Karthika Vijayan, Haizhou Li, and Tomoki Toda

Speech-to-Singing Voice Conversion

The challenges and strategies for improving vocal conversion processes

©ISTOCKPHOTO.COM/TRAFFIC_ANALYZER

Speech-to-singing (STS) conversion is the task of converting the read lyrics of a song, spoken in natural manner, to proper singing. The most important aspect of the task is to change the prosody of the natural speech to match with that of proper singing, while retaining the linguistic content and the speaker's identity. STS conversion is a challenging task because speaking and singing are different in many ways.

Introduction

STS conversion is an enabling technology for many innovative services and applications. It can be employed to beautify the singing renditions of amateur singers with inadequate singing skills, if we consider STS as an extreme scenario of converting bad singing to good-quality singing. It can be used to automatically generate reference singing for vocal learners and to personalize singing synthesis. As the state-of-the-art singing synthesis systems, such as Vocaloid [1] and Realivox [2], mostly generate singing vocals in a fixed voice, the idea of personalized singing is certainly appealing to the public, hence creating a strong commercially motivated drive. Besides its applications in the entertainment industry, STS conversion also serves as a bridge between speech and singing analyses. It provides valuable insights into the production and perception of speech and singing voices that are useful in many music information processing applications.

Though STS conversion is eagerly sought after, its realization is far from easy. Speech and singing vocals are produced by the same human voice production system, and hence, they share several similar characteristics. However, because of the distinctive vocal production processes between speaking and singing, they manifest through different acoustic characteristics [3]–[5]. STS conversion has to address several research problems including the temporal alignment between two very different signals, mapping of dissimilar acoustic characteristics, preserving fine attributes of speakers during conversion, and so on. All these make STS conversion a challenging task.

In this article, we describe the major challenges that need to be overcome for effective STS conversion. We first present the fundamental differences between speech and singing voices. We

Digital Object Identifier 10.1109/MSP.2018.2875195
Date of publication: 24 December 2018

then look into the methodology of STS conversion by introducing two prominent technical frameworks, i.e., template- and model-based approaches. Later, we present the evaluation strategies to assess the quality of synthesized singing vocals using objective and subjective measures. Finally, we summarize the tools and resources currently available for STS conversion study, and we discuss some implementation issues and future directions.

Speech versus singing

The major challenges in STS conversion stem from the differences between speaking and singing. The human voice production system that generates speech and singing signals can be effectively described by the source-filter model [3]. A comparative study between speech and singing can be performed by analyzing the different characteristics of glottal excitation and the vocal tract system.

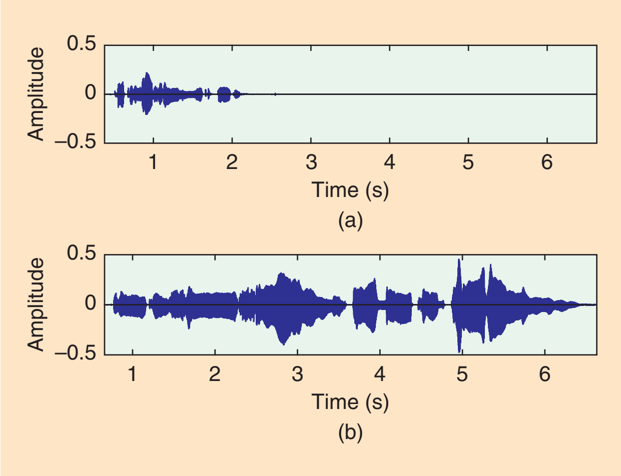

FIGURE 1. An illustration of the differences in vocal effort and dynamic ranges of amplitude between speech and singing vocals for “take a sad song and make it better,” uttered by the same person: (a) the speech vocal and (b) the singing vocal.

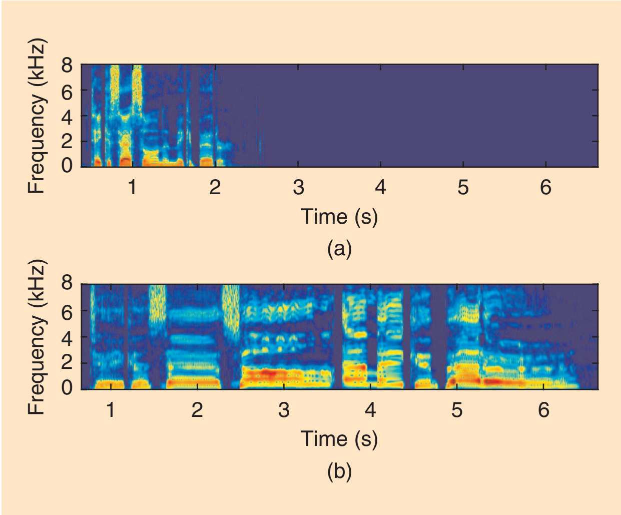

FIGURE 2. An illustration of the differences in spectral characteristics between speech and singing, corresponding to the utterances in Figure 1: (a) a spectrogram of the speech vocal and (b) a spectrogram of the singing vocal.

Singing requires a higher level of vocal effort, more active breathing during exhalation, and a larger range of variation in loudness than speaking. The dynamic range of short-time energy (instantaneous amplitude) of singing is thus larger than that of speech. This is achieved by trained singers with the careful management of subglottal pressure [6]. Figure 1 illustrates the difference between amplitudes of a speech and a singing sample.

For singing of high-level vocal effort, such as opera singing, a singer may not be able to rely only on the subglottal pressure variations because of physiological constraints. In such cases, a trained singer places his/her larynx in a particularly raised position by changing the form of articulators in the back end of the oral cavity, forcing the formants in the high-frequency section of the voice spectrum to cluster together. This peculiarly strong formant in the high-frequency spectrum of the singing voice is termed the *singing formant* [4]. The occurrence of the singing formant can be observed from Figure 2(b) (between 2 and 4 kHz from 2.5 to 3.5 s), which is absent in the corresponding speech spectrum in Figure 2(a). A trained singer, thus, manipulates the subglottal pressure with increased flexibility and introduces a singing formant efficiently to produce good-quality singing.

Smooth and soothing pitch variation is another integral factor of singing. A trained singer can carefully control the subglottal air pressure and volume to change the rate of vibrations of the vocal folds at the glottis [5]. A fine variation of pitch rendering adds expression to the singing, e.g., vibrato, preparation, and overshoot. Vibrato is manifested as quasi-periodic frequency modulations in pitch contour, overshoot is a deflection exceeding the target note observed after a note change and preparation is a deflection occurring just before a note change in the direction opposite to that of the note change [7]. Aided by the melody of singing, a singer produces a smooth fundamental frequency (F0) contour (pitch contour) with a larger dynamic range than that of speech signals, as can be observed in Figure 3(a) and (b). The fine characteristics of pitch of singing vocals are illustrated in Figure 3(b), as vibrato (between 2.5 and 3.5 s), overshoot (around 1.2 s), and preparation (around 1.5 and around 4 s). The particulars of pitch in singing affect the amplitudes and frequencies of formants as well, which results in the modulation of formants by the F0 contour [5]. Finally, singing follows a specific rhythm and melody by sustaining the vowels over the required duration that speaking does not require.

In terms of signal generation, singing has a close affinity to speech. They are both produced through the human voice production system, sharing many common acoustic properties. Therefore, the singing voice processing techniques have benefited from the recent advances in speech processing. The major differences between speaking and singing are manifested in characteristics of glottal excitation (subglottal pressure, pitch, and strength of excitation), the vocal tract system (singing formant), coupling between the vocal tract and excitation (modulation of formants by pitch), and the duration of voicing [3]. Such differences call for studies to adapt the parameterization and acoustic modeling methods from the speech voice to the singing voice. In short, the singing voice presents a unique set of research problems that deserve systematic study.

STS conversion

An STS conversion system is designed to take the read speech from a user as input (user speech) and generate prosody characteristics (singing prosody) for synthesis of singing vocals as output (synthesized singing). The basic idea of STS conversion includes the manipulation of parameters of user speech with respect to the reference prosody of the song, according to some predesigned transformation schemes. The transformed parameters that resemble those of singing are then used to generate output singing vocals. Because singing adheres to specific rhythm and melody, sustained notes, and vibrato, it is described by a structured prosody pattern. An important aspect of STS conversion is to transform the prosody of user speech into that of singing. Another equally important aspect is the preservation of the user's speaker identity, such as spectral characteristics, during the transformation.

The STS conversion techniques can be broadly classified into two categories, i.e., the template-based conversion and the model-based conversion, depending on how the reference prosody of singing is generated. The template-based framework uses reference prosody as a template that is extracted from high-quality singing. The parameters of user speech are then converted to those of singing vocals, using learned mapping schemes [8], [9]. On the other hand, the model-based framework generates singing prosody via prosody control models that are built on reference prosody described by musical scores, such as Musical Instrument Digital Interface (MIDI), and prior knowledge, such as musical pitch transitions and vibrato. With the mapping schemes or models, the parameters of user speech are then converted to those of singing vocals [10], [11]. In Figure 4, we illustrate the two general frameworks.

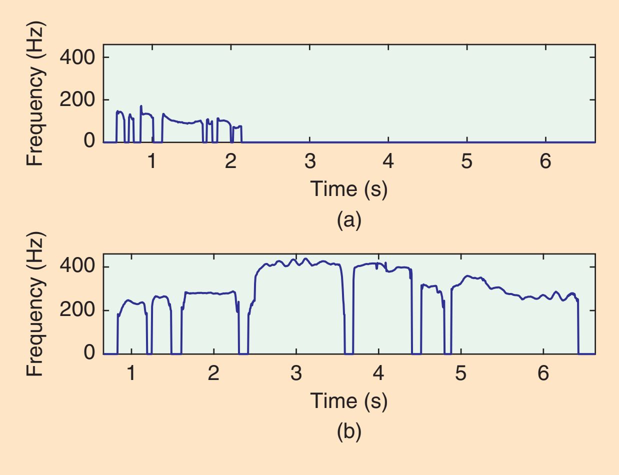

FIGURE 3. An illustration of the differences in the F0 contour between speech and singing, corresponding to the utterances in Figure 1: (a) the F0 contour of the speech vocal and (b) the F0 contour of the singing vocal.

The two conversion frameworks share a similar workflow. We convert the acoustic parameters from user speech to singing, then we synthesize the singing with a vocoder. The major difference between the two frameworks lies in the way they generate the singing prosody for synthesized singing from the reference. Once we have generated the intended singing prosody, the two frameworks face several common challenges. As illustrated in Figure 4, first we need to align the user speech to the musical rhythm and melody temporally [10]–[12]. We then map the phonetic content from user speech to synthesized singing without losing the speaker's identity. Finally, singing is a performing

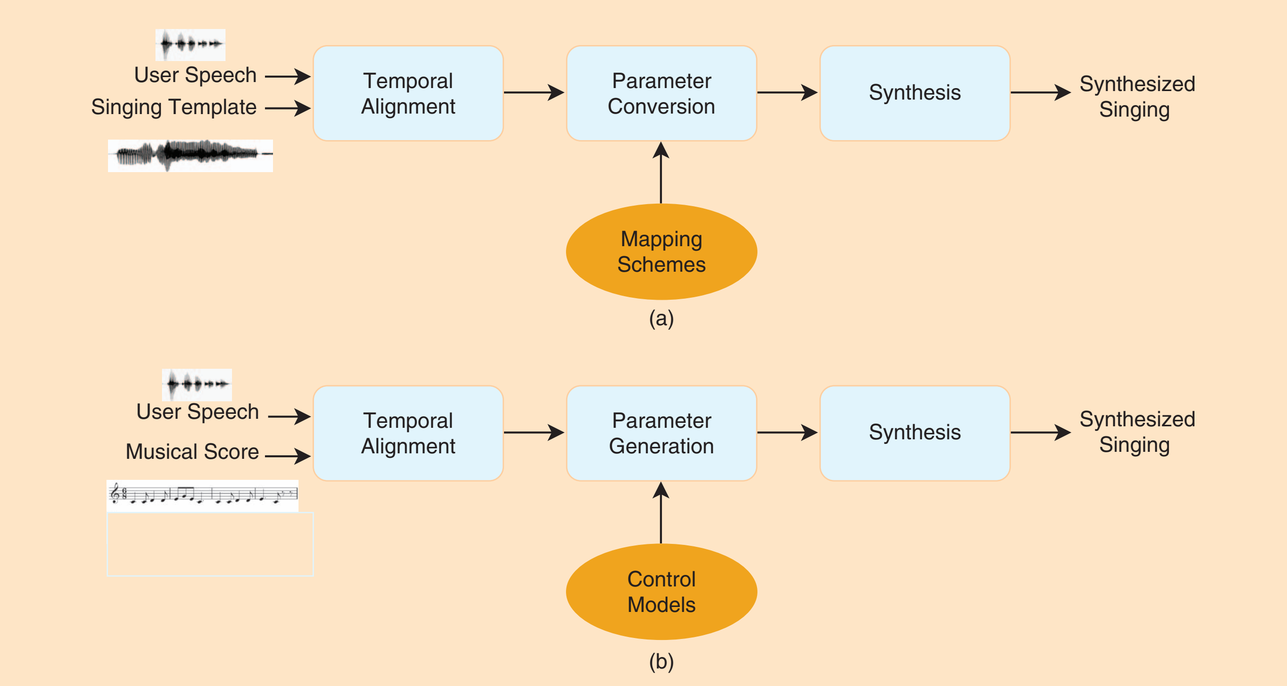

FIGURE 4. The (a) template- and (b) model-based frameworks for STS conversion.

art and a creative process that express an emotion, feeling, or taste that may not exist in the user's speech. The study of STS conversion is also about how to enable such expressions. Next, we discuss the fundamentals of the two conversion frameworks in detail.

Template-based STS conversion

The template-based approach assumes that a well-sung vocal is available. It uses the prosody derived from a natural singing vocal as the reference, thus minimizing the possible prosody errors in the synthesized singing. An important processing in the template-based approach is to synchronize the words between the user's speech and the singing template in time, which we call temporal alignment. The synchronization information obtained from the temporal alignment are needed for the subsequent frame-level parameter conversion.

Temporal alignment

Aligning linguistic content between the user's speech and the singing template is a crucial first step toward an accurate STS conversion because errors in temporal alignment are perceivable as distortions in synthesized singing. While manual alignment offers a high accuracy in general, it is not practical in real-time applications. Dynamic time warping is an effective algorithm for time-series alignment. The challenge is how to overcome the mismatch between speech and singing [13].

A dual alignment scheme was recently studied for effective speech to singing alignment [9], [13]. The dual alignment utilizes the read speech of lyrics of songs by the trained singer (singer's speech) to build a bridge between the user's speech and the singing template. Notice that we only need to manually align the singing template with the singer's speech once, which offers near perfect accuracy [13]. In dual alignment, the user's speech is first aligned with the singer's speech using dynamic time warping. Once such speech-to-speech alignment is obtained, the STS alignment can be established automatically. The problem of temporal alignment between the user's speech and the singing template is thus reduced to the alignment between two sets of speech signals via the user's speech and singer's speech.

To accurately align the user's speech to a singing template, we benefit from the understanding of speech and singing characteristics. The major differences between speech and singing are constituted by the properties of the glottal excitation source and the singing formant [4], [5]. It is advantageous to extract features that only represent the common properties between speech and singing, such as the voice activity contour and the lexical pause, and to remove their differences. Because the speech and singing share the same linguistic content, one can explore the use of multiple features, which we call *tandem features*, to represent their common characteristics. The analysis of signals for extraction of tandem features may include the following steps:

1) normalizing the short-time energy over the signals to nullify energy variations
2) performing source-filter decomposition to obtain smoothed spectral envelope, thus removing the glottal excitation source characteristics
3) restricting the smoothed spectrum to the low-frequency region to avoid the singing formant.

It has been observed that the temporal alignment performance of tandem features is superior to traditional features, such as the mel-frequency cepstral coefficients and the linear prediction cepstral coefficients [14].

Parameter extraction and conversion

Prominent signal analysis techniques, such as source-filter decomposition, can be employed to extract the parameters of the glottal excitation source and the vocal tract system from speech and singing signals. The parameters are then converted from the user's speech to the singing template and passed to the vocoder to generate time-domain signals. Because the analysis and conversion of the parameters of the signals are more effective in the spectral domain, we apply a short-time analysis to obtain the spectral parameters.

For the conversion of parameters from speech to singing, the characteristics of the excitation source and vocal tract system are considered separately. The properties of the glottal excitation source mostly contribute to the prosody and, hence, follow the melody of the song in singing voices. The properties of the vocal tract system, on the other hand, mostly characterize the timbre and provide valuable cues on speaker identity. The template-based framework extracts the F0 contour, representing excitation source parameters, from the singing template and retains them as the reference prosody for STS conversion. The smoothed spectral envelope, representing vocal tract system parameters, is extracted from the user's speech to preserve the speaker identity of the user in synthesized singing, except that these spectral characteristics have to be modified to resemble those of singing vocals. This is a particularly challenging task, involving speaker-dependent mapping of spectral characteristics from speaking to singing styles.

A simple solution to such speaker-dependent mappings can be provided by inducing the properties of the singing spectrum onto the speech spectrum. The most important spectral characteristics of singing voices are identified as the singing formant and the amplitude modulations of formant trajectories by the F0 contour [3]–[5]. The effect of the singing formant is introduced by emphasizing the speech spectrum around 3 kHz by a frequency-weighting function that resembles a bandpass filter. Also, the amplitude modulation in the temporal envelope of the speech signal at each vibrato in the F0 contour is estimated and added to the temporal envelope of the synthesized singing [10], [11]. The parameters for such manipulations can be obtained a priori empirically. Generally, a fixed set of control parameters do not provide a justified representation of a wide range of singing vocals in different genres. To estimate the parameters, other advanced techniques can be employed, such as partial least squares, Gaussian mixture models, exemplar-based representations, frequency warping, and deep learning [15]. While such techniques were studied for speaker identity conversion in general, they can be repurposed for voice conversion from speech to singing [16]. For training these voice conversion schemes, parallel databases of spoken and sung utterances that are temporally aligned at the frame level are generally required [17], [18].

To summarize, the converted parameters consist of the excitation source parameters from the singing template, and the spectral parameters are adjusted for singing from the user's speech. The source parameters represent the reference prosody of the song and are used as singing prosody for synthesized singing. The spectral parameters are converted from the user's speech, which carries forward the speaker's identity [8], [9], [13]. In Figure 5, we illustrate three spectrograms that are involved in the process. The smoothed spectral characteristics of the user's speech are preserved in synthesized singing, while the pitch harmonics resemble those in the singing template.

Model-based STS conversion

Instead of a singing template, the model-based framework uses musical scores, such as MIDI files, as the source of reference. The musical scores define accurate pitch transitions in the melody of singing vocals. Similar to the template-based conversion, the model-based conversion framework needs to align the user's speech to musical scores. The singing prosody and spectral parameters are generated from the musical score and the user's speech, respectively, by suitable control models [10], [11].

Temporal alignment

As a part of song writing, the lyrics writers align the lyrical words with the musical notes. Such alignment information is available for model-based STS conversion [10], [11]. Therefore, the actual task of temporal alignment is to align the user's speech to the lyrical words, either manually or with automatic speech recognition. Once the user's speech is aligned with the lyrical words, it is also aligned with the musical notes.

Parameter generation

The parameters from the user's speech can be extracted in the same way as that in the template-based framework. However, in the model-based framework, the excitation source parameters representing singing prosody (F0 contour and fine characteristics of pitch) are generated from the synthetic musical score.

The F0 contour for synthesized singing is generated from the reference prosody described by musical scores using an F0 control model. This control model transforms the unnaturally flat and discontinuous synthetic musical score into a human-realistic and continuous F0 contour, resembling that of natural singing vocals [19]. This can be achieved by inducing pitch variations found in natural singing onto the synthetic musical score [7]. The major characteristics of the F0 contour in singing voices are the gross representation of pitch in the melody and fine variations in the pitch that beautify the singing [5]. The fine attributes of the F0 contour include the vibrato, overshoot, preparation, and fine fluctuations [7]. With the F0 control model, we generate these fine attributes that are aligned with the musical scores to form a realistic F0 contour.

Vibrato is manifested as quasi-periodic frequency modulation in the F0 contour and can be generated using a second-order oscillatory system producing quasi-periodic sinusoidal variations. Overshoot and preparation are exhibited in the F0 contour as deflections associated with changes in musical notes. They are generated by second-order damping systems producing deflections of different polarities in the F0 contour [10], [11]. The parameters of these systems are estimated by least squares approximation of synthetic F0 contours with respect to natural F0 contours. The fine fluctuations in F0 are generated by high-pass filtering an amplitude-normalized white noise signal. The synthetically generated vibrato, overshoot, preparation, and fine fluctuations are added to the musical score to produce a human-realistic F0 contour that is later used as a singing prosody to generate synthesized singing outputs [10], [11].

Template-based versus model-based frameworks

The template-based and model-based frameworks mainly differ in the way they generate the singing prosody. The template-based framework employs the reference prosody extracted from the actual singing. On the other hand, the model-based framework generates the singing prosody from reference to musical scores using control models.

The template-based framework benefits from the near-perfect reference prosody that is directly obtained from the natural singing vocals. The reference prosody retains the pitch-rendering techniques in human singing such as vibrato, overshoot, and preparation. The model-based framework relies on the quality of the control models. In general, the control models generate singing prosody for output singing vocals that are less expressive than natural singing.

Yet, in practical implementations, recording high-quality reference singing templates for every song with variations, such

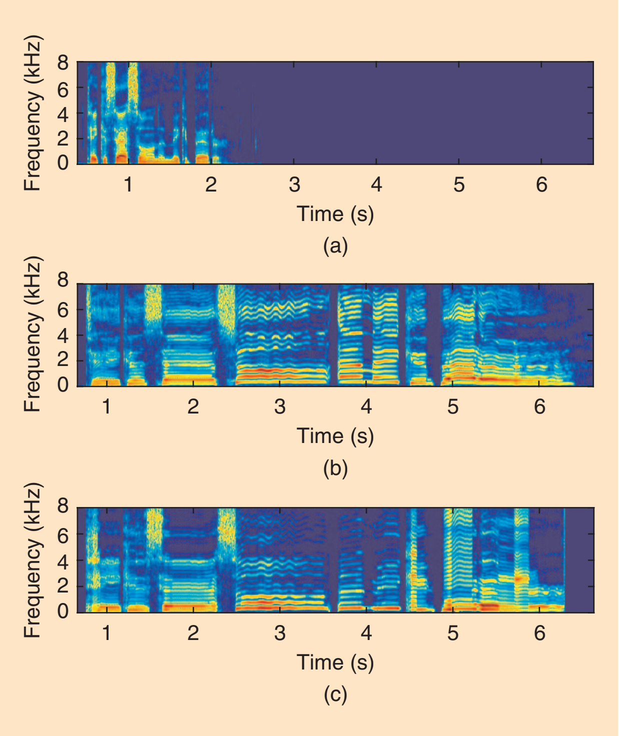

FIGURE 5. An illustration of the spectrograms in the template-based STS conversion: (a) the user's speech, (b) the singing template, and (c) the synthesized singing.

as classical singing versus Broadway-style singing or male versus female rendering, proves to be a tedious task. In contrast, preparing synthetic musical scores for a song will be relatively easier. Hence, the model-based framework is more scalable to a large song database in practical system deployment.

Evaluation of singing quality

The evaluation of perceptual quality of synthesized singing is necessary for the development of more sophisticated STS systems. The evaluation of singing quality is a research topic in itself. The perceptual quality of singing is generally evaluated using subjective tests in which listeners score the singing signals with respect to their intonation, rhythm, voice quality, and pronunciation. However, human listeners may not always be available. In such a scenario, objective evaluation strategies for the assessment of singing quality become extremely useful.

Objective evaluation

In the three processing modules of the processing pipeline in Figure 4, we evaluate the output of the modules against their respective ground truth in the objective evaluation. The evaluation of the synthesis module, also called the *vocoder*, is worth a separate study in speech synthesis [20], [21]. Here, we only discuss the evaluation of temporal alignment and parameter conversion.

The accuracy of temporal alignment between lyrical words of the user's speech and the target singing affects parameter conversion and, consequently, the perceptual quality of the synthesized singing. Let us take the template-based conversion as an example. We use the accuracy of temporal alignment between the user's speech and the singing template as one of the objective evaluation metrics.

The performance of temporal alignment can be reported over a database of parallel spoken-sung signals. The syllable-level and word-level manual transcriptions for each spoken and sung audio pair are required to set up the ground truth. The evaluation of temporal alignment between speech and singing signals can be reported using the average word boundary error, which is defined as the timing error between the ground truth and the aligned word boundaries [13]. A more detailed evaluation metric for temporal alignment can be defined as the timing error between the ground truth and the aligned syllable boundaries, which is termed the *average syllable boundary error*. To throw further insight into the effectiveness of a temporal alignment strategy, several other metrics can be defined, e.g., the percentage of gross alignment error or the percentage of syllable/word boundaries causing alignment errors of less than an acceptable error threshold.

The accuracy of parameter conversion is another contributing factor to the perceptual quality of synthesized singing. In parameter conversion, we convert the spectra of user speech to those of singing while preserving the speaker's identity. The resultant spectral characteristics represent the voice quality and pronunciation of the lyrics in synthesized singing. We may want to evaluate the parameter conversion in two aspects. One is to assess how well synthesized singing maintains the user's speaker identity. To do this, we can use techniques in speaker verification such as i-vector [22] and x-vector [23] to compare the speaker characteristics between the synthesized singing with the user's speech. Another factor is to assess the perceptual quality of the synthesized singing. A recent study on perceptual evaluation of singing quality (PESnQ) [24] suggests a systematic approach to the problem without the need of a reference singing. The PESnQ technique evaluates singing quality with a set of parameters covering pitch, rhythm, vibrato, voice quality, pronunciation, and volume, which provides objective evaluation close to human judgment.

Subjective evaluation

Assessing the perceptual quality of singing vocals can be done best by human evaluation. Upon availability of an expert panel of judges, samples of synthesized singing can be evaluated for their quality and various attributes. When an expert panel of judges is not available, the perceptual evaluation is conducted by collecting opinions of average listeners, who are music enthusiasts able to appreciate singing and notions of pitch, rhythm, and intonation. The common measures employed for perceptual evaluation of audio samples include the mean opinion score (MOS) and the best-worst score (BWS).

To provide an MOS, the listeners are asked to rate each synthesized sample according to a rating scale for assessing perceptual quality. They are typically asked to rate with respect to two factors:

1) how well the speaking user's identity is manifested in synthesized singing
2) how well the attributes of singing quality from the singing template are expressed in synthesized singing.

In the latter case, the listeners may be invited to evaluate different aspects of singing quality, e.g., voice quality, intonation, rhythm, intensity variations, and so on.

In many experiments, we have the need to compare among several synthesized singing samples. We use comparative statistics to rank the human preferences. For example, we use preference tests such as AB or ABX to identify detectable differences between the perceptual quality of samples. We know that humans are good at identifying the extremes, but their ability in ranking the preferences to fine details is limited. The BWS is a solution to provide comparative perceptual quality of samples. It was proposed for applications in economic surveys to evaluate the quality of products [25]. In STS conversion, the listeners are asked to choose the best and worst sounding samples from a set of audio signals after repeated listening to samples in all possible permutations [13]. If the sample i has appeared in many comparative groups, a BWS can be calculated for sample i by aggregating the statistics as, $\text{BWS}_i = (B_i - W_i)/N_i$, where B_i and W_i denote the number of times the item i is chosen as best and worst, respectively, by listeners. N_i represents the total number of appearances of item i in the set of trials [13], [25]. A more positive BWS indicates that the corresponding sample is more appealing to the listeners. The combination of MOS and BWS can reveal information about absolute and relative perceptual quality of singing samples synthesized by an STS conversion system.

Implementation issues

We have discussed the common frameworks for STS conversion and how to evaluate the quality of their outputs. In real-world applications, we also face many other technical challenges. We now discuss the issues concerning the input and the output of the system.

As the inputs to the system, the user's speech and/or the singing template are typically assumed to be pure vocals recorded in noise-free circumstances. However, the user's speech is often corrupted by noise, while the singing template may come with background music. In such cases, other signal processing modules such as speech enhancement and vocals–music separation are needed as a front-end application to the processing pipeline to prepare the pure vocals.

It is noted that a user's speech at runtime can be spontaneous or impromptu in a way that does not exactly follow the reference lyrical words. In such cases, the forced temporal alignment discussed in the section "STS Conversion" will fail. This calls for a smart temporal alignment algorithm that is able to strategically distribute the spoken content to the target song with the designated singing prosody for the best effect. In practice, spoken words can be obtained via an automatic speech recognition system, while singing rhythm information, such as the timing of notes, beats, and onsets, can be acquired from musical scores or via music information processing techniques. While the systems generate dry vocals, typically for final applications, the vocals are mixed with music in the same way that a recorded vocal would be mixed, i.e., processed further by audio production techniques, such as room reverberation [26], noise shaping [27], and audio equalization to improve the listening quality.

Tools and resources

Singing databases are necessary for the study of temporal alignment and transformation schemes. We report two databases having parallel recordings of read lyrics and singing vocals by trained singers. The National University of Singapore (NUS) sung and spoken lyrics corpus consists of 48 English songs of pop genre, read and sung by 12 singers. The manually marked phoneme labels for all recordings are also available as part of this database [17]. Similarly, the NUS–human language technology spoken lyrics and singing corpus contains 100 English pop songs, read and sung by ten singers. For the recordings of speech and singing, manually marked sentence transcriptions are also provided [18]. These databases can be used to study algorithms for temporal alignment and parameter conversion between speaking and singing. In addition, databases of singing vocals are needed for the training of transformation schemes, control models, and singing vocoders. Examples of such databases include the Smule database, the Center for Research in Entertainment and Learning database from the University of California, San Diego [31], the Basque database [32], the Isophonics singing voice data set [33], and the RAVDESS database from Science of Music, Auditory Research, and Technology laboratory at Ryerson University [34].

We also rely on computer-assisted processing tools and utilities. Wavesurfer and Praat are useful tools for analysis, visualization, and editing of audio signals. Plug-ins can be developed in these tools for automatic pitch tracking, which can be used for extraction of the F0 contour from the singing templates.

Effective parameterization of source and spectral properties is critical to the STS conversion task. This is achieved by an analysis–synthesis framework. During analysis, we obtain short-time frequency-specific parameters of the voice production system from speech or singing, i.e., the F0 contour and the smoothed spectrum. During synthesis, we reconstruct time-domain signals from these parameters. The speech transformation and representation by adaptive interpolation of weighted spectrogram (STRAIGHT) analysis represents an effective solution to the required analysis-synthesis [20]. It computes the smoothed spectral envelope (SP) representing the vocal tract system, and the F0 and the aperiodicity component (AP) representing the glottal excitation characteristics. There have been several alternatives to STRAIGHT. WORLD [21] presents a vocoding alternative that is computationally more efficient than STRAIGHT. A WaveNet vocoder trained on STRAIGHT parameters offers a more natural voice quality [28]. Note that a speaker-independent WaveNet vocoder involves training on a considerable amount of training samples. However, a speaker-adaptive WaveNet vocoder can offer a high-quality target voice with a quick adaptation process [29].

An iPhone application, named *Sing 4 Singapore*, was announced in 2014 [30] as the first near real-time STS conversion implementation that offers three English and two Chinese songs. This application allows a user to read the lyrics of songs, one line at a time. The application converts the read lyrics into the singing vocals and plays back the song in the user's voice together with the background music, as soon as the user finishes the reading [30].

Conclusions and future directions

In this article, we summarized the challenges in STS conversion that are created by the differences between speaking and singing. We presented the two major frameworks for STS conversion, which differ from each other in the way they generate the singing prosody. With the advent of STS conversion technology, we advocate that everyone can sing like a professional.

While STS conversion has made major progress recently, many research problems remain to be resolved. To improve the perceptual quality of synthesized singing, better temporal alignment and parameter conversion are expected. Furthermore, many applications may require real-time implementation of the algorithms that have not been well studied in the past.

The state-of-the-art STS conversion is mostly implemented as a modular system that consists of multiple signal processing modules in a processing pipeline. The recent studies on deep learning over large database have opened up opportunities in many ways. Inspired by the success in other signal processing pipelines, we consider that an end-to-end architecture for STS conversion will allow us to optimize the process in a systematic manner by reducing the artifacts introduced by the individual modules. We foresee that a deep-learning approach to STS conversion will become an active topic in the near future.

Authors

Karthika Vijayan (vijayan.karthika@nus.edu.sg) received her Ph.D. degree from the Indian Institute of Technology (IIT) Hyderabad, in 2016. She is a research fellow at the Department of Electrical and Computer Engineering, National University of Singapore. Her research interests include speech and singing signal processing and characterization. She is a member of the International Speech Communication Association and the Asia Pacific Signal and Information Processing Association (APSIPA). She received several awards including Research Excellence from IIT Hyderabad (2014 and 2015) and Springer book prize at the 2017 APSIPA–Annual Summit and Conference. She is a Member of the IEEE.

Haizhou Li (haizhou.li@nus.edu.sg) received his Ph.D. degree from South China University of Technology, Guangzhou, in 1990. He is a professor in the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include speech information processing and natural language processing. He is currently the editor-in-chief of *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2015–2018). He has served as the president of the International Speech Communication Association (2015–2017), and the president of Asia Pacific Signal and Information Processing Association (2015–2016). He is a Fellow of the IEEE.

Tomoki Toda (tomoki@icts.nagoya-u.ac.jp) received his D.E. degree from Nara Institute of Science and Technology, Japan, in 2003. He is a professor of the Information Technology Center at Nagoya University, Aichi, Japan. His research interests include speech, music, and sound processing. He has served as a member of the Speech and Language Technical Committee of the IEEE Signal Processing Society (SPS) (2007–2009, 2014–2016) and as an associate editor of *IEEE Signal Processing Letters* (2016–2018). He has received more than ten paper and achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 European Association for Signal Processing–International Speech Communication Association Best Paper Award (from *Speech Communication*). He is a Member of the IEEE.

References

[1] H. Kenmochi and H. Ohshita, "VOCALOID - commercial singing synthesizer based on sample concatenation," in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 4009–4010.

[2] Wikipedia. (2008). Realivox. [Online]. Available: https://en.wikipedia.org/wiki/Realivox

[3] B. Lindblom and J. Sundberg, "The human voice in speech and singing," in *Springer Handbook of Acoustics*, New York: Springer, Jan. 2014, pp. 703–746.

[4] J. Sundberg, "The level of the 'singing formant' and the source spectra of professional bass singers," *Q. Progress Status Report: STL-QPSR*, vol. 11, no. 4, pp. 21–39, Jan. 1970.

[5] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, P. F. MacNeilage, Ed. New York: Springer, 1983, pp. 39–55.

[6] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. America*, vol. 91, no. 5, pp. 2936–2946, May 1992.

[7] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 55–73, Nov. 2015.

[8] L. Cen, M. Dong, and P. Chan, "Segmentation of speech signals in template-based speech to singing conversion," in *Proc. APSIPA Annu. Summit and Conf.*, Xi'an, China, Oct. 2011.

[9] L. Cen, M. Dong, and P. Chan, "Template-based personalized singing voice synthesis," in *Proc. 2012 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4509–4512.

[10] 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. 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2007, pp. 215–218.

[11] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using STRAIGHT," in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 4005–4006.

[12] T. L. Nwe, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *Proc. 2010 IEEE Int. Conf. Multimedia and Expo*, Singapore, July 2010, pp. 1421–1426.

[13] K. Vijayan, M. Dong, and H. Li, "A dual alignment scheme for improved speech-to-singing voice conversion," in *Proc. APSIPA Annu. Summit and Conf.*, Kuala Lumpur, Malaysia, Dec. 2017, pp. 1547–1555.

[14] K. Vijayan, X. Gao, and H. Li, "Analysis of speech and singing signals for temporal alignment," in *Proc. APSIPA Annu. Summit and Conf.*, Honolulu, Hawaii, Dec. 2018.

[15] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[16] S. W. Lee, Z. Wu, M. Dong, X. Tian, and H. Li, "A comparative study of spectral transformation techniques for singing voice synthesis," in *Proc. INTERSPEECH*, Singapore, Sept. 2014, pp. 2499–2503.

[17] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. APSIPA Annu. Summit and Conf.*, Kaohsiung, Taiwan, Oct. 2013, pp. 1–9.

[18] X. Gao, B. Sisman, R. K. Das, and K. Vijayan, "NUS-HLT spoken lyrics and singing (SLS) corpus," in *Proc. Int. Conf. Orange Technologies (ICOT)*, Bali, Indonesia, Oct. 2018.

[19] S. W. Lee, S. T. Ang, M. Dong, and H. Li, "Generalized F0 modelling with absolute and relative pitch features for singing voice synthesis," in *Proc. 2012 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 429–432.

[20] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, Oct. 2011.

[21] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans.*, vol. E99-D, pp. 1877–1884, Jul. 2016.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[23] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. 2016 IEEE Spoken Language Technology Workshop (SLT)*, San Diego, California, Dec. 2016, pp. 165–170.

[24] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proc. APSIPA Annu. Summit and Conf.*, Kuala Lumpur, Malaysia, Dec. 2017, pp. 577–586.

[25] T. Flynn and A. Marley, *Best-Worst Scaling: Theory and Methods*. Cheltenham, UK: Edward Elgar Publishing, Inc., 2014.

[26] A. Tajadura-Jiménez, P. Larsson, A. Väljamäe, D. Västfjäll, and M. Kleiner, "When room size matters: Acoustic influences on emotional responses to sounds," *Emotion*, vol. 10, no. 3, pp. 416–422, 2010.

[27] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 7, pp. 1177–1184, July 2018.

[28] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proc. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, Dec. 2017, pp. 712–718.

[29] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *Proc. INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 1978–1982.

[30] M. Dong, S. W. Lee, H. Li, P. Chan, X. Peng, J. W. Ehnes, and D. Huang, "I2R speech2singing perfects everyone's singing," in *Proc. INTERSPEECH*, Singapore, Sept. 2014, pp. 2148–2149.

[31] Center for Research in Entertainment and Learning. (2008). Singing voice research database. [Online]. Available: http://crel.calit2.net/projects/databases/svdb

[32] X. Sarasola, E. Navas, D. Tavarez, D. Erro, I. Saratxaga, and I. Hernaez," A singing voice database in Basque for statistical singing synthesis of *bertsolaritza,"* in *Proc. Language Resources and Evaluation Conf. (LREC),* Portoroz, Slovenia, 2016, pp. 756–759.

[33] Isophonics. (2014). Singing voice audio dataset. [Online]. Available: http://isophonics.net/SingingVoiceDataset

[34] Science of Music, Auditory Research and Technology, Ryerson Univ. (2018). RAVDESS. [Online]. Available: https://smartlaboratory.org/ravdess/