VOCALISTENER AND VOCAWATCHER: IMITATING A HUMAN SINGER BY USING SIGNAL PROCESSING

Masataka Goto, Tomoyasu Nakano, Shuuji Kajita, Yosuke Matsusaka, Shin'ichiro Nakaoka, and Kazuhito Yokoi

National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

In this paper, we describe three singing information processing systems, *VocaListener*, *VocaListener2*, and *VocaWatcher*, that imitate singing expressions of the voice and face of a human singer. VocaListener can synthesize natural singing voices by analyzing and imitating the pitch and dynamics of the human singing. VocaListener2 imitates temporal timbre changes in addition to the pitch and dynamics. In synchronization with the synthesized singing voices, VocaWatcher can generate realistic facial motions of a humanoid robot, the *HRP-4C*, by analyzing and imitating facial motions of a human singing that are recorded by a single video camera. These systems that focus on "imitation" are not only promising for representing human-like naturalness, but also useful for providing intuitive control means.

Index Terms— Music, singing information processing, singing synthesis, singing robot

1. INTRODUCTION

In light of the growing importance of singing synthesis, the first goal of this research is to find ways to more easily synthesize human-like singing voices with natural expressions. Both amateur and professional musicians have started to use singing synthesizers as their main vocals, and songs sung by computer singers rather than human singers have become popular, often appearing on popular music charts in Japan [1]. As music synthesizers generating various instrumental sounds are already widely used and have become indispensable to popular music production, it is historically inevitable that singing synthesizers will become more widely used and likewise indispensable to music production. The only element of uncertainty is whether it will happen soon or farther into the future with more advanced technologies. Regardless of how soon this happens, increasing the naturalness and expressiveness of synthesized singing voices will contribute to the popularity of singing synthesis.

In addition to singing synthesis, the second goal of this research is to enable a humanoid robot to sing with realistic facial expressions and natural synthesized singing voices. Such a robot singer is an important and attractive humanoid robot application for the entertainment scene. In our experiences in exhibitions, a lot of people simply want to see humanoid robots walking, moving, singing, dancing, etc. Supported by such interests of the general public, robot singers are promising applications of robot technologies. It is, however, difficult to achieve realistic and natural expressions since it requires state-of-the-art integration of different technologies such as robot engineering, music processing, and image processing.

To overcome this difficulty of increasing the naturalness of singing voices and robot motions, this research focuses on the "imitation" of a human singer. As shown in Figure 1, we developed three singing information processing systems [2], a singing synthesis system *VocaListener* [3] to imitate the pitch and dynamics of the singer's voice, a singing synthesis system *VocaListener2* [4] to imitate timbre changes in addition to the pitch and dynamics, and a robot motion generation system *VocaWatcher* [5] to imitate facial expressions of the singer's face. VocaListener or VocaListener2 can be used independently of VocaWatcher to synthesize natural singing voices only. On the other hand, VocaWatcher should be used together with VocaListener or VocaListener2 to enable a humanoid robot to sing along with synthesized singing voices and with realistic facial expressions.

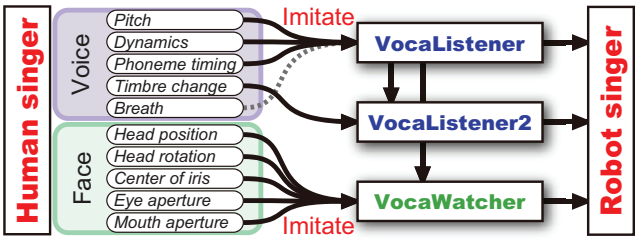

Fig. 1. VocaListener, VocaListener2, and VocaWatcher.

We first developed VocaListener [3] as a system to imitate the pitch and dynamics of a target human singing. With the help of the given lyrics of the song being sung, it automatically estimates the musical score of the song from the target singing. Because VocaListener can estimate expressive control parameters of various commercial singing synthesizers based on Yamaha's VOCALOID or VOCALOID2 technology [1], it easily synthesizes various singing voices that have identical pitch, dynamics, and lyrics, but different timbres. Thanks to the estimated natural expressions of the target human singing, synthesized singing voices can be human-like and natural without time-consuming manual adjustment of the control parameters. Temporal timbre changes of the target singing, however, are not imitated.

As an extension of VocaListener, we developed VocaListener2 [4], a system that imitates timbre changes, not only the pitch and dynamics, of the target human singing. Given a song, VocaListener2 constructs a *voice timbre space* by using various singing voices that are synthesized by VocaListener to have the identical pitch, dynamics, and lyrics, but different timbres. Temporal timbre changes of the target singing are represented as a trajectory in this voice timbre space, and the trajectory is used for synthesizing imitated singing voices.

Furthermore, for a robot singer, we developed VocaWatcher [5], a system that imitates facial expressions of a human singer's face during singing by analyzing a video clip of a person singing that is recorded by a single video camera. VocaWatcher can control the mouth, eye, and neck motions of a biped humanoid robot, the *HRP-4C* [6], by imitating corresponding human motions that are estimated without using any markers in the video clip. The HRP-4C has a realistic female facial appearance and body shape (160 height and 46 kg weight with 44 degrees of freedom). The imitated facial motions can be precisely synchronized, at a phoneme level, with synthesized singing voices by using the phoneme timing provided by VocaListener or VocaListener2.

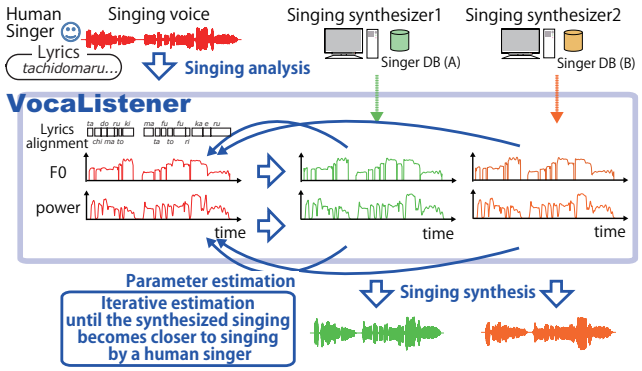

Fig. 2. Overview of VocaListener that iteratively estimates parameters of pitch (F0) and dynamics (power) for different singing synthesizers from a singing voice of a human singer and its lyrics.

2. VOCALISTENER: SINGING SYNTHESIS SYSTEM IMITATING PITCH AND DYNAMICS

Given audio signals from a singing voice and the written text of the song's lyrics, VocaListener [3] can imitate the singing voice by estimating singing synthesis parameters (pitch and dynamics) for various singing synthesizers such as VOCALOID and VOCALOID2 software [7]. We named this approach *singing-to-singing synthesis* [3] because a user need only provide a singing voice along the lyrics without a musical score for singing synthesis.

The most popular approach for singing synthesis is *lyrics-to-singing (text-to-singing) synthesis* where a user provides note-level score information of the melody with the corresponding lyrics to synthesize a singing voice [7–9]. To improve naturalness and provide original expressions, some systems [7] enable a user to adjust singing synthesis parameters such as pitch (F0) and dynamics (power). The manual parameter adjustment, however, is not easy and requires considerable time and effort. Another approach is *speech-to-singing synthesis* where a speaking voice reading the lyrics of a song is converted into a singing voice by controlling acoustic features [10] according to a given score. This approach is interesting because a user can synthesize singing voices having the user's voice timbre, but various voice timbres cannot be used.

Janer *et al.* [11] took an approach similar to our *singing-to-singing synthesis*. Their method analyzes acoustic features of the input user's singing and directly converts them into synthesis parameters. Their method is not robust, though, with respect to different singing synthesis conditions. For example, even if we specify the same parameters, the synthesized results always differ when we change to another singing synthesizer or a different singer database because each database includes a different set of voice waveforms. The ability to imitate a user's singing is therefore limited.

To overcome such limitations on robustness, VocaListener iteratively estimates singing synthesis parameters so that after a certain number of iterations the synthesized singing can become more similar to the user's singing in terms of pitch and dynamics (Figure 2). In short, VocaListener can synthesize a singing voice while listening to its own generated voice through an original feedback-loop mechanism. First, acoustic features (F0 and power) estimated from the target singing are converted into synthesis parameters (pitch and dynamics) for each singing synthesizer based on VOCALOID or VOCALOID2. Second, these parameters are fed to each singing synthesizer to obtain a tentative synthesized singing, which is analyzed and compared with the target singing. Synthesis parameters are then updated so that the compared differences can be made smaller. Until the synthesized singing is sufficiently close to the target singing, the system repeats the parameter updating and its synthesis.

To generate note-level score information regarding the melody,

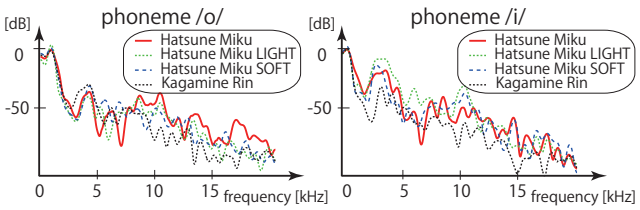

Fig. 3. Examples of differences in the spectral envelope due to differences of phonemes and singer databases (DBs).

VocaListener supports an accurate lyrics-to-singing synchronization function. Given the user's singing and the corresponding lyrics without any score information, VocaListener synchronizes them automatically to determine each musical note that corresponds to a phoneme of the lyrics. This synchronization is called *phonetic alignment* and is estimated through Viterbi alignment with an originally adapted/trained acoustic model (a phoneme-level hidden Markov model). The alignment results should then be adjusted iteratively so that each voiced section of the synthesized singing can be the same as the original voiced section of the target singing. Although synchronization errors with this model occasionally occur, we also provide an interface that lets a user easily correct such errors just by pointing them out.

In addition, VocaListener also supports pitch and style modification functions, *off-pitch correction*, *pitch transposition*, *vibrato adjustment*, and *singing smoothing*, to improve synthesized singing as if the user's singing skill were improved. A user can select whether to use these functions based on personal preference.

3. VOCALISTENER2: SINGING SYNTHESIS SYSTEM IMITATING VOICE TIMBRE CHANGES

Given a singing voice and the song's lyrics, VocaListener2 [4] can imitate temporal timbre changes of the singing voice as well as pitch and dynamics. This is also based on the *singing-to-singing synthesis* approach and is an extension of VocaListener, which deals with only pitch and dynamics.

Much previous work has been done on manipulating voice timbre such as speaking voice conversion [12, 13], emotional speech synthesis [14–16], singing voice conversion [17], and singing voice morphing [18]. However, these approaches cannot deal with intentional temporal timbre changes during singing. In contrast, VOCALOID [7] enables a user to adjust singing synthesis parameters to manipulate acoustic features (e.g., the spectrum) of synthesized singing for each instant of time. The manual parameter adjustment is not easy, though, and requires considerable time and effort.

Although some commercial singing synthesizers can synthesize voices of different styles (timbres) for a singer, their intermediate voices cannot be synthesized. For example, a singing synthesis software named *Hatsune Miku Append* (referred to as MIKU Append) can synthesize voices of six styles (DARK, LIGHT, SOFT, SOLID, SWEET, and VIVID) that have the same individuality as with *Hatsune Miku* and differ in voice timbre. Each voice or style is called a *singer database (DB)*. An intermediate voice (*e.g.,* between LIGHT and SOLID), however, cannot be synthesized.

To deal with these problems, VocaListener2 automatically controls temporal voice timbre changes by using signal processing techniques to manipulate the spectral envelope shapes of synthesized singing voices. We define voice timbre changes as differences in spectral envelope shape, such as spectral differences between *Hatsune Miku* and *MIKU Append*. As shown in Figure 3, such differences are caused not only by different voice timbres (*e.g.,* between Hatsune Miku and MIKU Append (LIGHT)), but also by different phonemes (*e.g.,* between /o/ and /i/), pitch, and dynamics. But when the phoneme, pitch, and dynamics are set to be the same by VocaL-

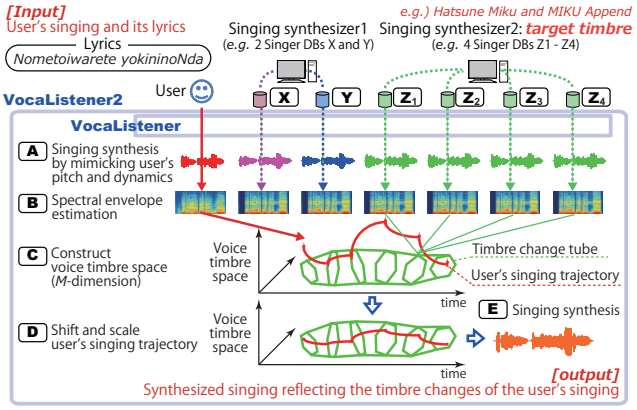

Fig. 4. Overview of VocaListener2 that automatically synthesizes a singing voice by imitating the pitch, dynamics, and timbre changes of a user's singing voice.

istener, only voice timbre changes can be represented as differences in a spectral envelope shape by VocaListener2.

Figure 4 shows an overview of VocaListener2. It consists of VocaListener, singing analysis (A–D), and singing synthesis (E). First, given a target human singing (a user's singing voice) and the song's lyrics, VocaListener is used to synthesize temporally synchronized singing voices having the same pitch and dynamics from many different singer DBs (A). Second, VocaListener2 estimates the spectral envelope of each sample of synthesized singing (B) by using the STRAIGHT speech manipulation system. Since differences in these envelopes represent only the voice timbre changes under the same phoneme, pitch, and dynamics, it then constructs an M-dimensional *voice timbre space* (C) by using a subspace method ($M = 3$ in our current implementation). At each temporal frame, any singing voice can be represented as a point in this space. The target user's singing or each synthesized singing using a singer DB is therefore represented as a temporal trajectory in this space. Although all singer DBs are used to construct the voice timbre space, a subset of singer DBs having the same singer individuality (e.g., seven singer DBs of Hatsune Miku and MIKU Append (DARK, LIGHT, SOFT, SOLID, SWEET, and VIVID)) has to be selected to synthesize an output singing voice that has temporal timbre changes similar to the target singing while keeping the individuality of the selected singer DBs (Z_1–Z_4). An M-dimensional *timbre change tube*, where M-polytope at each temporal frame forms a tube along the time axis, is therefore constructed by using such selected singer DBs. The trajectory of the target user's singing is then adjusted (shifted and scaled) so that it can be inside the tube (D). Finally, the output singing voice is synthesized from spectral envelopes corresponding to this adjusted trajectory (E).

4. VOCAWATCHER: ROBOT MOTION GENERATION SYSTEM IMITATING FACIAL EXPRESSIONS

Given a video clip of a singing performance recorded by a single video camera, VocaWatcher [5] enables the female HRP-4C robot to imitate the human singer by generating realistic facial expressions synchronized with synthesized singing voices. In the input video clip, a female human singer stood in front of the microphone and a fixed camera so that the upper half of her body could be recorded.

Because music is an important and attractive application for humanoid robots, various robots have been developed that play musical instruments [19–21]. Although singing humanoid robots with synthesized singing voices have also been developed, the movements of such robots [22–24] are not natural because of limitations of manual control. Our HRP-4C humanoid robot was used to sing a song at

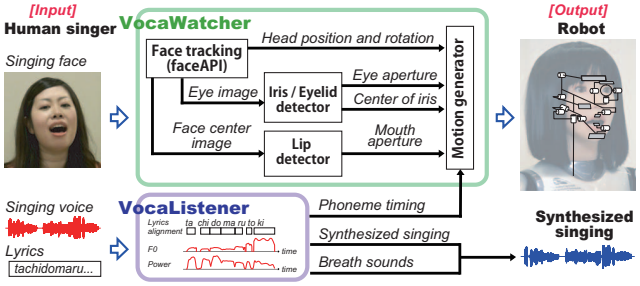

Fig. 5. Overview of VocaWatcher that automatically generates natural facial expressions of a robot singer by imitating a human singer.

a public exhibition, CEATEC JAPAN 2009, with well known techniques such as key poses, semi-automatic motion generation, and a commercial singing synthesizer [24], but its singing performance was not natural enough. Thus, none of these earlier robots provide realistic facial expressions synchronized with a naturally synthesized singing voice. As shown in computer graphics, motion capture techniques for imitating human motions are an effective way to generate realistic motions for robots. Previous approaches, however, required many reflective markers [25] or person-specific training data [26], and did not use audio signal processing to help facial control.

VocaWatcher uses both image and audio signal processing to produce a realistic singing performance with the HRP-4C robot (Figure 5). Our robot can sing naturally by using VocaListener to synthesize singing voices that imitate the singing voice of a human singer in the video clip. VocaListener2 can also be used, but it is yet to be tested for this purpose. The robot can also generate synchronous mouth, eye, and neck motions to imitate the facial expressions of the same human singer in the clip. VocaWatcher requires neither facial markers nor multiple cameras and can utilize audio-based timing information — i.e., phoneme timing provided by VocaListener — to improve robot motions.

Figure 5 shows an overview of VocaWatcher. VocaWatcher first detects the head position and rotation by using commercial face-tracking software, *faceAPI* from Seeing Machines. It also detects the iris and eyelid by using our original detection technique based on a subpixel algorithm, and the mouth aperture by tracking upper and lower lips on the basis of a particle filter. The detected motions are then used to generate motions of the neck (three joints), each eye (two joints), and the mouth (four joints). In addition, the mouth shape and motions are controlled so that they can be synchronized with each phoneme (vowel) of the synthesized singing voice. By using the vowel sequence and precise timing information provided by VocaListener, the mouth motions are generated so that predefined key mouth shapes corresponding to vowels and a breath can be reproduced at the beginning of each vowel or breath. Because the breath was not supported by original VocaListener, we extended it to imitate breath sounds that make the robot singing more natural because a human singer often opens the mouth during breathing.

Figure 6 compares an original human singer and a robot singer whose motions were generated by VocaWatcher. To make the robot performance more attractive, arm motions were manually choreographed. This live demonstration was open to the public at CEATEC JAPAN 2010 [5].

5. CONCLUSION

We have described our research aimed at building easy-to-use expressive singing synthesis systems for music production and a realistic facial motion generation system for a robot singer. Demonstrations are available at http://staff.aist.go.jp/t.nakano/*systemname*/ where *systemname* should be replaced with either VocaListener, VocaListener2, or VocaWatcher.

(a) at the start of singing, voicing /ta/ sound

(b) closing eyes, voicing /ru/ sound

(c) voicing long /ra/ sound

Fig. 6. Examples of the face of a human singer (on the left) and the face of the robot singer (on the right) with generated facial expressions imitating the human singer.

We believe that user interfaces will play an important role in singing synthesis systems. Although the lyrics-to-singing (text-to-singing) synthesis is currently the most popular, some users have difficulty using it or feel it is too troublesome dealing with the piano-roll or score editor. As real-time recordings of live MIDI input are intuitive and indispensable for musicians to use general sound synthesizers, our singing-to-singing synthesis can serve as an easier, more efficient option for a user who can sing along with a song or use solo singing voices without accompaniments. Similarly, when a user can personally control robot singers in the future, facial input like that of VocaWatcher will be an easier option than manual motion adjustment. Such intuitive options without requiring time-consuming manual adjustment are important to help users focus on how to express the user's message or intention through a song.

This research starts from the "imitation" of natural human expressions. Of course, imitation places limitations on freedom, but we think that once we can represent a high degree of naturalness by using control parameters for synthesizing singing voices or generating robot motions, we will be able to take the next step of modeling the naturalness in a control parameter space to generate new natural representations beyond mere imitations.

Because of the human-like naturalness of singing voices synthesized by VocaListener and VocaListener2 and the facial expressions generated by VocaWatcher, we found that some people at first sight do not realize that these are computer-generated results. On the other hand, we are aware that some people feel a sense of creepiness that is exemplified by the well known "uncanny valley". We are not worried about the uncanny valley because it just means that we are in a transitional stage in the development of future technologies that can go beyond the uncanny valley. We believe that it is important to keep working toward such future technologies.

6. REFERENCES

[1] H. Kenmochi, "VOCALOID and Hatsune Miku phenomenon in Japan," in *Proc. of InterSinging 2010*, pp. 1–4, 2010.

[2] M. Goto *et al.*, "Singing information processing based on singing voice modeling," in *Proc. of ICASSP 2010*, 2010.

[3] T. Nakano and M. Goto, "VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation," in *Proc. of SMC 2009*, pp. 343–348, 2009.

[4] T. Nakano and M. Goto, "VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," in *Proc. of ICASSP 2011*, pp. 453–456, 2011.

[5] S. Kajita *et al.*, "VocaWatcher: Natural singing motion generator for a humanoid robot," in *Proc. of IROS 2011*, 2011.

[6] K. Kaneko *et al.*, "Cybernetic Human HRP-4C," in *Proc. of Humanoids 2009*, pp. 7–14, 2009.

[7] H. Kenmochi and H. Ohshita, "Vocaloid – commercial singing synthesizer based on sample concatenation," in *Proc. of Interspeech 2007*, pp. 4011–4010, 2007.

[8] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[9] K. Saino *et al.*, "An HMM-based singing voice synthesis system," in *Proc. of Interspeech 2006*, pp. 1141–1144, 2006.

[10] T. Saitou *et al.*, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Proc. of WASPAA 2007*, pp. 215–218, 2007.

[11] J. Janer, J. Bonada and M. Blaauw, "Performance-driven control for sample-based singing voice synthesis," in *Proc. of DAFx-06*, pp. 41–44, 2006.

[12] T. Toda, A. Black and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. on ASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[13] F. Villavicencio and E. Maestre, "GMM-PCA based speaker-timbre conversion on full-quality speech," in *Proc. of SSW-7*, pp. 56–61, 2010.

[14] O. Türk and M. Schröder, "A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis," in *Proc. of Interspeech 2008*, pp. 2282–2285, 2008.

[15] T. Nose, M. Tachibana and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Trans. on Information and Systems*, vol. E92-D, no. 3, pp. 489–497, 2009.

[16] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[17] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Proc. of Interspeech 2010*, pp. 2162–2165, 2010.

[18] H. Kawahara *et al.*, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," in *Proc. of ICASSP 2009*, pp. 3905–3908, 2009.

[19] I. Kato *et al.*, "The robot musician WABOT-2 (Waseda Robot-2)," *Robotics*, vol. 3, pp. 143–155, 1987.

[20] K. Chida *et al.*, "Development of a new anthropomorphic flutist robot WF-4," in *Proc. of ICRA 2004*, pp. 152–157, 2004.

[21] T. Mizumoto *et al.*, "Thereminist robot: Development of a robot theremin player with feedforward and feedback arm control based on a theremin's pitch model," in *Proc. of IROS 2009*, pp. 2297–2302, 2009.

[22] Y. Kuroki *et al.*, "A small biped entertainment robot exploring attractive applications," in *Proc. of ICRA 2003*, pp. 471–476, 2003.

[23] K. Murata *et al.*, "A robot singer with music recognition based on real-time beat tracking," in *Proc. of ISMIR 2008*, pp. 199–204, 2008.

[24] M. Tachibana, S. Nakaoka and H. Kenmochi, "A singing robot realized by a collaboration of VOCALOID and Cybernetic Human HRP-4C," in *Proc. of InterSinging 2010*, pp. 9–14, 2010.

[25] F. Wilbers, C. Ishi and H. Ishiguro, "A blendshape model for mapping facial motions to an android," in *Proc. of IROS 2007*, pp. 542–547, 2007.

[26] P. Jaeckel, N. Campbell and C. Melhuish, "Facial behavior mapping — from video footage to a robot head," *Robotics and Autonomous Systems*, vol. 56, pp. 1042–1049, 2008.