Learning-Based Jam Session System for A Guitar Trio

Masatoshi Hamanaka*, Masataka Goto**,***, and Nobuyuki Otsu***,*
*University of Tsukuba, ** "Information and Human Activity," PRESTO, JST
***National Institute of Advanced Industrial Science and Technology (AIST)
m.hamanaka@aist.go.jp

Abstract

This paper describes a jam session system that enables a human player to interplay with virtual players, each of which imitates musical reactions of a human player. Previous session systems have parameters for altering a way of reacting but have not been able to imitate such reactions. Our system can obtain the reaction model of a human player—that is, the characteristic way that player reacts to the other players—by learning the relationship between MIDI data of music the player listens to and MIDI data of music improvised by the player. Experimental results show that the reaction model of any player participating in a guitar trio session can be learned from the MIDI recording of that session.

1 Introduction

This study aims to make a jam session system in which virtual players react as if they were actual human players with various characters. We want to make it possible for human players to interact whenever they like with a virtual player imitating anyone they want to perform with, a familiar, professional, or deceased player... even themselves. What is most important in imitating players is to acquire the *reaction model* of the target human player—that is, to acquire a model of the characteristic manner in which that target player reacts to the performances of other players. The imitating virtual player can then improvise according to this reaction model.

Previous session systems have not been able to imitate a human player's reaction. Some systems [1][2] concentrate on following the performance of a human soloist without considering the individual character of the virtual player. Although JASPER [3] has a set of rules that determine the system reactions and VirJa Session [4] has parameters for altering a way of reacting, those systems cannot learn the reaction model of an actual player.

Our jam session system makes it possible for the reaction model of a target human player to be acquired from a MIDI recording of a session in which that player participated. The system statistically learns the relationship between the MIDI data of the music the target player listens to and MIDI data of the music improvised by that player. In other words, the system learns the relationship between the input and the output of the target player. The main advantage of this approach is that it is not necessary to examine the target player directly: all we need to build the model is a recording of a session in which that player participated.

2 Learning-Based Session System

Our system deals with constant-tempo 12-bar blues performed by a guitar trio consisting of a human guitarist and two virtual guitarists. The three players take the solo part one after another without a fixed leader-follower relationship. We chose the guitar trio configuration because we can use the performance of any player in MIDI recordings when learning the reaction model.

The system has two session modes, a learning mode and a session mode. In the learning mode (in Section 2.1), the system acquires reaction models in non-real time. These models are stored in a database and different reaction models can be assigned to the two virtual players before the session play. In the session mode (Section 2.2), a human player can interact with the virtual players in real time. As shown in Figure 1, each virtual player listens to the performances of all the players, including its own, and uses the reaction model to determine what its next reaction (output performance) will be.

2.1 Learning Mode

To create a virtual player that reacts as the actual human player does, it is necessary to acquire the actual player's individual reaction model. The main issue in acquiring the reaction model is to learn the relationship between the input and the output of the target player in MIDI recordings. This can be formulated as a problem of obtaining the mapping from the input to the target player's output. The direct

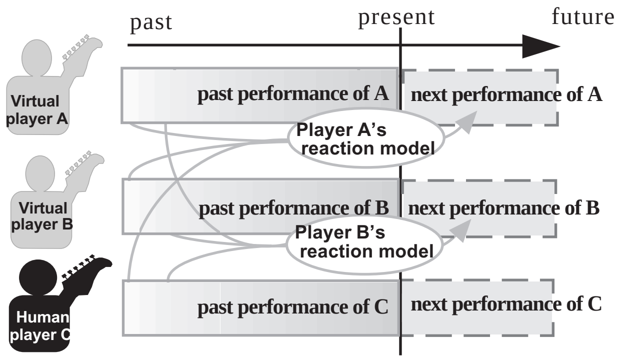


Figure 1: Player's reaction model.

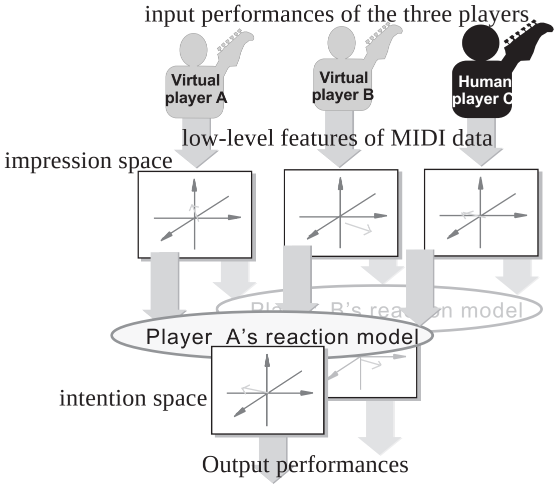

input performances of the three players
Virtual player A
Virtual player B
Human player C
low-level features of MIDI data
impression space
Player B's reaction model
Player A's reaction model
intention space
Output performances

Figure 2: Player architecture.

MIDI-level learning of this mapping, however, is too difficult because the same MIDI-level situation rarely occurs more than once and the mapping itself is too sparse. We therefore introduce two intermediate subjective spaces: an *impression space* and an *intention space* (Figure 2) .

1) Impression space

The impression space represents the subjective impression derived from the MIDI input. By applying principal component analysis (PCA) to the results of subjective evaluations of various MIDI performances, we determined three coordinate axes of the impression space. PCA is a statistical method for reducing the number of dimensions while capturing the major variance in a large data set. While listening the performance, a subject evaluated 10 impression words subjectively by rating them on a scale of one to seven.

The three selected axes of the impression space represent qualities that can be described as *appealing*, *energetic*, and *heavy*. To obtain a vector in this space, an *impression vector* corresponding to the MIDI input, we use canonical correlation analysis (CCA) . This analysis maximizes the correlation between various low-level features of the MIDI input (such as pitch, counts of notes, tensions, and pitch bend) and the corresponding subjective evaluation. Since an impression vector is obtained from each player's performance, we have at every moment three impression vectors (Figure 2) .

The impression space is necessary for learning the relationship between various input performances and the corresponding output performances. If we represent the input performances as short MIDI segments without using the impression space, the same MIDI segments will not be repeated in a different session. The impression space enables to abstract subjective impressions from input MIDI data and those impressions can be repeated. Even if two segments of the input MIDI data differ, they can be represented as a similar vector in the impression space as long as they give the same impression.

Figure 3 shows the transition of the rated values for the impression word "*appealing*." The black line represents the value calculated by the system and the gray line represents the value evaluated by a human listener. For 92 percent of the performance, the calculated and subjectively evaluated values do not differ by more than 1.

2) Intention space

The intention space represents the intention of the player improvising the output. A vector in this space, an *intention vector*, determines the feeling of the next output. It is used to select short MIDI phrases from a phrase database, and the output MIDI performance is generated by connecting the selected phrases.

Without the intention space, it is difficult to learn

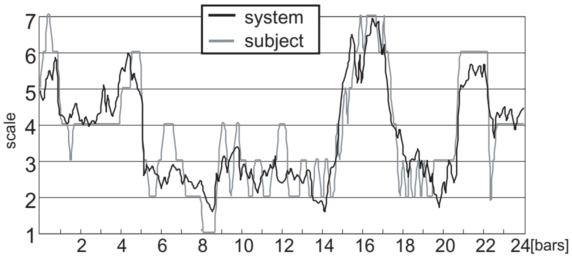


Figure 3: Transition of the rated values calculated by the system and evaluated by a subject (for the impression word "appealing").

the relationship between impression vectors and output MIDI data because in actual MIDI recordings various output can occur when the input data gives the same impression. The intention space makes it easier to learn the player's reaction model.

The intention space is constructed by using multidimensional scaling (MDS) [5], such that intention vectors are distributed with proximities proportional to subjective similarities of short phrases corresponding to those vectors. The dimensions of this space were determined from the MDS results to be three.

Because the number of the short phrases is limited, those phrases are sparsely placed in the intention space. When generating the output, the system selects the output phrase close to the determined intention vector: an appropriate phrase can be selected even if the phrase database does not have a phrase that is exactly placed on the intention vector.

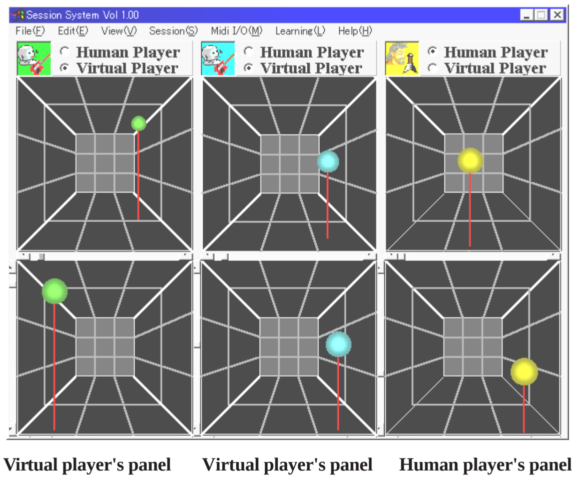
3) Reaction model

We can regard the mapping from the impression space to the intention space as the reaction model. To learn this mapping function statistically, we obtained various training sets from the target session recordings. These sets are pairs of impression vectors obtained from the three players during a sequence of past twelve bars and the corresponding next intention vector. For this learning we use Gaussian radial basis function (RBF) networks [6]. The RBF networks have one hidden layer with nonlinear inputs, and each node in the hidden layer computes the distance between the input vector and the center of the corresponding radial basis function. The RBF networks have good generalization ability and can learn a nonlinear mapping function we are dealing with.

2.2 Session Mode

Using the reaction model acquired in the learning mode, each virtual player improvises while reacting to the human player and the other virtual player. The processing flow of each virtual player can be summarized as follows:

1. The low-level features of MIDI performances of all the players are calculated at every 1/48 bar.
2. Every 1/12 bar the three impression vectors are obtained from the low-level features.
3. At the beginning of every bar, the intention vector is determined by feeding the reaction model past impression vectors.
4. The output performance is generated by connecting short phrases selected from a phrase database that contains a hundred phrases, fifty solo-style phrases, and fifty backing-style phrases. Each phrase is selected, according to the determined intention vector, by considering the fitness for the chord progression. A virtual player can start a solo performance at any bar.



○ Human Player
● Virtual Player

○ Human Player
● Virtual Player

● Human Player
○ Virtual Player

Figure 4: Screen snapshot of the system output.

Note that the reaction model can predict the next intention vector from the impression vectors gathered during the past twelve bars in real time: a virtual player thus does not get behind the other players.

3 Experimental Results

We have implemented the proposed system on a personal computer (with a Pentium III 650 MHz processor) and Figure 4 shows a screen snapshot of the system output. In this figure there are three columns (called *player panels*) , each corresponding to a different player. The toggle switch on the top of each panel in this figure indicates whether the panel is for a virtual player or a human player, and in each panel there are two boxes representing three-dimensional spaces: the upper box is the impression space and the lower box is the intention space. The sphere in each box indicates the current value of the impression or intention vector.

In our experiments, after recording a session performance of three human guitarists playing MIDI guitars, we first made the system learn the reaction model of each of them. We used a metronome sound to keep the tempo (120 M.M.) when recording, and the total length of this recording session was 144 bars. We then let five human guitarists use the system in the session mode. The system indeed enabled each human

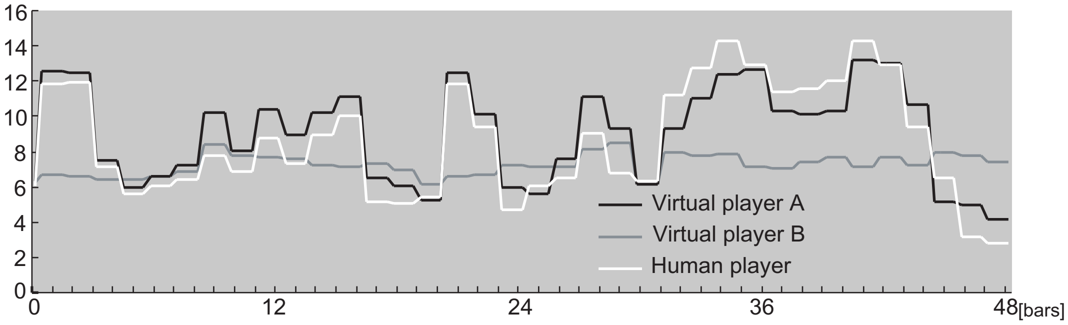


Figure 5: Transition of intention vectors of three players. (This figure charts one component of intention vectors that have three components.)

guitarist to interact with two virtual guitarists, each with a different reaction model.

To find out how well a virtual player could imitate a human player, we asked a human player to perform with a virtual player A imitating him and a virtual player B imitating a different player. The human player and the virtual player imitating him tended to take the solo at almost the same time and to perform phrases that felt similar. Figure 5 shows the transition of intention vectors of three players during 48 bars where the intention vectors of the virtual player A and the human player are particularly similar. Examining all the values of the intention vectors during the session, we compared the distances between the intention vectors of the virtual players and the human player. Over 144 bars the average distance between the intention vectors of the human player and the virtual player imitating him was significantly smaller than that between the intention vectors of the human player and the virtual player imitating a different player. These results showed that our system learned the reaction model from the MIDI recordings of sessions.

Furthermore, five guitarists who performed with the system remarked that each virtual player performed characteristically. In particular, the human player who participated a jam session with a virtual player that imitated himself remarked that it was not comfortable to play with his virtual player because he felt himself mimicked. We think that this means the virtual player's RBF networks actually predicted the human player's intention.

4 Conclusion

We have described a session system in which a human guitarist and two virtual guitarists imitating human guitarists interact with each other. It is based on the learning of a reaction model that is the mapping from the past performances of all the three players to the next performance of the imitated player. The experimental results showed that our system can imitate musical reactions of human players. We plan to extend the system so that it can be applied to other musical instruments, such as piano and drums.

5 Acknowledgments

We thank Daisuke Hashimoto, Shinichi Chida, and Yasuhiro Saita for cooperating in the experiment as players.

References

[1] M. Nishijima and K. Watanabe: Interactive music composer on neural networks, Proc. of ICMC, pp. 53-56, 1992.

[2] Y. Aono, H. Katayose and S. Inokuchi: An improvisational accompaniment system observing performer's musical gesture, Proc. of ICMC, pp. 106-107, 1995.

[3] S. Wake, H. Kato, N. Saiwaki and S. Inokuchi: Cooperative Musical Partner System Using Tension-Parameter: JASPER (Jam Session Partner) , Trans. of IPS Japan, Vol. 35, No. 7, pp. 1469-1481, 1994. (in Japanese)

[4] M. Goto, I. Hidaka, H. Matsumoto, Y. Kuroda and Y. Muraoka: A Jazz Session System for Interplay among All Players — VirJa Session (Virtual Jazz Session System) —, Proc. of ICMC, pp. 346-349, 1996.

[5] J.B. Kruskal and M. Wish, Multidimensional Scaling, Sage Publications, 1978.

[6] S.Chen, C.F.N. Cowan and P.M. Great: Orthogonal Least Sequares Learning Algorithm for Radial Basis Function Networks, IEEE Transactions on Neural Networks, Vol.4, pp.246-257, 1991.