each other. In the constructed application, it is possible to collect information of user preferable rendering parameters in the server. By analyzing collected information, we try to recommend personalized preferable rendering parameters.

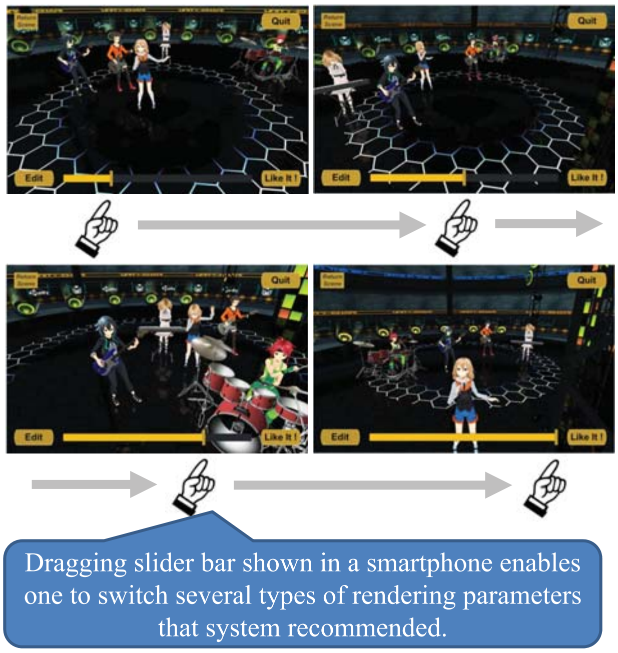
Quit
Edit
Like It!
Dragging slider bar shown in a smartphone enables one to switch several types of rendering parameters that system recommended.

Fig. 2. View image of demo 1 (smartphone-based implementation). Unity-chan stage: ©Unity Technologies Japan/UCL, characters: ©Gemba Inc. Taiki Mitsuishi, VR system construction: NTT-AT, Kouki Mitsuishi

Figure 2 shows the user viewing images represented on a smartphone. By dragging a slider bar, it is enable to switch the system recommended J kinds of rendering parameters. By pushing ”Like” button, the information of user preferable rendering parameter is sent to the server.

2.2. Demo2: HMD-based implementation

As another implementation of music staging AI, we constructed a system that user enables to dive into virtually visualized live performance. Similar to the demo 1, five kinds of separated signals were estimated from stereo music signal. They are virtually placed at the pre-defined avatar’s position. Figure 3 shows the user wearing an HMD (HTC VIVE) and user viewing image examples. By wearing an HMD and a headphone, user enables to walk and teleport to an arbitrary place by using user-holding controller. Corresponding to the user position, audio/visual are changed.

3. REQUIREMENTS FOR SPACE/EQUIPMENTS

To perform demonstrations, wide area (2.5 m × 2.5 m) are needed as shown in Fig. 4. To avoid conflicting persons, we would like to surround the demonstration area by tables. Since some PCs were used in demonstration, AC outlets (100 V, 800 W) are required.

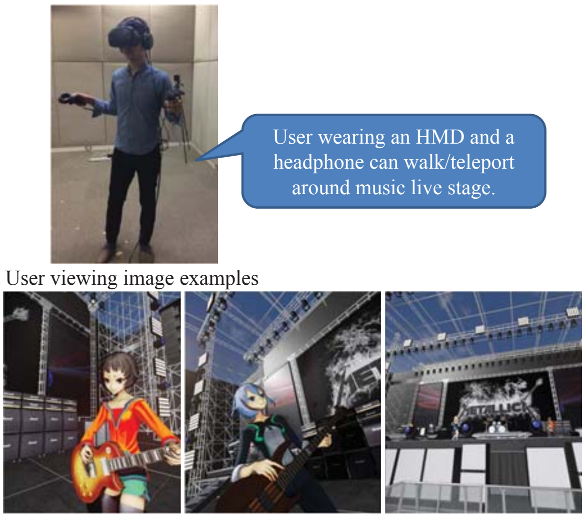
User wearing an HMD and a headphone can walk/teleport around music live stage.
User viewing image examples

Fig. 3. View image of demo 2 (HMD-based implementation) . characters: ©Gemba Inc. Taiki Mitsuishi, VR system construction: NTT-AT, Kouki Mitsuishi

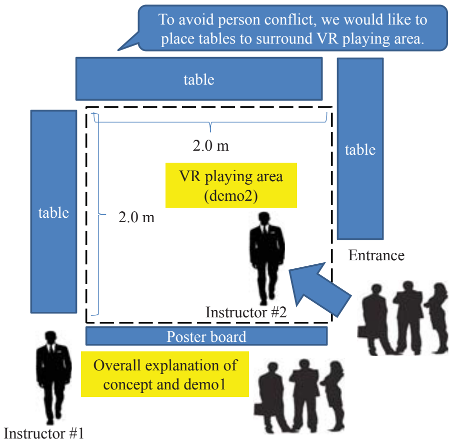
To avoid person conflict, we would like to place tables to surround VR playing area.
table
2.0 m
VR playing area (demo2)
table
table
2.0 m
Entrance
Instructor #2
Poster board
Overall explanation of concept and demo1
Instructor #1

Fig. 4. Placement of demonstration equipments

4. CONTACT INFORMATION

Kenta Niwa, NTT Media Intelligence Laboratories
E-mail: niwa.kenta@lab.ntt.co.jp
Phone number: +81 422 59 7026

5. REFERENCES

[1] K. Niwa, T. Nishino and K. Takeda, “Encoding large array signals into a 3D sound field representation for selective listening point audio based on blind source separation”, *ICASSP 2008*, pp.181–184, 2008.

[2] K. Niwa, Y. Koizumi, K. Kobayashi and H. Uematsu, “Binaural sound generation corresponding to omnidirectional video view using angular region-wise source enhancement”, *ICASSP 2016*, pp. 2852–2856 2016.

[3] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi and Y. Hioka, “Supervised source enhancement composed of nonnegative auto-encoders and complementarity subtraction”, *ICASSP 2017* (accepted).

[4] K. Ohtani, K. Niwa and K. Takeda, “Single Dimensional Control of Spatial Audio Object Arrangement”, *12th western pacific acoustics conference (WESPAC) 2015*, 2015