# SINGING INFORMATION PROCESSING BASED ON SINGING VOICE MODELING

*Masataka Goto, Takeshi Saitou, Tomoyasu Nakano, and Hiromasa Fujihara*

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, JAPAN <m.goto [at] aist.go.jp>

**ABSTRACT**

In this paper, we propose a novel area of research referred to as *singing information processing*. To shape the concept of this area, we first introduce singing understanding systems for synchronizing between vocal melody and corresponding lyrics, identifying the singer name, evaluating singing skills, creating hyperlinks between phrases in the lyrics of songs, and detecting breath sounds. We then introduce music information retrieval systems based on similarity of vocal melody timbre and vocal percussion, and singing synthesis systems. Common signal processing techniques for modeling singing voices that are used in these systems, such as techniques for extracting the vocal melody from polyphonic music recordings and modeling the lyrics by using phoneme HMMs for singing voices, are discussed.

*Index Terms*— Music, singing information processing, singing voice modeling, vocal melody

## 1. INTRODUCTION

As research on music information processing [1, 2, 3], including research on music information retrieval [4], has continued to rapidly expand, research activities related to singing have also become more vigorous. Such activities are attracting attention not only from a scientific point of view, but also from the standpoint of industrial applications. Singing-related research is highly diverse, ranging from basic research on the features unique to singing to applied research such as that on the synthesis of singing voices, lyrics recognition, singer identification, retrieval of singing voices, and singing-skill evaluation. In this paper, we refer to this broad range of singing-related studies as *singing information processing* and introduce examples of these studies with the focus on signal processing techniques for modeling singing voices.

Singing possesses aspects of both speech and music, and there are many unsolved research problems from the viewpoint of either field. For example, singing voices generally fluctuate more than speaking voices, and musical accompaniment, which is closely interlinked with singing, is usually included at a relatively high volume. Because of these characteristics, the automatic recognition of singing is the most difficult class of speech recognition from a technical point of view. In fact, the automatic recognition of lyrics in vocals has not yet been fully achieved. Furthermore, from the viewpoint of music recognition and understanding, large fluctuations and variations in singing cause various difficulties compared to a similar analysis of musical instruments. Technically speaking, there are many difficult and deeply interesting problems in this regard. Similarly, in the research on singing synthesis, many problems still exist, since, in addition to conveying content in the form of language as in speaking, singing synthesis requires dynamic, complex, and expressive changes in the voice pitch, intensity, and timbre of singing. In this way, the study of singing information processing is a genuine frontier of science.

Moreover, while music is an important type of content from the viewpoints of industry and culture, singing is the most important element of music. The results of singing-related research should therefore have a major impact on society. In fact, singing processing techniques to correct pitch by signal processing is already being used on a routine basis in the production of commercial music (popular music, in particular). It has become an absolute necessity for correcting pitch at points in a song where the singer is less than skillful and for making corrections to achieve a desired effect. Song retrieval using singing voices is becoming practical, and a service that can return the name of a song in response to a melody that is sung or hummed by the user can be easily used on a mobile phone or on the web. The synthesis of singing voices has also been attracting attention in recent years [5], and songs created using singing-synthesis technology are now being posted in large numbers on video sharing services likes *Nico Nico Douga* in Japan and *YouTube* [6, 7]. Even compact discs featuring compilations of songs created using singing-synthesis technology are often sold in Japan. Other singing information processing techniques are also being applied, for example, to a function for evaluating (scoring) a person's singing in the *karaoke* industry. There are a great number of people who listen to music with a focus on singing, especially in the case of popular music, and we can expect a wide variety of singing-related applications to appear in the future.

In this paper, we 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, we first describe this research field of singing information processing by introducing several systems we have developed. We then discuss the techniques used in our systems to model singing voices.

## 2. SINGING INFORMATION PROCESSING SYSTEMS

Since the concept of singing information processing is broad and still emerging, we could provide many examples of various types of systems. The following subsections briefly explain nine singing information processing systems we have built. They are grouped into three categories: systems for listening to singing voices, systems for music information retrieval based on singing voices, and systems for singing synthesis.

### 2.1. Systems for Listening to Singing Voices

The following systems deal with important aspects for listening to or understanding singing voices, such as the lyrics of a song [8], the singer [9], singing skill [10], identical lyric phrases, and breath [11].

#### *2.1.1. LyricSynchronizer: Automatic synchronization of lyrics with polyphonic music recordings*

*LyricSynchronizer* (Figure 1) is a system that displays scrolling lyrics with the phrase currently being sung highlighted during playback of a song [12]. 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

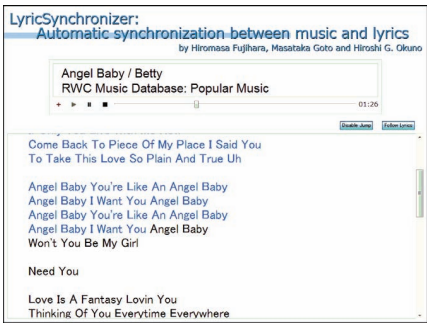


**Fig. 1**. LyricSynchronizer: Automatic synchronization of lyrics with polyphonic music recordings [12].

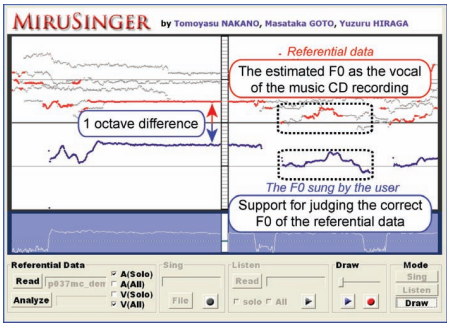


**Fig. 2**. MiruSinger: Singing skill visualization and training [14].

first segregates the vocal melody from polyphonic sound mixtures, detects vocal sections, and then applies the Viterbi alignment (forced alignment) technique to those sections to locate each phoneme [12].

#### *2.1.2. Singer ID: Singer identification for polyphonic music recordings*

Our Singer ID system automatically identifies the name of the singer who sang the input song in the form of polyphonic musical audio signals [13]. Even if 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 that are reliable enough for classification to improve the robustness [13]. After training a Gaussian mixture model (GMM) for each singer, the identity of the singer is determined on the basis of likelihood.

#### *2.1.3. MiruSinger: Singing skill visualization and training*

*MiruSinger* (Figure 2) 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 better [14]. 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 research on automatic singing skill evaluation for unknown melodies, in which a sung phrase can be categorized into good or poor classes by using a support vector machine (SVM) [15]. The vocal melody of the target song is also estimated in polyphonic sound mixtures.

#### *2.1.4. 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 [16]. 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 hidden Markov models (HMMs) [16].

#### *2.1.5. Breath Detection: Automatic detection of breath sounds in unaccompanied singing voice*

Our automatic breath detection system finds each breath sound in unaccompanied solo singing [17]. 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 the 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 [17].

### 2.2. Systems for Music Information Retrieval Based on Singing Voices

Traditional approaches for building music information retrieval systems are to use bibliographic information, such as titles and artist names. Because of the rapid and widespread diffusion of online or portable music, there is a great demand for content-based music information retrieval that can extract favorite songs from a large amount of music [18]. VocalFinder is an example of a system using such content-based retrieval based on singing voices. When retrieving drum patterns by using voices, vocal percussion is a useful input method [19].

#### *2.2.1. VocalFinder: Music information retrieval based on singing voice timbre*

*VocalFinder* (Figure 3) is a music information retrieval system that can search a database for songs that have similar vocal timbres [20]. 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 [20].

#### *2.2.2. Voice Drummer: Music notation of drums using vocal percussion input*

*Voice Drummer* (Figure 4) is a percussion instrument notation system that uses oral percussion patterns as input [21]. A user sings out a drum pattern (beatboxing), which is analyzed and matched with entries in a drum pattern database, 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

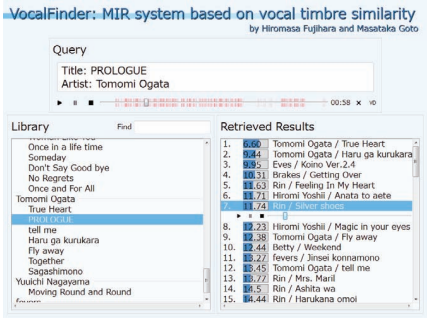


**Fig. 3**. VocalFinder: Music information retrieval based on singing voice timbre [20].

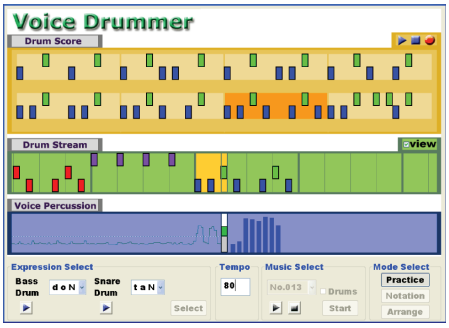


**Fig. 4**. Voice Drummer: Music notation of drums using vocal percussion input [21].

expressions of bass and snare drums is obtained from expression experiments [21].

### 2.3. Systems for Singing Synthesis

When synthesizing singing voices, most research approaches have focused on *text-to-singing (lyrics-to-singing) synthesis* [22, 23]. In the following, we introduce two systems based on two new approaches, *speech-to-singing synthesis* and *singing-to-singing synthesis*.

#### *2.3.1. SingBySpeaking: Speech-to-singing synthesis*

*SingBySpeaking*[1] 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 its musical score [24].

The system is based on the speech manipulation system STRAIGHT [25] 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 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 and the amplitude modulation of formants in synchronization with vibrato [24].

#### *2.3.2. VocaListener: Singing-to-singing synthesis*

*VocaListener*[2] is a singing synthesis system that automatically estimates singing synthesis parameters (pitch and dynamics) of commercial singing synthesizers by mimicking a user's singing voice [26]. Since a natural voice is provided by the user, the synthesized singing voice mimicking it can be human-like and natural without time-consuming manual adjustment. We call this approach *singing-to-singing synthesis*.

The system repeatedly updates singing synthesis parameters so that the synthesized singing can more closely mimic the user's singing. It supports a highly-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. Moreover, the system has functions to help modify the user's singing by correcting off-pitch phrases or changing vibrato [26].

## 3. TECHNIQUES FOR SINGING VOICE MODELING

Providing an important foundation for singing information processing, several common techniques for modeling singing voices are used in our systems that were briefly described in Section 2. The following subsections describe four key techniques.

### 3.1. Extracting Singing Voices in Polyphonic Music

To model singing voices from polyphonic sound mixtures, the extraction of vocal melody is important. We therefore use a predominant-F0 estimation method called *PreFEst* [27] to estimate F0 of the vocal melody. The harmonic structure of the vocal melody is also extracted and its corresponding audio signal of vocal melody is resynthesized by sinusoidal synthesis. This technique is used by five systems dealing with the polyphonic music input: LyricSynchronizer, Singer ID, MiruSinger, Hyperlinking Lyrics, and VocalFinder. More advanced techniques have also been developed [28, 29].

### 3.2. Modeling Lyrics

We use three different techniques to model the lyrics or phonemes in five systems: LyricSynchronizer, Hyperlinking Lyrics, Voice Drummer, SingBySpeaking, and VocaListener.

First, LyricSynchronizer and Hyperlinking Lyrics need to identify each phoneme of the lyrics or phrase in a polyphonic sound mixture. By using PreFEst, they first extract the vocal melody and then apply a vocal activity detection (VAD) technique using GMMs to remove non-vocal sections. They then perform forced (Viterbi) alignment between vocal vowels and a phoneme network of the lyrics or phrase by using phoneme HMMs. VocaListener also uses phoneme HMMs for the lyrics alignment. Using singing voices where each phoneme is hand-labeled, these phoneme HMMs specialized for singing voices were developed by adapting phoneme HMMs for speech recognition to singing voices [12] or training from scratch [16, 26].

Second, SingBySpeaking needs to identify each phoneme of the lyrics in monophonic speech voices. It therefore uses phoneme HMMs for the lyrics alignment, but those HMMs need not be specialized for singing voices [24].

Third, Voice Drummer uses HMMs to identify a drum pattern sung by a user, each of which corresponds to a different pattern in a drum pattern database. Onomatopoeic expressions for bass and snare drums are represented by HMMs [21].

### 3.3. Modeling Singer and Singing Voice Timbre

Two systems, Singer ID and VocalFinder, deal with singing voice timbre estimated from polyphonic sound mixtures. By using PreFEst, they extract the vocal melody to reduce the influence of accompaniment sounds, and then select frames that are reliable enough for identification or similarity calculation. In Singer ID, GMMs with LPC-derived mel cepstral coefficients (LPMCCs) are used to model the voice timbre for each singer [13]. In VocalFinder, GMMs with LPMCCs and $\Delta$F0 are used to calculate similarity between every pair of singers [20].

[1] Examples of synthesized singing are available at http://www.interspeech2007.org/Technical/synthesis_of_singing_challenge.php.

[2] A demonstration video including examples of synthesized singing is available at http://staff.aist.go.jp/t.nakano/VocaListener/.

The VAD technique mentioned in Section 3.2 can be considered as the modeling of vocal (singing voice) timbre. It uses both vocal and non-vocal GMMs, which are trained on feature vectors extracted from vocal and non-vocal sections of training data, respectively [12, 13, 20].

### 3.4. Modeling F0

The F0 of singing voices is modeled for purposes of both generation and analysis.

In SingBySpeaking, the F0 of the vocal melody is automatically generated by using the F0 control model. The generated F0 contour has two kinds of changes unique to singing: (a) global F0 changes that correspond to musical notes and (b) local F0 changes that include four types of F0 fluctuations: overshoot, vibrato, preparation, and fine fluctuation [30].

In VocalFinder, ΔF0 is used as an element of feature vectors for GMMs to model the dynamics of F0's trajectory. A singing voice tends to have temporal variations in its F0 and such temporal information is expected to express the singer's characteristics [20].

Vibrato is also detected by analyzing F0 to modify and improve the mimicked singing during vibrato in VocaListener [26], or to visualize each vibrato for training in MiruSinger [14].

## 4. CONCLUSION

We have described our singing information processing systems and the signal processing techniques behind those systems. Far from being a thing of the past, the era of singing-related research lives on. Indeed, we expect research related to singing information processing to progress rapidly in the years to come because every system and technique still needs further research. A wide variety of research problems not discussed in this paper remain to be solved. It will become increasingly important that all kinds of knowledge surrounding singing voices, such as psychology, physiology, and vocal pedagogy, be considered in combination with signal processing, machine learning, interface techniques, and other key techniques.

As explained in the introduction, singing possesses aspects of both speech and music. At present, the research fields of spoken language information processing and music information processing, while mutually influencing each other, do not have many points of contact. Looking forward, we aim to establish a field that can be called "audio information processing" in which speech and music are processed in a comprehensive manner instead of being treated as separate entities. Research on singing information processing is one solid approach to that end — and it might hold the key to success in this field. In the future, we believe that this research field will attract the interest of even more people as it continues to develop.

**ACKNOWLEDGMENTS:** This research was supported by CrestMuse, CREST, JST.

## 5. 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, "Active music listening interfaces based on signal processing," in *Proc. of ICASSP 2007*, 2007.

[4] M. Casey *et al.*, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

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

[6] 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.

[7] 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.

[8] M.-Y. Kan *et al.*, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 2, pp. 338–349, 2008.

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

[10] 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.

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

[12] 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.

[13] H. Fujihara *et al.*, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. of ISMIR 2005*, pp. 329–336, 2005.

[14] 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.

[15] 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.

[16] 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.

[17] T. Nakano *et al.*, "Analysis and automatic detection of breath sounds in unaccompanied singing voice," in *Proc. of ICMPC 2008*, pp. 387–390, 2008.

[18] M. Suzuki *et al.*, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

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

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

[21] T. Nakano *et al.*, "Voice Drummer: A music notation interface of drum sounds using voice percussion input," in *Proc. of UIST 2005 (Demos)*, pp. 49–50, 2005.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[28] H. Fujihara *et al.*, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and Viterbi search," in *Proc. of ICASSP 2006*, pp. V–253–256, 2006.

[29] H. Fujihara, M. Goto and H. G. Okuno, "A novel framework for recognizing phonemes of singing voice in polyphonic music," in *Proc. of WASPAA 2009*, 2009.

[30] T. Saitou, M. Unoki and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, no. 3–4, pp. 405–417, 2005.