SINGING INFORMATION PROCESSING

Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST), Japan
m.goto [at] aist.go.jp

ABSTRACT

This paper introduces *singing information processing*, which is defined as music information processing for singing voices. As many people listen to music with a focus on singing, singing is one of the most important elements of music. Singing information processing attracts attention not only from a scientific point of view but also from the standpoint of commercial applications, such as singing synthesis, automatic singing pitch correction, query-by-humming, and singing skill evaluation for *karaoke*. The concept of singing information processing is broad and still emerging.

Index Terms— Singing information processing, music information research

1. INTRODUCTION

As music information research [1–4] has continued to develop, research activities related to singing have become more vigorous. Such activities are attracting attention not only from a scientific point of view, but also from the standpoint of commercial applications. Singing-related research is highly diverse, ranging from basic research on the features unique to singing to applied research such as that on singing synthesis, lyrics recognition, lyrics alignment, singer identification, retrieval of singing voices, singing skill evaluation, singing training, singing voice conversion, singing impression estimation, and singer robot. I have named this broad range of singing-related studies *singing information processing* [5].

Since singing is one of the most important elements of music, singing information processing has a major impact on society from the viewpoints of industry and culture. In fact, automatic pitch-correction technology for vocals is already being used on a routine basis in the production of commercial music (popular music, in particular). It has become essential for correcting pitch at points in a song where the singer is less than skillful and for achieving a desired artificial effect. A function for evaluating (scoring) a person's singing in the *karaoke* industry is also popular. More recently, singing-synthesis systems have become widely used and people actively enjoy songs with synthesized singing voices as the main vocals.

In this paper, I treat all music-related sounds uttered from a person's mouth — whether they are generated by regular singing or even by "vocal percussion" (mimicking drum sounds) — as *"singing"*. In the following, I describe this research field of singing information processing by introducing several systems we have developed.

2. SINGING SYNTHESIS

Achieving the synthesis of singing voices is one of the oldest research problems [6] taken up in the field of singing information processing. In recent years, the research into singing synthesis has developed in diverse ways.

Singing synthesis has been actively studied by European research institutions, especially in the second half of the 1980s and throughout the 1990s. Much of this research has focused on synthesizing ever better opera singing and to clarify the mechanism of opera singing through such attempts at synthesis. Representative examples of this research include synthesized singing voices of a virtual castrated male singer (castrato) created by IRCAM (Institute for Research and Coordination Acoustic/Music) in France for the movie "Farinelli Il Castrato" [7], a singing-synthesis system using formant synthesis by Sundberg at KTH (Sweden's Royal Institute of Technology) [8], and a singing-synthesis system using a physical model of an articulatory vocal tract by Cook [9]. More recently, singing-synthesis systems based on corpus-based synthesis methods have been proposed [10–16]. These methods include concatenative synthesis methods and hidden Markov model (HMM) synthesis methods. Other examples include a system for analyzing, processing, and synthesizing scat singing [17], singing morphing that blends different singing styles [18–20], and singing synthesis that imitates the user's singing performance [21, 22].

Since 2007, singing synthesis technology represented by Yamaha's *VOCALOID* [12] has gained much attention in Japan. 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 and are now being posted in large numbers on video-sharing services like *Niconico* (`http://www.nicovideo.jp/video_top/`) and *YouTube* (`http://www.youtube.com`) [23, 24]. Compact discs featuring compilations of songs created using singing-synthesis technology are often sold and appear on popular music charts in Japan [25]. In particular, *Hatsune Miku* [26] is the most popular software package based on VOCALOID and features a cute synthesized voice with an illustration of a cartoon girl. Although Hatsune Miku is a virtual singer, she has already performed in many 'live' concerts with human musicians.

When synthesizing singing voices, most research approaches have focused on *text-to-singing (lyrics-to-singing) synthesis* [10, 12, 13]. In the following, I introduce two other approaches, *speech-to-singing synthesis* [27] and *singing-to-singing synthesis* [22].

2.1. SingBySpeaking: Speech-to-singing synthesis

SingBySpeaking is a speech-to-singing synthesis system that can synthesize a converted singing voice when given a speaking voice reading the lyrics of a song and the song's musical score [27].

This system is based on the STRAIGHT speech manipulation system [28] and comprises three models controlling three acoustic features unique to singing voices: the fundamental frequency (F0), phoneme duration, and spectrum. Given the musical score and its

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

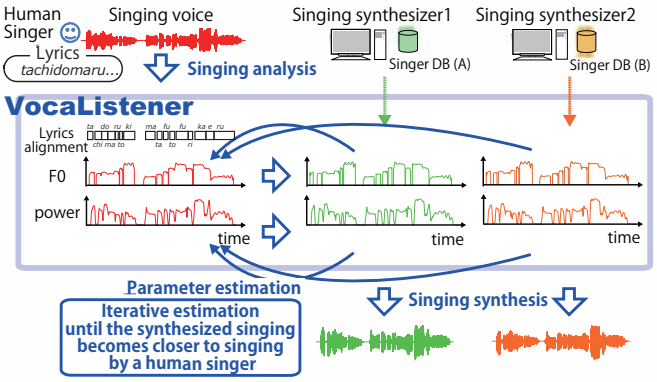

Fig. 1. VocaListener: Singing-to-singing synthesis [22, 30].

tempo, the F0 control model generates the natural F0 contour. The duration control model lengthens the duration of each phoneme by considering the duration of its musical note. The spectral control model controls both the singing formant [29] and the amplitude modulation of formants in synchronization with vibrato [27].

2.2. VocaListener: Singing-to-singing synthesis

VocaListener (Figure 1) is a singing synthesis system that imitates the pitch and dynamics of a target human singing [22, 30]. We call this approach *singing-to-singing synthesis*. 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 a commercial singing synthesizer based on Yamaha's VOCALOID technology [12], 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. The system also has functions to help modify the user's singing by correcting off-pitch phrases or changing vibrato. Temporal timbre changes of the target singing, however, are not imitated.

As an extension of VocaListener, we then developed VocaListener2 [31], 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 to synthesize imitated singing voices.

Furthermore, for a robot singer, we developed VocaWatcher [30, 32], 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 [33], 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.

3. LYRIC TRANSCRIPTION AND SYNCHRONIZATION

Considering that lyrics make up one of the most important elements of information conveyed by singing, there are various research approaches that aim to give computers the ability to understand lyrics in different ways.

Transcribing lyrics from musical audio signals — that is, automatically writing out the lyrics of an unknown song from the audio signals of that song — is a challenging research problem. This problem can be treated as the singing version of automatic speech recognition. When targeting singing, however, difficulties such as large fluctuations and musical accompaniment come into play. Because of these difficulties, high-performance lyric transcription applicable to singing that includes accompaniment has yet to be realized. However, since the ability to recognize lyrics could make a variety of compelling applications possible such as song retrieval by lyrics, research in this area is expected to accelerate in the years to come.

Examples of research targeting singing without musical accompaniment include an automatic music-transcription system with lyrics recognition for a solo singing voice [34], a robust speech-modeling method applicable to even high-pitch signals such as those in singing [35], query-by-humming music retrieval using not only pitch (melody) but also lyrics [36], and a method that exploits the fact that the lyrics of popular songs are often repeated in a chorus [37]. However, research on singing that includes musical accompaniment is rare, though research has been done on the recognition of phonemes in singing under the limited conditions of known phoneme boundaries [38]. Additionally, while not strictly lyrics recognition, research has been reported on indentifying the sung language [39].

If the text of lyrics is known in advance, we can take the lyric synchronization approach (i.e., lyrics-to-audio alignment approach). This means assigning a temporal correspondence between lyrics and musical audio signals of a song. For example, we can use an HMM-based forced-alignment method as is applied in speech recognition, but this suffers from difficulties unique to singing. When compared to lyric transcription, lyric synchronization presents a simpler problem so it can be achieved at a practical level of performance with relative ease and, except for some cases [40], it can target singing with musical accompaniment. Research into lyric synchronization targeting singing with accompaniment can be divided into two categories: that using no forced alignment [41–43] and that using forced alignment [44–48]. When textual chord information is additionally given, it can be used to improve the accuracy of lyric synchronization [49].

In the following, I introduce three lyric-based systems, *LyricSynchronizer*, *Hyperlinking Lyrics*, and *VocaRefiner*.

3.1. LyricSynchronizer: Automatic synchronization of lyrics with polyphonic music recordings

LyricSynchronizer (Figure 2) is a system that displays scrolling lyrics with the phrase currently being sung highlighted during playback of a song [47, 48]. Because the lyrics are automatically synchronized with the song, a user can easily follow the current playback position even on a small screen. Moreover, a user can click on a word in the lyrics shown on a screen to jump to and listen from that word.

Achieving this is difficult because most singing voices are accompanied by other musical instruments. It is therefore necessary to focus on the vocal part in polyphonic sound mixtures by reducing the influence of accompaniment sounds. To do this, the system first segregates the vocal melody from polyphonic sound mixtures, de-

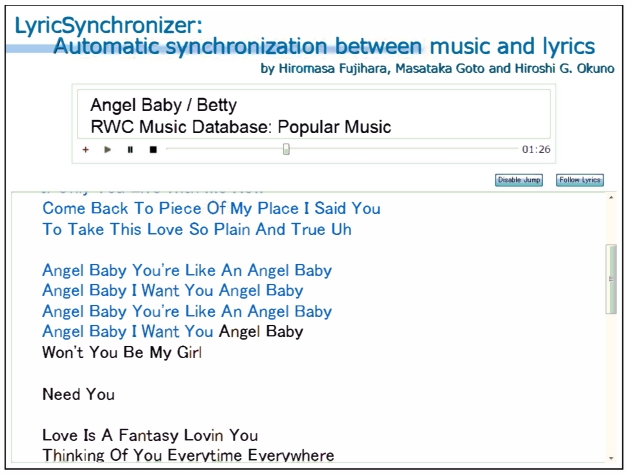

Fig. 2. LyricSynchronizer: Automatic synchronization of lyrics with polyphonic music recordings [47, 48].

tects vocal sections, and then applies the Viterbi alignment (forced alignment) method to those sections to locate each phoneme [47,48].

3.2. Hyperlinking Lyrics: Creating hyperlinks between phrases in song lyrics

Hyperlinking Lyrics is a system for creating a hyperlink from a phrase in the lyrics of a song to the same phrase in the lyrics of another song [50]. This can be used in various applications, such as song clustering based on the meaning of the lyrics and a music playback interface that will enable a user to browse and discover songs on the basis of lyrics.

Given a song database consisting of songs with their text lyrics and songs without their text lyrics, the system first extracts appropriate keywords (phrases) from the text lyrics without using audio signals. It then estimates the start and end times of these keywords in audio signals by using HMMs.

3.3. VocaRefiner: Interactive singing recording by integrating multiple singing recordings

VocaRefiner (Figure 3) is a system for enabling a singer to make a better singing recording by integrating multiple recordings of a song he or she has sung repeatedly [51]. It features a function called clickable lyrics, with which the singer can click a word in the displayed lyrics to start recording from that word as can be done on LyricSynchronizer. Clickable lyrics facilitate efficient multiple recordings because the singer can easily and quickly repeat recordings of a phrase until satisfied. Each of the recordings is automatically aligned to the music-synchronized lyrics for comparison by using a phonetic alignment technique.

VocaRefiner also features a function, called three-element decomposition, that analyzes each recording to decompose it into three essential elements: F0, power, and spectral envelope. This enables the singer to select good elements from different recordings and use them to synthesize a better recording by taking full advantage of the singer's ability. Pitch correction and time stretching are also supported.

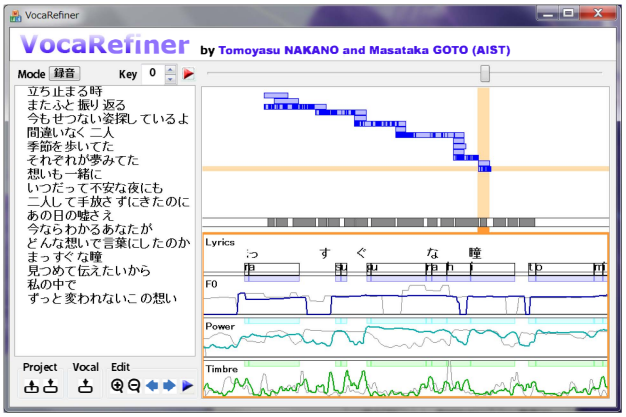

Fig. 3. VocaRefiner: Interactive singing recording by integrating multiple singing recordings [51].

4. VOCAL TIMBRE ANALYSIS

The timbre of a singing voice is an important element in singing not only because it can provide information as to who is actually singing but also because it can greatly affect the impression a person receives on listening to a singing voice [52].

Many studies on identifying the singer of a song have been made since the beginning of the 2000s [53–64]. These studies aim to specify the singer name for an input song from a set of singers registered beforehand in a database. This is generally accomplished by first extracting features and then applying a machine-learning technique. Some studies apply a vocal-activity detection method and use only the segments judged to have singing for identification purposes [54–56, 58, 61], other studies reduce the effects of musical accompaniment in addition to vocal-segment detection [57, 59, 63, 64], and still another study makes use of the individual characteristics of vibrato [60].

Methods that can estimate vocal-timbre similarity between singing voices in different songs — i.e., determine the extent to which a singing voice in one song resembles that in another song — have a wide range of applications. A variety of methods have been proposed, including one on the automatic clustering of songs based on similarity between singing voices [65] and another on vocal timbre analysis based on topic modeling using latent Dirichlet allocation (LDA), which can be used for cross-gender vocal-timbre similarity [66].

Furthermore, singing voice conversion that changes the timbre of a singing voice is also a popular research topic and has been studied [67–71].

In the following, I introduce two systems related to vocal timbre, *Singer ID* and *VocalFinder*.

4.1. Singer ID: Singer identification for polyphonic music recordings

The Singer ID system automatically identifies the name of the singer who sang the input song in the form of polyphonic musical audio signals [63, 64]. Even if the singer names for some songs are not available as metadata, the system enables users to retrieve those songs based on the singer names, for example. This is especially useful when artist names in the metadata are not singer names.

Like LyricSynchronizer, this system also segregates the vocal melody from polyphonic sound mixtures, and then selects frames

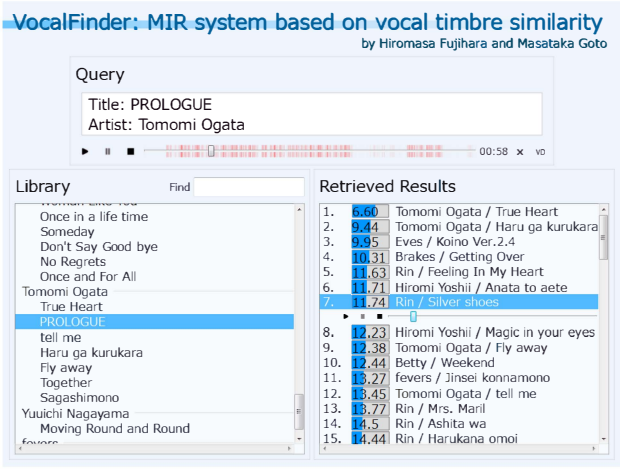

Fig. 4. VocalFinder: Music information retrieval based on singing voice timbre [64, 72].

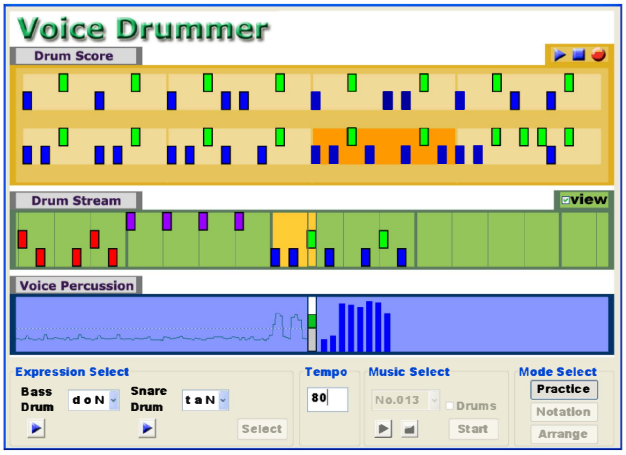

Fig. 5. Voice Drummer: Music notation of drums using vocal percussion input [89].

that are reliable enough for classification to improve the robustness. After training of a Gaussian mixture model (GMM) for each singer, the identity of the singer is determined on the basis of likelihood.

4.2. VocalFinder: Music information retrieval based on singing voice timbre

VocalFinder (Figure 4) is a music information retrieval system that can search a database for songs that have similar vocal timbres [64, 72]. Given a query song presented by a user, a list of songs with vocals having similar voice timbre to the query song is shown. With this system, we can find a song by using its musical content (i.e., vocal timbre) in addition to traditional bibliographic information.

To achieve this, we developed a method for extracting feature vectors that represent the characteristics of singing voices and calculating the vocal-timbre similarity between two songs by using the mutual information content of their feature vectors.

5. MUSIC INFORMATION RETRIEVAL BASED ON SINGING VOICES

Given the huge number of songs now available, music information retrieval is becoming increasingly important. The VocalFinder system described in Section 4.2 is an example of music information retrieval based on singing voices.

"Humming" can be treated as singing without lyrics, and music retrieval based on humming, or Query-By-Humming (QBH), is a method of retrieving songs using humming as a search key. The dawn of research on "content-based music information retrieval" using other than bibliographic information (such as titles and artist names) occurred in the 1990s [73–75], and it was this research that set the foundation for many studies on QBH (see [76–79] for details). A variety of QBH methods have been proposed, including a method for expressing the dynamic fluctuations of the F0 locus by a phase plane (F0 - ΔF0 plane) [80], a method for expressing F0 by probabilistic models without estimating F0 from the singing voice [81], and a method for improving retrieval performance by learning from user singing [82]. There have also been attempts at using features other than F0 for music retrieval, such as a method that uses the lyrics uttered by a singing voice [36] and a method that uses mel-frequency cepstral coefficients (MFCC) or formant frequencies [83].

Another style of music information retrieval is drum-pattern retrieval by "voice percussion" [84–89], which refers to a type of singing that expresses the sounds of a drum by using onomatopoeic words like "don-tan-do-do-tan" or that acoustically imitates those sounds in a more faithful manner. The latter is often called "beatboxing," and Query-By-Beatboxing is a method for recognizing the sounds so uttered and retrieving corresponding drum sounds or a drum pattern. In the following, I introduce a system for this sort of retrieval, *Voice Drummer*.

5.1. Voice Drummer: Music notation of drums using vocal percussion input

Voice Drummer (Figure 5) is a percussion instrument notation system that uses oral percussion patterns as input [89]. A user sings out a drum pattern (beatboxing), which is analyzed and matched with entries in a drum pattern database; this analysis and matching is based on onset timing patterns and intended drum types (bass or snare drums). As real-time feedback, the system shows the graphical score of recognized (retrieved) patterns. The user can also sing along to an existing musical piece without drums so that its drum patterns can be arranged according to the sung patterns.

The system uses onomatopoeia as internal representation of drum sounds, and retrieves a sung drum pattern from the pattern database by using HMMs. A pronunciation dictionary of onomatopoeic expressions of bass and snare drums is obtained from expression experiments.

6. SINGING SKILL EVALUATION

Singing skill plays an important role in characterizing singing, and clarification of the acoustic features that determine singing skill is expected to not only benefit singing instruction and training but to also find application in music information retrieval and singing synthesis.

Studies to reveal acoustic features related to singing skill have been done. Vibrato and the singer's formant are typical requirements for a good singing voice [90]. By investigating acoustic and perceptual effects of vocal training in amateur singing, it was found that two

kinds of F0 fluctuations (vibrato and overshoot) and singer's formant were changed by vocal training [91]. Conditions for "vocal ugliness" (unpleasant singing) have also been identified. These include irregular vibrato, pitch dissonance, weak fundamental amplitude in the voice source spectrum, and a lack of singer's formant [8]. The manner in which breaths are taken while singing and the positioning of breath sounds may also be associated with singing skill. Detailed studies of such an association have not yet been performed, but the acoustic analysis of breath sounds and their automatic detection have been studied [92, 93].

The most common approach for automatically evaluating singing skill is the karaoke scoring function. This popular function generally scores a singing voice based mainly on the difference between that voice and the musical score (pitch) of the song in question. Moreover, functions for detecting vibrato and portamento have been added. Research on automatic evaluation based on agreement of pitch and note duration has also been done [94]. In addition, studies on automatically evaluating singing skills without the use of musical scores have been made. There is an SVM-based method using acoustic features related to relative pitch and vibrato [95] and a GMM-based method using MFCC or mel-filter-bank coefficients as acoustic features [96]. Singing enthusiasm can also be evaluated by detecting related acoustic features [97].

In addition to the above automatic scoring function for karaoke, other applications related to singing skill are being considered, such as those for improving singing skill (in support of singing training). In particular, methods have been proposed for visualizing the results of user-singer analysis and providing feedback in real time [98–100]. It has been reported that pitch in singing can be improved by displaying the extent to which pitch in the user's singing departs from the target pitch [98].

In the following, I introduce two systems related to the above topics, *MiruSinger* and *Breath Detection*.

6.1. MiruSinger: Singing skill visualization and training

MiruSinger (Figure 6) is a singing skill visualization system that analyzes and visualizes vocal singing with reference to the vocal part of a target song that a user wants to sing more skillfully [100]. As real-time feedback, the system visualizes the characteristics of singing skills, such as F0 (the fundamental frequency) and vibrato sections of the user's singing voice, showing comparison with the F0 trajectory of the vocal part estimated in polyphonic sound mixtures.

Each vibrato is detected on the basis of our method developed for automatic singing skill evaluation for unknown melodies, in which a sung phrase can be categorized into good or poor classes by using an SVM [95]. The vocal melody of the target song is also estimated in polyphonic sound mixtures.

6.2. Breath Detection: Automatic detection of breath sounds in unaccompanied singing voice

Our automatic breath detection system finds each breath sound in unaccompanied solo singing [93]. Detected breath sounds can be suppressed as noise, or can be used as valuable cues for applications such as the segmentation and structural analysis of music and the evaluation of a singer's skill.

The system uses HMMs with MFCC, ΔMFCC, and Δpower as acoustic features to detect breath sounds as variant time-length events. We also did a detailed acoustic analysis of breath sounds and found that the spectral envelopes of breath sounds remain similar within the same song, and their long-term average spectra have a notable spectral peak at about 1.6 kHz for male singers and 1.7 kHz for female singers [93].

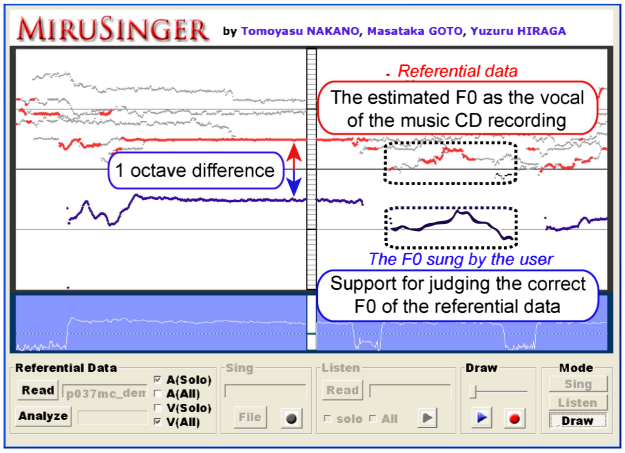

Fig. 6. MiruSinger: Singing skill visualization and training [100].

7. CONCLUSION

In this paper, I have provided an overview of the research field regarding singing information processing and have described examples of our singing information processing systems. It is important to realize, though, that these examples are just brief excerpts taken from a wide variety of systems and methods that have been developed by many researchers. We expect research related to singing information processing to progress rapidly in the years to come because every system and method available today still needs further refinement. A wide variety of research problems not discussed in this paper remain to be solved. It will be increasingly important that all kinds of knowledge concerning singing voices — such as psychology [101], physiology [102], and vocal pedagogy [103] — be considered in combination with signal processing, machine learning, human-computer interaction, and other key technologies. As this field of research continues to develop in the future, I expect to see an increasing number of researchers contributing to the advances realized.

Acknowledgments: I thank (in alphabetical order by surname) Hiromasa Fujihara Yuzuru Hiraga Tomoyasu Nakano, Hiroshi G. Okuno, and Takeshi Saitou, who have worked with me to build the systems presented in this paper. This work was supported in part by CREST, JST.

8. REFERENCES

[1] M. Goto and K. Hirata, "Invited review: Recent studies on music information processing," *Acoustical Science and Technology (edited by the Acoustical Society of Japan)*, vol. 25, no. 6, pp. 419–425, 2004.

[2] A. Klapuri and M. Davy, eds., *Signal Processing Methods for Music Transcription*. Springer, 2006.

[3] M. Goto, "Grand challenges in music information research," in *Dagstuhl Follow-Ups: Multimodal Music Processing* (M. Muller, M. Goto and M. Schedl, eds.), pp. 217–225, Dagstuhl Publishing, 2012.

[4] M. Goto, "Frontiers of music information research based on signal processing," in *Proc. of the 12th IEEE International Conference on Signal Processing (IEEE ICSP 2014)*, 2014.

[5] M. Goto, T. Saitou, T. Nakano and H. Fujihara, "Singing information processing based on singing voice modeling," in *Proc. of ICASSP 2010*, pp. 5506–5509, 2010.

[6] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Computer Music Journal*, vol. 20, no. 3, pp. 38–46, 1996.

[7] P. Depalle, G. Garcia and X. Rodet, "A virtual castrato," in *Proc. of ICMC '94*, pp. 357–360, 1994.

[8] J. Sundberg, "The KTH synthesis of singing," *Advances in Cognitive Psychology. Special issue on Music Performance*, vol. 2, no. 2-3, pp. 131–143, 2006.

[9] P. R. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing*. PhD thesis, Stanford University, 1991.

[10] 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.

[11] D. Schwarz, "Corpus-based concatenative synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 92–104, 2007.

[12] H. Kenmochi and H. Ohshita, "Vocaloid — commercial singing synthesizer based on sample concatenation," in *Proc. of Interspeech 2007*, pp. 4009–4010, 2007.

[13] K. Saino, H. Zen, Y. Nankaku, A. Lee and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proc. of Interspeech 2006*, pp. 1141–1144, 2006.

[14] K. Saino, M. Tachibana and H. Kenmochi, "A singing style modeling system for singing voice synthesizers," in *Proc. of Interspeech 2010*, pp. 2894–2897, 2010.

[15] T. Nose, M. Kanemoto, T. Koriyama and T. Kobayashi, "A style control technique for singing voice synthesis based on multiple-regression HSMM," in *Proc. of Interspeech 2013*, pp. 378–382, 2013.

[16] M. Umbert, J. Bonada and M. Blaauw, "Generating singing voice expression contours based on unit selection," in *Proc. of SMAC 2013*, 2013.

[17] H. Kawahara and H. Katayose, "Scat singing generation using a versatile speech manipulation system STRAIGHT," *The Journal of the Acoustical Society of America (The 141st Meeting of the Acoustical Society of America)*, vol. 109, no. 5, pp. 2425–2426, 2001.

[18] H. Kawahara, "Application and extensions of STRAIGHT-based morphing for singing voice manipulations based on vowel centred approach," in *Proc. of the 19th International Congress on Acoustics 2007 (ICA 2007)*, pp. 2018–2021, 2007.

[19] M. Morise, M. Onishi, H. Kawahara and H. Katayose, "v.morish'09: A morphing-based singing design interface for vocal melodies," in *Proc. of ICEC 2009*, pp. 185–190, 2009.

[20] H. Kawahara, M. Morise, H. Banno and V. G. Skuk, "Temporally variable multi-aspect N-way morphing based on interference-free speech representations," in *Proc. of APSIPA ASC 2013*, 2013.

[21] 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.

[22] 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.

[23] M. Hamasaki, H. Takeda and T. Nishimura, "Network analysis of massively collaborative creation of multimedia contents: Case study of Hatsune Miku videos on Nico Nico Douga," in *Proc. of uxTV'08*, pp. 165–168, 2008.

[24] Cabinet Office, Government of Japan, "Virtual idol," *Highlighting JAPAN through images*, vol. 2, no. 11, pp. 24–25, 2009. `http://www.gov-online.go.jp/pdf/hlj_img/vol_0020et/24-25.pdf`.

[25] H. Kenmochi, "VOCALOID and Hatsune Miku phenomenon in japan," in *Proc. of the First Interdisciplinary Workshop on Singing Voice (InterSinging 2010)*, pp. 1–4, 2010.

[26] Crypton Future Media, "What is the HATSUNE MIKU movement?." `http://www.crypton.co.jp/download/pdf/info_miku_e.pdf`, 2008.

[27] T. Saitou, M. Goto, M. Unoki and M. Akagi, "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.

[28] H. Kawahara, I. Masuda-Kasuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[29] J. Sundberg, *The Science of Singing Voice*. Northern Illinois University Press, 1987.

[30] M. Goto, T. Nakano, S. Kajita, Y. Matsusaka, S. Nakaoka and K. Yokoi, "VocaListener and VocaWatcher: Imitating a human singer by using signal processing," in *Proc. of ICASSP 2012*, pp. 5393–5396, 2012.

[31] 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.

[32] S. Kajita, T. Nakano, M. Goto, Y. Matsusaka, S. Nakaoka and K. Yokoi, "VocaWatcher: Natural singing motion generator for a humanoid robot," in *Proc. of IROS 2011*, 2011.

[33] K. Kaneko, F. Kanehiro, M. Morisawa, K. Miura, S. Nakaoka and S. Kajita, "Cybernetic Human HRP-4C," in *Proc. of Humanoids 2009*, pp. 7–14, 2009.

[34] C.-K. Wang, R.-Y. Lyu and Y.-C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," in *Proc. of Eurospeech 2003*, pp. 1197–1200, 2003.

[35] A. Sasou, M. Goto, S. Hayamizu and K. Tanaka, "An autoregressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition," in *Proc. of ICASSP 2005*, pp. I–237–240, 2005.

[36] M. Suzuki, T. Hosoya, A. Ito and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[37] M. McVicar, D. P. Ellis and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. of ICASSP 2014*, pp. 3141–3145, 2014.

[38] M. Gruhne, K. Schmidt and C. Dittmar, "Phoneme recognition in popular music," in *Proc. of ISMIR 2007*, pp. 369–370, 2007.

[39] W.-H. Tsai and H.-M. Wang, "Automatic identification of the sung language in popular music recordings," *J. New Music Res.*, vol. 36, no. 2, pp. 105–114, 2007.

[40] A. Loscos, P. Cano and J. Bonada, "Low-delay singing voice alignment to text," in *Proc. of ICMC 99*, 1999.

[41] C. H. Wong, W. M. Szeto and K. H. Wong, "Automatic lyrics alignment for cantonese popular music," *Multimedia Systems*, vol. 4-5, no. 12, pp. 307–323, 2007.

[42] M. Müller, F. Kurth, D. Damm, C. Fremerey and M. Clausen, "Lyrics-based audio retrieval and multimodal navigation in music collections," in *Proc. of ECDL 2007*, pp. 112–123, 2007.

[43] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 338–349, 2008.

[44] K. Chen, S. Gao, Y. Zhu and Q. Sun, "Popular song and lyrics synchronization and its application to music information retrieval," in *Proc. of MMCN'06*, 2006.

[45] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *Proc. of ISM 2006*, pp. 257–264, 2006.

[46] D. Iskandar, Y. Wang, M.-Y. Kan and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proc. ACM Multimedia 2006*, pp. 659–662, 2006.

[47] H. Fujihara and M. Goto, "Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection," in *Proc. of ICASSP 2008*, 2008.

[48] H. Fujihara, M. Goto, J. Ogata and H. G. Okuno, "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[49] M. Mauch, H. Fujihara and M. Goto, "Integrating additional chord information into HMM-based lyrics-to-audio alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 200–210, 2012.

[50] H. Fujihara, M. Goto and J. Ogata, "Hyperlinking Lyrics: A method for creating hyperlinks between phrases in song lyrics," in *Proc. of ISMIR 2008*, pp. 281–286, 2008.

[51] T. Nakano and M. Goto, "VocaRefiner: An interactive singing recording system with integration of multiple singing recordings," in *Proc. of SMC 2013*, pp. 115–122, 2013.

[52] A. Kanato, T. Nakano, M. Goto and H. Kikuchi, "An automatic singing impression estimation method using factor analysis and multiple regression," in *Proc. of ICMC SMC 2014*, pp. 1244–1251, 2014.

[53] B. Whitman, G. Flake and S. Lawrence, "Artist detection in music with minnowmatch," in *Proc. of NNSP 2001*, pp. 559–568, 2001.

[54] A. L. Berenzweig, D. P. W. Ellis and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proc. of AES-22 Intl. Conf. on Virt., Synth., and Ent. Audio*, 2002.

[55] Y. E. Kim and B. Whitman, "Singer identificatin in popular music recordings using voice coding features," in *Proc. of ISMIR 2002*, pp. 164–169, 2002.

[56] T. Zhang, "Automatic singer identification," in *Proc. of ICME 2003*, vol. I, pp. 33–36, 2003.

[57] M. A. Bartsch, *Automatic Singer Identification in Polyphonic Music*. PhD thesis, The University of Michigan, 2004.

[58] N. C. Maddage, C. Xu and Y. Wang, "Singer identification based on vocal and instrumental models," in *Proc. of ICPR'04*, vol. 2, pp. 375–378, 2004.

[59] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 330–341, 2006.

[60] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 519–530, 2007.

[61] J. Shen, B. Cui, J. Shepherd and K.-L. Tan, "Towards efficient automated singer identification in large music databases," in *Proc. of SIGIR'06*, pp. 59–66, 2006.

[62] A. Mesaros, T. Virtanen and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. of ISMIR 2007*, 2007.

[63] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. of ISMIR 2005*, pp. 329–336, 2005.

[64] H. Fujihara, M. Goto, T. Kitahara and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, 2010.

[65] W.-H. Tsai, H.-M. Wang, D. Rodgers, S.-S. Cheng and H.-M. Yu, "Blind clustering of popular music recordings based on singer voice characteristics," in *Proc. of ISMIR 2003*, pp. 167–173, 2003.

[66] T. Nakano, K. Yoshii and M. Goto, "Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity," in *Proc. of ICASSP 2014*, pp. 5239–5343, 2014.

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

[68] H. Doi, T. Toda, T. Nakano, M. Goto and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," in *Proc. of APSIPA ASC 2012*, 2012.

[69] H. Doi, T. Toda, T. Nakano, M. Goto and S. Nakamura, "Evaluation of a singing voice conversion method based on many-to-many eigenvoice conversion," in *Proc. of Interspeech 2013*, pp. 1067–1071, 2013.

[70] K. Kobayashi, H. Doi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti and S. Nakamura, "An investigation of acoustic features for singing voice conversion based on perceptual age," in *Proc. of Interspeech 2013*, pp. 1057–1061, 2013.

[71] K. Kobayashi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti and S. Nakamura, "Regression approaches to perceptual age control in singing voice conversion," in *Proc. of ICASSP 2014*, pp. 7954–7958, 2014.

[72] H. Fujihara and M. Goto, "A music information retrieval system based on singing voice timbre," in *Proc. of ISMIR 2007*, pp. 467–470, 2007.

[73] T. Kageyama, K. Mochizuki and Y. Takashima, "Melody retrieval with humming," in *Proc. of ICMC 93*, pp. 349–351, 1993.

[74] A. Ghias, J. Logan, D. Chamberlin and B. Smith, "Query by humming: Musical information retrieval in an audio database," in *Proc. of ACM Multimedia 1995*, vol. 95, pp. 231–236, 1995.

[75] T. Sonoda, M. Goto and Y. Muraoka, "A WWW-based melody retrieval system," in *Proceedings of ICMC 98*, pp. 349–352, 1998.

[76] R. Dannenberg, W. Birmingham, B. Pardo, C. Meek, N. Hu and G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the musart testbed," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 687–701, 2007.

[77] E. Unal, E. Chew, P. Georgiou and S. Narayanan, "Challenging uncertainty in query by humming systems: A fingerprinting approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 359–371, 2008.

[78] J.-S. R. Jang and H.-R. Lee, "A general framework of progressive filtering and its application to query by singing/humming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 350–358, 2008.

[79] W.-H. Tsai, Y.-M. Tu and C.-H. Ma, "An FFT-based fast melody comparison method for query-by-singing/humming systems," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2285–2291, 2012.

[80] Y. Ohishi, M. Goto, K. Ito and K. Takeda, "A stochastic representation of the dynamics of sung melody," in *Proc. of ISMIR 2007*, pp. 371–372, 2007.

[81] M. Suzuki, T. Ichikawa, A. Ito and S. Makino, "Novel tonal feature and statistical user modeling for query-by-humming," *Journal of Information Processing*, vol. 17, pp. 95–105, 2009.

[82] D. Little, D. Raffensperger and B. Pardo, "A query by humming system that learns from experience," in *Proc. of ISMIR 2007*, pp. 335–338, 2007.

[83] A. Duda, A. Nurnberger and S. Stober, "Towards query by singing / humming on audio databases," in *Proc. of ISMIR 2007*, pp. 331–334, 2007.

[84] O. Gillet and G. Richard, "Drum loops retrieval from spoken queries," *Journal of Intelligent Information Systems*, vol. 24, no. 2–3, pp. 159–177, 2005.

[85] O. Gillet and G. Richard, "Indexing and querying drum loops databases," in *Proc. of CBMI 2005*, 2005.

[86] A. Kapur, M. Benning and G. Tzanetakis, "Query-by-beatboxing: Music retrieval for the dj," in *Proc. of ISMIR 2004*, pp. 170–177, 2004.

[87] A. Hazan, "Towards automatic transcription of expressive oral percussive performances," in *Proc. of IUI 2005*, pp. 296–298, 2005.

[88] E. Sinyor, C. McKay, R. Fiebrink, D. McEnnis and I. Fujinaga, "Beatbox classification using ace," in *Proc. of ISMIR 2005*, pp. 672–675, 2005.

[89] T. Nakano, M. Goto, J. Ogata and Y. Hiraga, "Voice Drummer: A music notation interface of drum sounds using voice percussion input," in *Proc. of UIST 2005 (Demos)*, pp. 49–50, 2005.

[90] W. T. Bartholomew, "A physical definition of "good voice-quality" in the male voice," *J. Acoust. Soc. Am.*, vol. 55, pp. 838–844, 1934.

[91] T. Saitou and M. Goto, "Acoustic and perceptual effects of vocal training in amateur male singing," in *Proc. of Interspeech 2009*, pp. 832–835, 2009.

[92] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 838–850, 2007.

[93] T. Nakano, J. Ogata, M. Goto and Y. Hiraga, "Analysis and automatic detection of breath sounds in unaccompanied singing voice," in *Proc. of ICMPC 2008*, pp. 387–390, 2008.

[94] P. Lal, "A comparison of singing evaluation algorithms," in *Proc. of Interspeech 2006*, pp. 2298–2301, 2006.

[95] T. Nakano, M. Goto and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. of Interspeech 2006*, pp. 1706–1709, 2006.

[96] P. Prasert, K. Iwano and S. Furui, "An automatic singing voice evaluation method for voice training systems," in *Proceeding of the 2008 Spring Meeting of the Acoustical Society of Japan*, pp. 911–912, 2008.

[97] R. Daido, S. Hahm, Ito, S. Makino and A. Ito, "A system for evaluating singing enthusiasm for karaoke," in *Proc. of ISMIR 2011*, pp. 31–36, 2011.

[98] D. M. Howard and G. F. Welch, "Microcomputer-based singing ability assessment and development," *Applied Acoustics*, vol. 27, pp. 89–102, 1989.

[99] D. Hoppe, M. Sadakata and P. Desain, "Development of real-time visual feedback assistance in singing training: a review," *Journal of Computer Assisted Learning*, vol. 22, pp. 308–316, 2006.

[100] T. Nakano, M. Goto and Y. Hiraga, "MiruSinger: A singing skill visualization interface using real-time feedback and music CD recordings as referential data," in *Proc. of ISM 2007 Workshops (Demonstrations)*, pp. 75–76, 2007.

[101] D. Deutsch, ed., *The Psychology of Music*. Academic Press, 1982.

[102] I. R. Titze, *Principles of Voice Production*. The National Center for Voice and Speech, 2000.

[103] F. Husler and Y. Rodd-Marling, *Singing: The Physical Nature of the Vocal Organ. A Guide to the Unlocking of the Singing Voice*. Hutchinson & Co, 1965.