Timbre Synthesis Technology that Serves Communication

Naotoshi Osaka, Ken-Ichi Sakakibara†*, and Takafumi Hikichi*

Abstract

Sounds other than spoken language play an important role in communication. Moreover, there are demands for high-quality timbre synthesis technology for creating multimedia contents. NTT Communication Science Laboratories is exploring technologies for synthesizing both musical instrument sounds and singing voices. This paper introduces three technologies: i) a physical model of the “sho,” a Japanese traditional music instrument, ii) a model for synthesizing throat singing, and iii) a sound synthesis system called “O^{t}kinshi” as a tool for content creation.

1. Introduction

Sound information is important in human-to-human communication. Awareness of this importance has increased recently, and, along with visual information, sound is being recognized as vital to the creation of multimedia content. This paper focuses on sound information other than speech as spoken language.

Sound information processing is still unfamiliar to most people. However, sound and music information processing will become popular, and even if you are not well versed in music, you will soon be able to handle and manipulate sound as you like. On the other hand, experts look forward to the development of sound representation technologies. Other possibilities include applications to both artistic music and commercial contents such as games and multimedia contents. The key to these applications is new timbre synthesis technology. NTT Communication Science Laboratories are conducting various research projects on implementing timbre synthesis technologies.

Two areas we are investigating at present are sound morphing [1] and sound hybridization. In sound morphing, a sound is changed smoothly and continuously from one sound to another or an intermediate timbre between the timbres of two given sounds is synthesized. In sound hybridization, elements of sound (loudness, pitch, and timbre) are taken from different sounds and a composite sound is synthesized.

We also are developing technology for synthesizing known musical instrument sounds. By investigating the sounding mechanism of musical instruments scientifically, we hope to understand a family of instruments or one instrument in particular. The advantage of this approach is that it lets us simulate musical performances. However, the technology goes beyond simple synthesis of the original instrument sound: it will enable us to synthesize impossible musical performances with, for example, infinitely fast finger movement.

We are also researching singing voices. From the viewpoint of new timbre synthesis and control, our goal is timbre synthesis of singing voices based on scientific investigation of the voicing mechanism in singing. These research topics are closely related to research on voicing mechanisms in spoken language and general musical instrument sound synthesis, and they should bridge the gap between music and speech research.

All of these technologies have been incorporated into a sound synthesis system.

† NTT Communication Science Laboratories
Atsugi-shi, 243-0198 Japan
E-mail: kis@brl.ntt.co.jp

2. Modeling the sound production mechanism of the sho

We think it important to research Japanese traditional musical instruments and model their sound production mechanisms because little scientific research has been done on such instruments. Furthermore, there is a need to understand why Japanese musical instruments have not been modified toward large volume, stable pitch, and quick control. Besides, these instruments are not treated in computer music composition.

The three main instruments in Gagaku music are the sho, hichiriki, and ryuteki [2]. From the viewpoint of musical acoustics, these resemble the harmonica, oboe, and flute, respectively. The sho resembles the harmonica in that metal reeds, called free reeds, are the source of sound generation, and it is played by blowing and drawing. However, it also has structural differences and, unlike the harmonica, involves fingering. It is difficult to find a Western instrument that has a similar mechanism. The sho was imported into Japan more than a thousand years ago. In spite of such a long history, no detailed research has been done on its sound production mechanism.

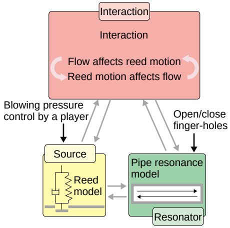
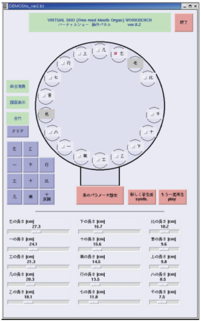


Fig. 1. Physical model of the sho.



乞の長さ [cm] 27.3
一の長さ [cm] 24.1
工の長さ [cm] 21.3
凢の長さ [cm] 20.3
乙の長さ [cm] 18.1
下の長さ [cm] 16.7
十の長さ [cm] 15.6
美の長さ [cm] 14.5
行の長さ [cm] 13.5
七の長さ [cm] 11.8
比の長さ [cm] 10.2
言の長さ [cm] 9.6
上の長さ [cm] 9.8
八の長さ [cm] 8.5
千の長さ [cm] 7.5

Fig. 2. Control panel of our system.

We measured the relationships among reed vibration, resonance frequencies of the pipe, and sound pitch or timbre. Based on the results, we constructed a physical model of the sho [3] and synthesized sounds by computer (Fig. 1). We have built a system controlled by blowing pressure and fingering that has a graphical user interface (Fig. 2). It can be used by beginners for practice. When you input chord names, their fingering positions are highlighted and you hear their synthesized sounds. Thus, you can understand the chords both visually and aurally.

Furthermore, by modifying the pipe lengths or reed characteristics, which is impossible with real instruments, the system can synthesize sounds with vibrato or tremolo. You can control parameters freely and seek your favorite sound timbres with dynamic features. Implementing the sound production mechanism on a computer in this way lets us extend the possibilities of conventional instruments and musical composition.

3. Synthesis and control of timbre of the singing voice

Each style of singing in the world (traditional European singing, Japanese Minyoh, Korean Pansori, Tyvan Khoomei, and so on) is characterized by a timbre that is, for the most part, unique to it. These various timbres may share an important and extensive domain in timbre space. On the other hand, the wide range of timbres in singing voices is produced by continuous control of the phonatory organs, which are common to all human beings. Therefore, modeling their timbre production mechanisms by extracting the parameters that govern the control of phonatory movements based on physiological observation could lead to a system for synthesizing and controlling the various timbres of singing voices.

The timbre of the human voice is varied by controlling both the voice source (*i.e.*, airflow produced by the vocal fold (VF) vibration) and the resonance cavity, whose shape is changed by the motion of the articulatory organs, such as the tongue, lips, and velum (Fig. 3). The voice source controlled by laryngeal adjustment especially contributes to the varied timbres of singing voices.

To model voice sources for timbre synthesis, we are studying throat singing in collaboration with the University of Tokyo. Throat singing, typified by Tyvan Khöömei and Mongolian Khöömij, is very different from traditional European singing in timbre and voice production mechanisms.

In throat singing, there are two different laryngeal voices: the drone and kargyraa. The drone is sometimes called "damigoe" in Japanese and its perceptual impression can be classified as pressed. Moreover, it is slightly different from the normal pressed voice. The kargyraa is a very low pitched voice that ranges out of the modal register.

We observed the laryngeal adjustments and movements of these two different laryngeal voices by simultaneously recording sound waveforms, electroglottography (EGG) waveforms, and high-speed digital images (4500 frames/s) through a flexible endoscope inserted into the nose cavity of a singer [4], [5]. As a result, we clarified that the false vocal folds (FVFs), which do not vibrate in normal phonation, vibrate and contribute to the generation of the special voice sources in throat singing (Fig. 4).

Based on these results and the laryngeal airflow estimated by inverse filtering, we propose a new laryngeal flow model as a signal model which is formulated by mathematical equations. By controlling its parameters, we can use this model to synthesize various laryngeal sources, such as normal and pressed voices and the drone and kargyraa voices in throat singing [6].

We also propose a new physical model of the laryngeal source called the 2 × 2-mass model, which is a self-oscillating model of the VF and FVF vibrations. The model (Fig. 5) was devised by attaching a two-mass model for FVFs to the ordinary two-mass model

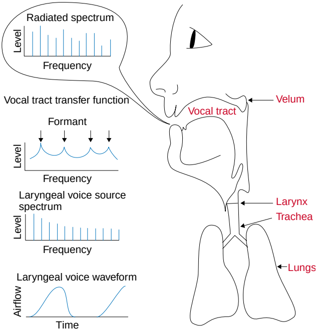


Fig. 3. Human speech organs and mechanism.

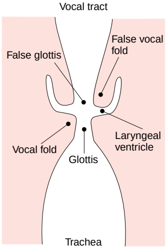


Fig. 4. Human larynx (coronal section).

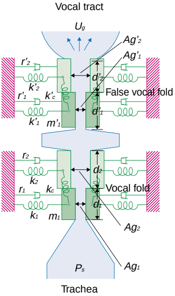


Fig. 5. 2 × 2-mass model.

for VFs with a laryngeal ventricle space between the models. By changing its control parameters, such as mass and stiffness, we can use the 2 × 2-mass model to synthesize various laryngeal voice sources [7].

4. Sound synthesis system *O*^{t}*kinshi*

The name *O*^{t}*kinshi* derives from the Japanese words meaning sound and voice system. There are two versions: a Linux version for researchers and a Windows version for ordinary users. Figure 6 shows the system configuration. Techniques for sound morphing, vibrato control (in which vibrato is added to or subtracted from a musical sound), and physical-model-based sound synthesis are implemented in software. General-purpose functions include a wave editor and spectral editor, whose functions are equivalent to those of a commercial wave editor.

Another feature of the system is its user interface [8]. Previous systems have had a clear distinction between music software and timbre software. Therefore, music creators have had to manipulate two types of software separately. This inconvenience was cleverly solved by defining a sound object hierarchically from the top (the music level) to the bottom (the timbre level). Figure 7 shows the data structure and the display of the "sound object." In the sound object, a music piece is represented by one icon. Double clicking an icon brings up a detailed panel. Each stage of the hierarchy, has an icon display (layer 0) and a panel display of the details (layer 1). In the detailed display panel, sound is expressed as a combination of multiple parts (channels). Each part can be traced to the lower layer by double clicking. The lowest layer represents a sound wave display. This straightforward expression of music to sound was not available before, and it makes the manipulation of both sound and music easier.

The system stores a user's operation history using the programming language "Ruby." The history is also displayed serially using operation icons. As a result, sounds can be modified and resynthesized using icons. Moreover, users can re-run the recorded and modified text for the first time. This function is very convenient and efficient when, after creating a particular sound, you want to have multiple sounds repeatedly with slightly changed parameters. We are making preparations so that those who want to use the prototype system can download it from the Internet. We are also preparing to provide the system to CCRMA (Center for Computer Research in Music

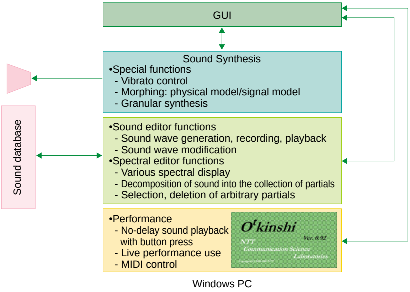

GUI
Sound Synthesis
•Special functions
- Vibrato control
- Morphing: physical model/signal model
- Granular synthesis
•Sound editor functions
- Sound wave generation, recording, playback
- Sound wave modification
•Spectral editor functions
- Various spectral display
- Decomposition of sound into the collection of partials
- Selection, deletion of arbitrary partials
•Performance
- No-delay sound playback with button press
- Live performance use
- MIDI control
O'kinshi
Ver. 0.92
NTT Communication Science Laboratories
Sound database
Windows PC

Fig. 6. System configuration.

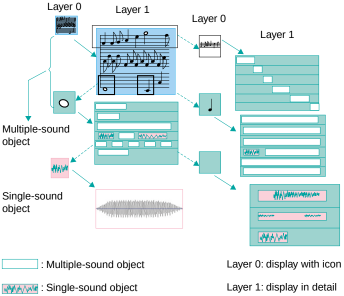


Fig. 7. Hierarchical representation of sound objects.

and Acoustics), Stanford University for reserch, music composition, and education.

5. Conclusions

In NTT Communication Science Laboratories, the range of research relevant to music timbre extends from the science of sound to prototype system development. We expect to clarify various sound mechanisms and establish sound synthesis technology and systems by applying. The results of our research will find uses in content, such as games and broadcast programs, and artistic representation such as concert music.

References

[1] N. Osaka, "Timbre interpolation of sounds using a sinusoidal model," ICMC 95 Proceedings, Banff, Canada, pp. 408-411, Sep. 1995.

[2] S. Sadie ed., The New Grove Dictionary of Music and Musicians, 9, Japan, III, 1, pp. 510-515, Macmillan Publishers Limited, London, 1980.

[3] T. Hikichi, N. Osaka, and F. Itakura, "Time-domain simulation of sound production of the sho," Journal of Acoustical Society of America, Vol. 113, No. 2, pp. 1092-1101, 2003.

[4] K.-I. Sakakibara, S. Adachi, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, M. Todoroki, H. Imagawa, and S. Niimi, "Analysis of vocal fold vibrations in throat singing," Tech. Rep. Musical Acoust., Vol. 19, No. 4, pp. 41-48, Acoust. Soc. Jpn., Sep. 2002 (in Japanese). http://www.acoustics.org/press/144th/Sakakibara.htm

[5] K.-I. Sakakibara, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, H. Imagawa, and S. Niimi, "Vocal fold and false vocal fold vibrations and synthesis of khoomei," Proc. of ICMC, pp. 135-138, Sep. 2001.

[6] K.-I. Sakakibara, T. Konishi, K. Kondo, E. Z. Murano, M. Kumada, H. Imagawa, and S. Niimi, "Vocal fold and false vocal fold vibrations and synthesis of khoomei," Proc. of ICMC, pp. 135-138, Sep. 2001.

[7] K.-I. Sakakibara, H. Imagawa, S. Niimi, and N. Osaka, "Synthesis of the laryngeal source of throat singing using a 2 × 2-mass model," Proc. of ICMC, pp. 5-8, Sep. 2002.

[8] N. Osaka and T. Hikichi, "Visual manipulation environment for sound synthesis, modification and performance," Proceedings of the 1999 International Computer Music Conference, pp. 429-432, 1999.

Naotoshi Osaka

Former Senior Research Scientist, leading Media Representation Research Group, NTT Communication Science Laboratories.

He received the M.S. degree in electrical engineering from Waseda University in 1978. His main research interests include telephone transmission performance and speech dialogue. He received a D.Eng degree for an objective model for telephone transmission performance. He is currently studying timbre synthesis for both sounds and speech. He is currently a professor at Tokyo Denki University. He is a senior member of IEEE and is also a member of the IEICE, IPSJ and the ASJ. He is also a vice-president of International Computer Music Association (ICMA).

Ken-Ichi Sakakibara

Research Scientist, Media Representation Research Group, NTT Communication Science Laboratories.

He received a B.Sc. and M.Sc. in mathematics from Kyoto University. He is currently interested in the timbre of musical sounds, singing voices, and voice quality. He is a member of the Acoust. Soc. Jpn., the Acoust. Soc. Am., the IPSJ, and the Jpn Soc. of Logop. and Phoniat. and also a member of the board of directors of the Jpn. Association of Vocalization Instructors.

Takafumi Hikichi

Research Scientist, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.S. and M.S. degrees in electrical engineering from Nagoya University, in 1993 and 1995, respectively. In 1995, he joined NTT Basic Research Laboratories. He was awarded the Awaya-Kiyoshi prize from the Acoustical Society of Japan in 2000. He is now working at the NTT Communication Science Laboratories. Mr. Hikichi is a member of the Acoustical Society of America, the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, and the Acoustical Society of Japan.