2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Implementation of Expressive Performance Rules on the WF-4RIII by modeling a professional flutist performance using NN

Jorge Solis, Kei Suefuji, Koichi Taniguchi, Takeshi Ninomiya, Maki Maeda and Atsuo Takanishi

***Abstract*—In this paper, the methodology for automatically generating an expressive performance on the anthropomorphic flutist robot is detailed. A feed-forward network trained with the error back-propagation algorithm was implemented to model the performance's expressiveness of a professional flutist. In particular, the *note duration* and *vibrato* were considered as performance rules (sources of variation) to enhance the robot's performance expressiveness. From the mechanical point of view, the vibrato and lung systems were re-designed to effectively control the proposed music performance rules. An experimental setup was proposed to verify the effectiveness of generating a new score with expressiveness from a model created based on the performance of a professional flutist. As a result, the flutist robot was able of automatically producing an expressive performance similar to the human one from a nominal score.**

I. Introduction

THE fact that music can be used as a means for expression and communication is often acknowledged. Yet this is one of the least understood aspects of music, at least as far as scientific explanation goes. Performers introduce some deviations from nominal values specified in the score, which characterizes their own performance. It is known that several performances of the same score often differ significantly, depending on performer's expressive intentions.

Studies in music performance use the word *expressiveness* to indicate the systematic presence of deviations from the musical notation as a communication means between musician and listener [1]. Such deviations represent the added value of a performance and are part of the reason that music is interesting to listen to and sounds alive. In fact, a score played with the exact values indicated in it lacks of musical meaning and is perceived dull as a text read without any prosodic inflexion. Indeed, human performers never respect tempo, timing, and loudness notations (some deviations are always introduced). A performance which is played accordingly to appropriate rules imposed by a specific musical praxis will be called *natural*. In order to understand how humans can express emotions while performing music; several researchers have tried to emulate the human music performance by proposing computational models [2] and by developing mechanical systems which nearly simulate the physiology of the organs involved on the performance of musical instruments [3-5].

From the computational point of view, the analysis of such systematic deviations has led to the formulation of different models that try to describe their structures and aim at explaining where, how, and why a performer modifies (sometimes in an unconscious way) what is indicated by the notation of the score. In recent years, several researchers on the computer music field have focused on the Artificial Intelligence (AI) approaches for developing automatic performance systems in order to capture the knowledge applied when performing a score by means of rules. In order to develop an automatic performance system, mostly two approaches have been proposed: the analysis-by-synthesis and the analysis-by-measurement [6]. Such kind of approaches mainly converts a music score into an expressive musical performance by applying rules (typically including time, sound and timbre deviations). Every rule tries to predict some deviation that a human performer inserts by quantitatively describing the deviations to be applied to a musical score. As a result, more attractive and human-like performances can be generated and simulated.

From the engineering point of view, several researchers have been developing musical performance robots that nearly imitate the function of the organs for playing musical instruments. In particular, authors have been developing an anthropomorphic flutist robot which imitates the human flute playing by emulating the human motor control required to play the flute [3]. The main idea of this approach is to emulate human dexterity and to coordinate the movements of each of the organs involved during the flute playing by mechanical means. For this purpose, the synchronization of all the simulated organs of the flutist robot is realized by reading the timing clock signal from the MIDI data generated from a PC sequencer and by generating an interrupt every 5ms on the PC controller [7]. Even that the Waseda Flutist robot has demonstrated to be able of nearly imitating the performance of an intermediate flutist, the robot's performance still lacks of a human-like expressiveness which is desirable for achieving a more natural performance. Up to now, we have focused on using the analysis-by-measurement method to enhance the expressiveness of the flutist robot's performance;

Manuscript received August 31, 2006. A part of this research was done at the Humanoid Robotics Institute (HRI), Waseda University. This research was supported (in part) by a Gifu-in-Aid for the WABOT-HOUSE Project by Gifu Prefecture. We would like to thanks Ms. Akiko Sato, a professional flutist, for her valuable advices while developing the flutist robot.

Jorge Solis is with the Waseda University, Mechanical Engineering Department, 3-4-1 Ookubo, Shinjuku-ku, 169-8555. Tokyo, Japan (e-mail: solis@kurenai.waseda.jp).

Kei Suefuji, Koichi Taniguchi, Takeshi Ninomiya and Maki Maeda are with the Waseda University, Graduate School of Science and Engineering.

Atsuo Takanishi is with the Waseda University, Mechanical Engineering Department and Humanoid Robotics Institute, 3-4-1 Ookubo, Shinjuku-ku, 169-8555. Tokyo, Japan (e-mail: takanisi@waseda.jp).

where the performance of a professional flutist is analyzed (based on the Fast Fourier Transform [8]) to extract musical parameters such as pitch, volume, tempo, etc. However, this approach cannot provide enough information to describe how performers actually add expression to their performances. In addition, every time the flutist robot is programmed to perform a new score, the recording from such score performed by professional flutist is required.

Therefore, in order to enhance the performance's expressiveness of the flutist robot, an analysis-by-synthesis approach was implemented to model a human performance. In particular, an Artificial Neural Network (ANN) has been implemented to model the musical expressiveness of a professional flutist. By using such a model, a set of performance rules can be extracted to produce an expressive musical performance. Specifically, the *note duration* and *vibrato* were considered as principal sources of variation required for an expressive performance.

The extracted performance rules where then implemented on the performance control system of the Waseda Flutist Robot No.4 Refined III (WF-4RIII). This new version of the flutist robot improved the vibrato and lung systems to effectively control the musical parameters extracted from the resultant performance rules. From the mechanical point of view; the lung system was re-designed to enable a better control of the air while breathing (to effectively add deviations on the *tempo*), and the vibrato system was re-designed to emulate more closely a human-like vocal cord (to effectively control the amplitude and frequency of the vibration added to the air beam).

This paper is organized as follows. In Section II, the way of modeling an expressive performance of a professional flutist using ANN is detailed. In the following section, the improvements of the mechanical system of WF-4RIII are described. Finally, a set of experiments were carried out to verify if the flutist robot could effectively enhance the expressiveness of its performance.

II. Modeling Human Music Performance

A. Analysis of Human Performance

From the computer music field, the research on music performance has been quite intensive in the 20th century, particularly in its last decades. As a result, several automatic performance systems have been developed to convert a music score into an expressive musical performance. As it was previously mentioned, mainly two strategies have been used for the design of such performance systems: analysis-by-synthesis and analysis-by-measurement.

Rules based on an analysis-by-measurement method are derived from measurements of real performances; usually recorded on audio CDs or played with MIDI-enabled instruments connected to a computer [9]. Often the data are processed statistically, such that the rules reflect typical rather than individual deviations from a deadpan performance, even though individual deviations may be musically highly relevant.

The second method implies that the intuitive, nonverbal knowledge and the experience of an expert musician are translated into performance rules. These rules explicitly describe musically relevant factors. The most important is the KTH rule system [10]. Machine learning is also another active research stream. Katayose [11] used some artificial intelligence inductive algorithms to infer performance rules from recorded performances. Similar approaches were proposed by Arcos [12] and Suzuki [13]. Several methodologies of approximation of human performances were developed using, a fuzzy logic approach [14], multiple regression analysis [15], or neural network techniques [16].

Up to now, the implementation of Artificial Intelligence (AI) approaches has demonstrated to be capable of generating high-quality human-like monophonic performances based on examples of human performers [2]. However, all of these systems have tested only by computer systems or midi-enabled instruments which limited the unique experience of a live performance. Therefore, we proposed to implement an AI approach to the performance control system of the WF-4RIII which can provided the unique experience of a live performance. In particular, a feed-forward neural network was implemented to model the musical expressiveness from a performance of a professional flutist. From such a model, a set of musical performance rules can be created and then used by the flutist robot in order to produce an expressive performance; even from a different nominal score (Figure 1). In this paper, the musical performance rules considered are: the *note duration* and *vibrato* (duration and frequency). In order to train the ANN, the teaching signal was obtained by analyzing the considered musical parameters from a recording of a professional flutist performance. In the following sub-sections, the way of analyzing the human performance is described.

B. Note duration

As one of the principal characteristics of an expressive performance, the deviations of *tempo* are added by performers to express emotions. For this purpose, we have

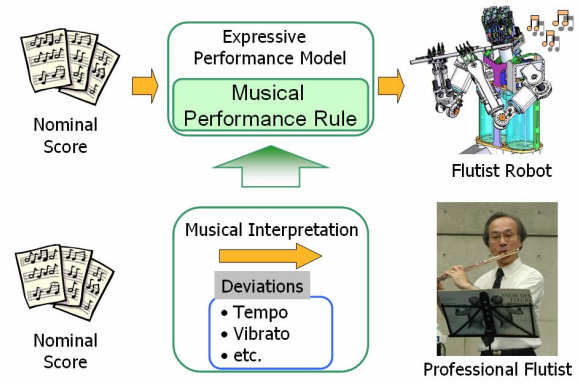

Fig. 1. The flutist robot may produce an expressive performance by extracting the performance rules modeled from a professional flutist.

proposed to analyze the duration of a note from the performance of a professional flutist. In order to analyze such musical parameter, we have recorded an expressive performance of a professional flutist. The recording was sampled at 44100k Hz with a resolution of 16 bits. Such recording was then analyzed by the short-time Fourier transform (STFT). We experimentally found that a frame size of 4096 points (frequency resolution of 10.77 Hz) with a Hanning window obtained a good compromise between resolution and processing speed.

By computing the STFT, the *note duration* can be easily obtained by comparing the volume of the fundamental frequency between two adjacent frames. In Fig. 2, the diagram flux used to determine the duration of a note is shown. Basically, when a note is found, the amplitude of the fundamental frequency is obtained and then compared with the amplitude of the fundamental frequency of contiguous frames to detect when the note has changed.

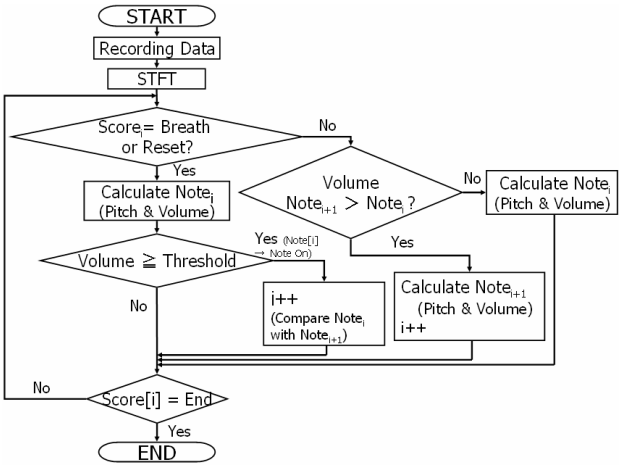

Fig. 2. Algorithm used to determine the *note duration*.

C. Vibrato: Duration and Frequency

Similar to the *note duration*, the *vibrato* plays a key role on producing an expressive performance. The v*ibrato* gives a pleasing flexibility, tenderness and richness to the tone. In flute playing, it is mainly used to add warmth and expressiveness to notes. Basically, the principal parameters of the *vibrato* are the rate and width of modulation. The first parameter is related to how fast the vibrato is being played while the second one is referred to how sharp/flat or how far from the note is being played.

Therefore, we proposed to extract the *vibrato* duration and frequency from the performance of a professional flutist. In order to compute them, a notch filter was applied to the original sound to reduce its noise. Then, for each note the frequency specified in cents was calculated using (1).

By computing the frequency specified in cents of each note of a score, one can easily analyze the duration and frequency of each harmonic of a note (in our case, up to the 3rd harmonic was considered). As an example, in Fig. 3, the analysis of the duration and frequency of the *vibrato* of note A4 is shown.

As a result from the analysis of the *note duration* and v*ibrato*, the professional flutist performance can be analyzed; from where most of the musical parameters can be obtained (such as pitch, note volume, note off, note on and vibrato duration/frequency). In the following section, the method for modeling the human performance is described.

$$f_{cent} = 1200 \log_2 (f_o / f_{average}) \quad (1)$$

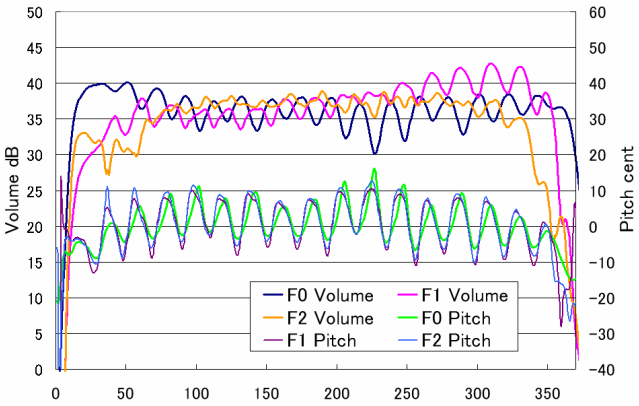

Fig. 3. Analysis of the *Vibrato* from the performance of a human player.

D. Artificial Neural Networks

In order to create an expressive performance, a feed-forward neural network trained with the error back-propagation algorithm was implemented using Borland Builder C++. Feed-forward neural networks are the most widely used models in many practical applications [16]. Such kind of network is divided into layers: input, hidden and output (Figure 4a). The input layer consists of just the inputs to the network. The hidden layer consists of any number of *neurons* placed in parallel. Each neuron (Figure 4b) performs a weighted summation of the inputs, which is then passed through a non-linear function know as an *activation* or *transfer function*. Mathematically the functionality of a hidden neuron is described as (2); where u_i is the internal state of the neuron, h_i is a threshold value and the number of inputs as n. The internal state u_i is represented as in (3); where x_j represents the inputs to neuron and $w_{i,j}$ are the weights between neurons i and j. In order to compute the final network output, the transfer function $\boldsymbol{f}$ is defined as in (4).

$$X_i^n = f(u_i^n - h_i^n). \quad (2)$$

$$u_i^n = \Sigma_{j=1}^n W_{i,j}^{n,n-1} \bullet X_j^{n-1} \quad (3)$$

$$f(x) = 1/(1 + e^{(a-x)}). \quad (4)$$

In our application, we modeled an expressive performance from a professional flutist; where the *note duration* and *vibrato* (duration and frequency) were considered as relevant performance rules in producing local deviation during the flute performance. The model was created by using a nominal score as input and the considered performance rules as outputs (Figure 4a). In order to train the ANN, the

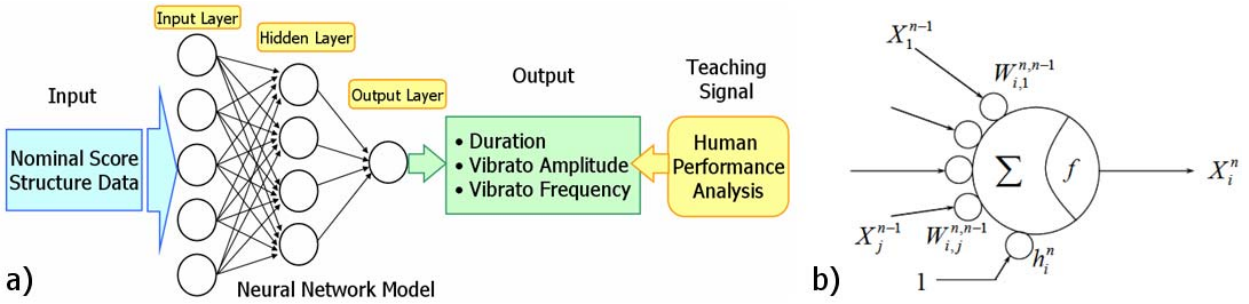

Fig. 4. a) Graphical representation of feed-forward NN trained with the error back-propagation algorithm; b) Representation of one unit in a NN.

back-propagation algorithm was considered. This kind of supervised learning incorporates an external teacher, so that each output unit is told what its desired response to input signals ought to be. During learning, the weight vectors (W_{ij}) are updated using (4); where $E(t)$ is the error between the output value and desired one and η is the learning rate. The ANN was trained to learn the extracted performance rules obtained from the analysis of the professional flutist performance. One of the critical issues while designing a neural network is the generalization; which helps preventing overfitting. Overfitting occurs when a network has memorized the trained set but has not learned to generalize the new inputs. In this paper; as a first approach, we have avoided such situation by defining a small number of hidden layer units and by limiting the number of learning steps (less than 10,000).

$$W_{ij}^{k} = W_{ij}^{k}(t-1) - \eta\left[\delta E(t) \Big/ \delta W_{ij}^{k}(t)\right]. \quad (5)$$

III. Waseda Flutist Robot No. 4 Refined III

The WF-4RIII was developed this year and it has a total of 43-DOFs which reproduced the lips, neck, lung, arms, fingers, vibrato and eyes required for the flute playing performance (Figure 5). Compared with the previous version, the WF-4RII [17], this new version has the same number of degrees of freedom and it has mainly improved the design of the vibrato and lung system in order to implement the performance rules previously described so that an effective control of the *note duration* and *vibrato* can be achieved.

A. New Vibrato Mechanism

The previous vibrato mechanism implemented on the WF-4RII was composed by a coil voice motor which presses directly a tube to add vibrations to the air beam pass through this mechanism [17]. However, human uses a more complicated mechanism to produce a vibrato. It is believed that mainly the vocal cord of human has an important role in producing it. In fact, by observing the laryngeal movement while playing a wind instrument using laryngo-fiberscope, the shape of the vocal cord of flutists differs from the level of expertise [18]. As it is shown in Fig. 6, the laryngoscopic view of the vocal folds demonstrated the differences among them.

Therefore, we believe that the control of the aperture of the glottis plays a key role in producing a human-like vibrato which will help in producing a performance with expressiveness. Thus, a new vibrato mechanism for the WF-4RIII has been designed similar to the shape and human vocal cord (Figure 7a). The vocal cord part was fabricated with a theromoplastic rubber Septon by Kuralay Co. Ltd due to its high stiffness and flexibility. In order to control the amplitude and its frequency of the aperture of the glottis, a DC motor linked to a couple of gears (which are attached to the vocal fold were used through a link) is used (Figure 7b). As a result, the new vibrato system reproduce quite similar the vibration of human vocal cords.

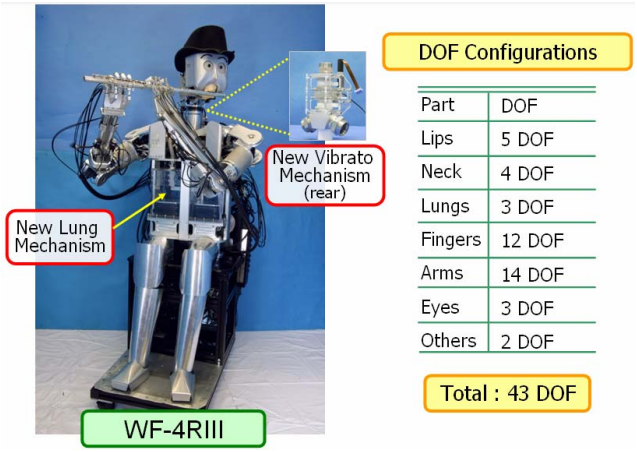

Fig. 5. The Waseda Flutist Robot No.4 Refined III

B. New Lung Mechanism

The previous lung system on the WF-4RII was implemented using two vane mechanisms which were controlled by an AC motor [17]. The breathing process was controlled by a couple of valve mechanisms which were located behind the robot. Even that the mechanical noise was effectively reduced, still some problems were found. In particular, the air conversion efficiency was too low (51%) which means some the existence of some loss of air from the lungs to the lips. Furthermore, the time required for the inhalation was too long (2.36 sec). Such kinds of problems make difficult the accurate control of the *note duration* while playing the flute.

Therefore, a new lung mechanism was designed on the WF-4RIII by using a bellow system located inside an acrylic container. Each of the bellow has not contact with the

container to assure a high airtightness (Figure 8a). In order to increase the inhalation speed, a crank mechanism was used and controlled by an AC servo motor attached to a link; which it moves a shaft connected to the bellow plate (Fig. 8b). A rubber was attached along the shaft. This new design enabled to improve the airtightness to achieve 85% air conversion efficiency and reduce the time required for the inhalation process to 0.64 sec.

Fig. 6. Laryngoscopic view of the vocal folds from different flutists.

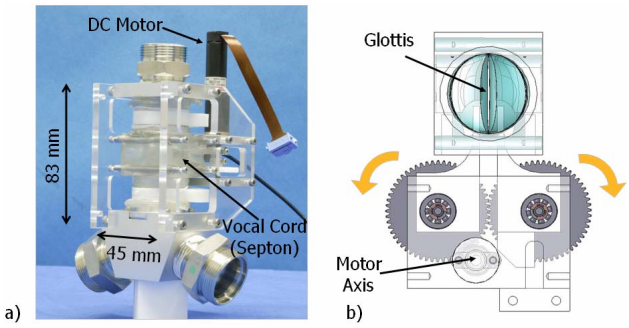

Fig. 7. a) New vibrato mechanism; b) 3D mechanism (top view)

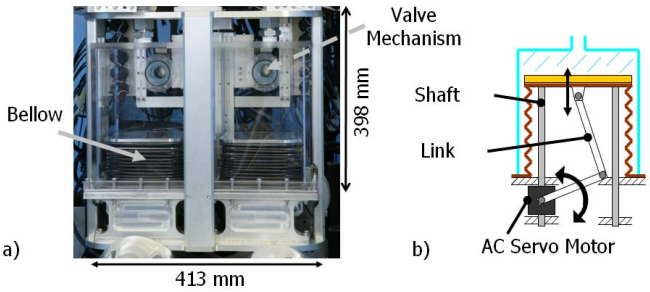

Fig. 8. a) New lung system of WF-4RIII; b) Lung mechanism detail.

IV. Experiments & Results

The experiments presented in this paper are focused in verifying the usefulness of implementing performance rules, modeled from an expressive performance of a professional flutist, to enhance the expressiveness of the WF-4RIII's performance. For that purpose, we would like to investigate if such an expressive performance model could be used by the flutist robot to automatically predict a new expressive score different from the one modeled with the neural network (we are assuming a score with similar style).

The proposed experiment was divided in three steps: at first, an expressive performance model was created from the recorded performance of a professional flutist. Then, we have verified how well the training data fits to the model. Finally; we confirmed how well the created model can predict a different score with expressiveness.

Therefore, we have recorded a professional flutist performing the *Sonata No.4 in C Major* composed by Handel; from where *note duration* and *vibrato* were extracted using the proposed algorithms. Such musical parameters were then use to train the feed-forward neural network (Figure 4b).

In order to verify how well the created model fits the training data, the obtained performance rules were used to create the music data which was converted into MIDI format so that it can be reproduced on a computer system. Such performance was recorded and compared with the professional flutist performance. In order to compare the differences between both performances, the correlation coefficient was computed. The correlation coefficient is a quantity that gives the quality of how well the predicted data fits to the original data. As a result from the comparison, a high correlation coefficient was found for all the considered musical parameters (0.86, 0.93 and 0.86 for the *note duration*, *vibrato* duration and frequency respectively). From this result, we can conclude that the implemented ANN could effectively be used for modeling the expressiveness of a professional flutist.

Finally, we have used the previously produced expressive performance model to automatically predict the required deviations from a different score (with similar musical style). In this case, the musical score *Le Cygne* (composed by Camille Saint-Saëns) was considered. The nominal score was inputted to the expressive performance model and the performance rules were automatically created. The outputs from the ANN were used to produce the music data which was then converted into midi format. The midi file was inputted on the WF-4RIII's control performance system. The robot's performance was recorded and then compared with the professional flutist's performance; where the *note duration* and *vibrato* parameters were extracted. In addition, the midi file was outputted on a midi device connected to a computer system and compared with the professional flutist performance.

The musical parameters obtained from both performances are shown in Fig. 9. Regarding the n*ote duration*, the robot's performance could nearly imitate the behavior found in the human one (correlation coefficient = 0.81). Regarding the *vibrato*, some differences between the human and robot performances were found; however, still the correlation coefficient is acceptable (0.71 and 0.72 respectively). Regarding the comparison between the professional flutist performance and the one reproduced on a midi device; a high correlation coefficient was found for all the considered music parameters (0.86, 0.85 and 0.81 for the *note duration*, *vibrato* duration and frequency respectively). In order to understand the differences between the results obtained from the flutist robot's performance and the one reproduced on a midi device, a t-test statistical analysis was performed; where no statistical difference was detected for the *note duration* and the *vibrato* duration parameters ($p>0.05$). Meanwhile, a statistical difference was detected regarding the *vibrato* frequency was

found ($p<0.05$). This means that the AI approaches could be used to predict an expressive score even when the musical rules are used to produce a live performance based on the WF-4RIIII without considerable differences.

From the results presented above, the implementation of feed-forward neural network enabled the WF-4RIII to automatically predict the required performance rules (*note duration* and *vibrato*) to produce an expressive performance from a nominal score; by using an expressive performance model generated from a professional flutist performing a different score (having a similar style).

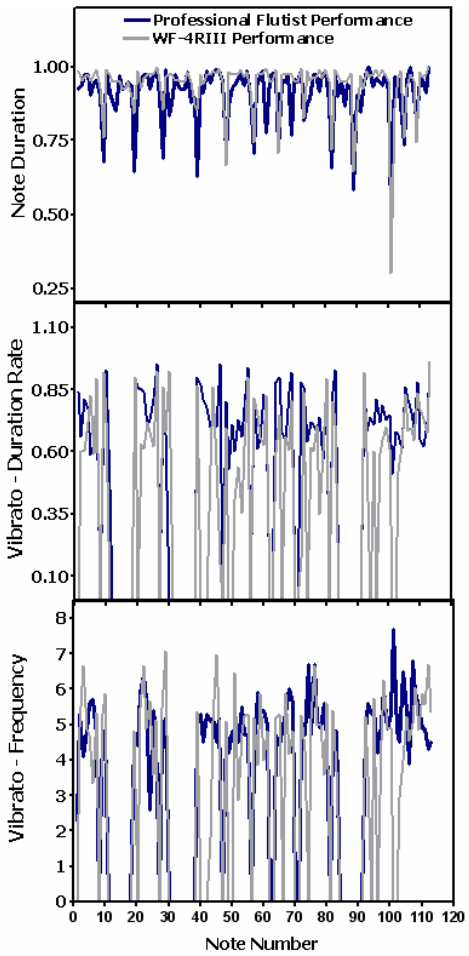

Fig. 9. Comparing the professional flutist performance vs. WF-4RIII.

V. Conclusions & Future Work

In this paper, the development of the WF-4RIII was detailed. From the computational point of view, a feed-forward neural network trained with the error back-propagation algorithm was implemented to create an expressive performance model from a professional flutist performance. As a result, an expressive performance was automatically produced from a nominal score and performed by the WF-4RIII. From the mechanical point of view; the vibrato and lung mechanism were re-designed to effectively control the music performance rules during the robot's performance. In particular, a human-like vocal cord was designed and the lung system was designed to improve the airtightness and to increase the inhalation speed.

Although the WF-4RIII was able of automatically generating an expressive performance, we require performing further improvements on the learning process of the ANN as well as on the performance control system. Regarding the first issue, we will implement more efficient methods [16] to avoid overfitting (i.e. model selection, early stopping, etc). Regarding the performance control system, a feedback signal must be considered during the learning process, so that the flutist robot can also autonomously improve its own performance. Therefore, as a future work, we will propose to implement the feedback-error-learning based on the implemented neural networks.

References

[1] A. Gabrielsson, "Music performance, the psychology of music," in The Psychology of Music, 2nd ed., New York: Academic, 1997, pp. 35-47.

[2] R. L. Mantaras and J.L. Arcos, "AI and Music: From composition to expressive performance," AI Magazine, 2002, pp. 43-58.

[3] J. Solis, K. Chida, K. Suefuji, and A. Takanishi, "The development of the anthropomorphic flutist robot at Waseda University," International Journal of Humanoid Robots, 2006, vol. 30(2), pp. 127-151.

[4] M. Kajitani, "Development of musician robots," Journal of Robotics and Mechatronics, 1989, vol. 1(3), pp. 254-255.

[5] K. Shibuya, "Analysis of Human KANSEI and development of a violin playing robot," in Workshop of the IEEE/RSJ Int. Conference on Intelligent Robots and Systems: Musical Performance Robots and Its Applications, 2006, Beijing, China.

[6] R. Bresin, "Virtual Virtuosity: Studies Automatic Music Performance," Ph.D Thesis, Kungl Tekniska Hogskolan, 2000, p. 32.

[7] K. Chida, I. Okuma, S. Isoda, Y. Saisu, K. Wakamatsu, K. Nishikawa, J. Solis, H. Takanobu, A. Takanishi, "Development of a new anthropomorphic flutist robot WF-4," in *Proc. of IEEE International Conference on Robots and Automation*, 2004. pp. 152-157.

[8] J. Solis, K. Chida, K. Suefuji, K. Taniguchi, S.M. Hashimoto, and A. Takanishi, "Imitation of human flute playing by the anthropomorphic flutist robot WF-4RII," in the Computer Music Journal, 2006, vol. 30(4).

[9] N.P. Todd, "A model of expressive timing in tonal music," Music Perception, 1995, vol. 3, pp. 1940-1949.

[10] A. Friberg, V. Colombo, L. Fryden, and J. Sundberg, "Performance rules for computer-controlled contemporary keyboard music," Comput. Music Journal, 1991, vol. 15(2), pp. 49-55.

[11] H. Katayose and S. Inokuchi, "Learning performance rules in a music interpretation system," Comput. Humanities, 1993, vol. 27, pp. 31-40.

[12] J.L. Arcos and R.L. de Mantaras, "An interactive case-based reasoning approach for generating expressive music," Appl. Intell., 2001, vol. 14(1), pp. 115-129.

[13] T. Suzuki, T. Tolunaga, and H. Tanaka, "A case based approach to the generation of musical expression," in Proc. IJCAI, 1999, pp. 642-648.

[14] R. Bresin, G.D. Poli, and R. Ghetta, "A fuzzy approach to performance rules," in Proc. XI Colloq. on Musical Informatics, 1995, pp. 163-168.

[15] O. Ishikawa, Y. Aono, H. Katayose, and S. Inokuchi, "Extraction of musical performance rule using a modified algorithm of multiple regression analysis," in Proc. KTH Symp. Grammars for Music Performance, 2000, pp. 348-351.

[16] C.M. Bishop. *Neural Networks for Pattern Recognition*. Great Britain: Oxford University Press, 2004, pp. 116-121.

[17] J. Solis, K. Suefuji, K. Chida, K. Taniguchi, and A. Takanishi, "The mechanical improvements of the anthropomorphic flutist robot WF-4RII to increase the sound clarity and to enhance the interactivity with humans," in *Proc. of the 16th CISM-IFToMM Symposium on Robot Design, Dynamics, and Control*, 2006, pp. 247-254.

[18] S. Mukai, "Laryngeal movement while playing wind instruments," in *Proc. of International Symposium on Musical Acoustics*, 1992, pp. 239-241.