SINGING EXPRESSION TRANSFER FROM ONE VOICE TO ANOTHER FOR A GIVEN SONG

Sangeon Yong, Juhan Nam

Graduate School of Culture Technology, KAIST
{koragon2, juhannam}@kaist.ac.kr

ABSTRACT

We present a vocal processing algorithm to automatically transfer singing expressions from one voice to another for a given song. Depending on singers' competence, a song can be rendered with great variations in terms of local tempo, pitch and dynamics. The proposed method temporally aligns a pair of singing voices using melodic and lyrical features that they have in common. Then, it conducts time-scale modification on the source voice according to the time-stretching ratio from the alignment result after smoothing. Once the two voices are aligned, the method modifies pitch and energy expressions of the source voice in a frame-by-frame manner using a pitch-synchronous overlap-add algorithm and a simple amplitude envelope matching. We designed our experiment to transfer singing expressions from a highly technical singer to a plain singer. The results show that our proposed method improves the singing quality effectively.

***Index Terms*—** singing voice, expression transfer, dynamic time warping, time-scale modification, phoneme classification, pitch-synchronous overlap-add

1. INTRODUCTION

Singing is a popular musical activity that many people enjoy, for example, in the form of karaoke. Depending on singing skills, a song can be rendered into touching music or just noisy sound. What if my bad singing can be transformed so that it sounds like a professional? In this paper, we present a vocal processing algorithm that automatically transfers singing expressions from one voice to another for a given piece of music.

Commercial vocal correction tools such as Autotune[1], VariAudio[2] and Melodyne[3] mainly focus on modifying pitch of singing voice. Some of them are capable of manipulating note onset timing or other musical expressions by editing transcribed MIDI notes. Although they provide automated controls to some degree, the correction process is often tedious and repetitive until satisfactory results are obtained.

There are some previous work that attempted to minimize the manual effort in modifying audio signals in musical expressions. Bryan et. al. proposed a variable-rate time-stretching method that allows users to modify the stretching ratio easily [1]. Given a user-guided stiffness curve, the method automatically computed time-dependent stretch rate via a constrained optimization program. Röebel et. al. proposed an algorithm to remove vibrato expressions [2]. They operated entirely based on spectral envelope smoothing without manipulation of individual partial parameters. While these methods provide more convenience in processing singing voice signals, they still require user guide or parametric control to some extent.

In this paper, we propose a method that modifies musical expressions of singing voice in a fully automatic manner using a target singing voice as a control guide. Assuming that both source and target voices sing the same song, the method transfers musical expressions from target to source in terms of tempo, pitch, and dynamics. Using reference recordings as a target to obtain expression parameters of singing voice has been previously attempted in singing voice morphing [3], singing voice synthesis [4], speech-to-singing and singing-to-speech conversion [5, 6], and vocal timbre adaption [7]. However, our method is distinguished from them in that it requires no additional information such as symbolic music scores and lyrics. Also, the method modifies only expressive elements in singing while preserving the timbre of source voice. We expect that the proposed vocal processing will be useful for not only sound production but also vocal training.

2. PROPOSED METHOD

Figure 1 illustrates the overview of the proposed singing expression transfer method. This section describes each of the processes that modifies tempo, pitch and dynamics in sequence.

This research is supported by Ministry of Culture, Sports, and Tourism and Korea Creative Content Agency in the Culture Technology Research & Development Program and by the Technology Innovation Program. No.10080667, funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea)

[1] http://www.antarestech.com/products/index.php

[2] https://www.steinberg.net/en/products/cubase/cubase_pro.html

[3] http://www.celemony.com/en/melodyne/what-is-melodyne

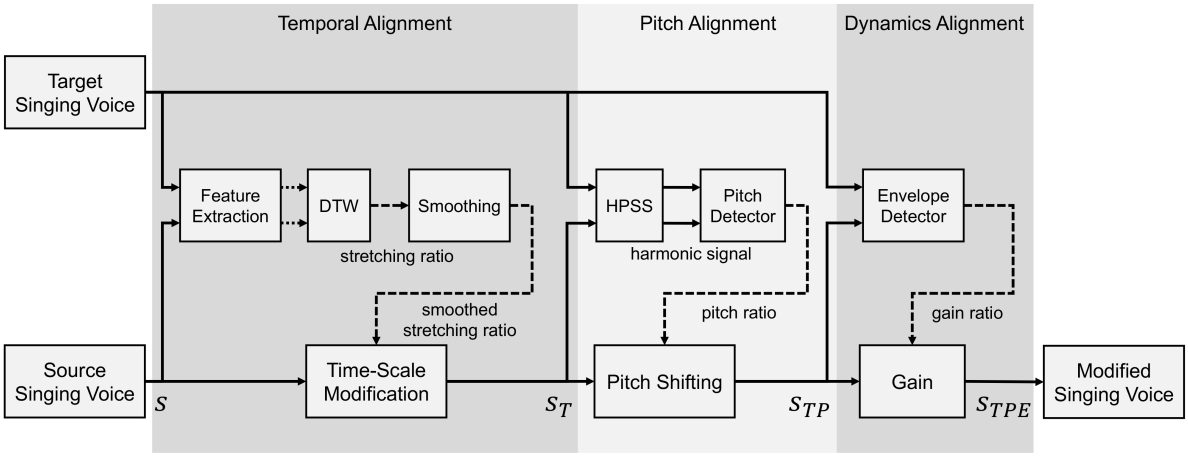

Fig. 1. Overview of the proposed singing expression transfer method

2.1. Feature Extraction For Temporal Alignment

The first step of the system is temporal alignment that synchronizes note timings between the two voices. This is actually the most important step because the subsequent steps relies on the aligned source for pitch and dynamics processing. We use dynamic time warping (DTW), a temporal alignment algorithm that is popularly used for music and audio data [8]. The issue here is what type of features will be used as input for the DTW.

Considering that the source and target voices are rendered from the same song, one straightforward approach is transcribing the audio signals into MIDI notes and use the melody notes for DTW [9]. However, the direct use of transcription results can be affected by the performance of the transcription algorithm. Another aspect to consider is exploiting the phonetic information from lyrics which is another common part in the two singing voices.

Our initial approach to embrace both melodic and lyrical features was simply using the spectrogram of two singing voices. The left-top in Figure 2 (a) shows the similarity matrix where each element was computed from cosine distance between every pair of the two magnitude spectra. While the alignment path returned from the DTW algorithm finds the onsets and offsets of notes quite successfully, it often failed to find a correct alignment path when either one voice has vibrato and pitch bending. For example, the alignment path has severe detour where the target voice has strong vibrato (in the range of 300 to 350 time frames). This detour causes audible artifacts when the system modifies the time scale of the source signal.

To solve the detour problem and improve the path accuracy, we propose to use two audio features that eliminate differences between two singers in musical expressions and timbre while preserving two common aspects, melody and lyrics. One is max-filtered constant-Q transform that handles the melodic aspect. Specifically, we used a constant-Q transform based on 88-band filterbanks, each of which is designed to cover one musical note with semi-tone resolution [10]. The max-filtering is applied to further alleviate pitch variation [2], particularly for the case that the two singing voices have more than one semi-tone in pitch difference, for example, by wrong note play or excessive pitch bending. The similarity matrix and alignment path in Figure 2 (b) show that the detour in the segment with strong vibrato becomes more diagonal.

The other feature is the phoneme score extracted from a phoneme classifier. This is meant to extract phonetic information in the lyrics while removing timbre difference between two voices. We used an open-source phoneme classifier that predicts frame-level phoneme probability distribution[4]. It uses 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC) with delta and double-delta as input feature and was trained with HTK speech recognition toolkit to predict the distribution of 39 phonemes as output. We used the output as a lyrical feature vector for temporal alignment. The similarity matrix and alignment path in Figure 2 (c) show that the phonetic feature helps alleviating the detour problem as well. Figure 2(d) shows the result when both melody and lyrics features are used. The alignment path is similar to that in Figure 2 (c) but it becomes even smoother.

2.2. Smoothing Time Stretch Ratio

Given the alignment path, we need to find a sequence of time-stretching ratios to apply them for a time-scale modification algorithm. Since the alignment path moves only three directions every frame, that is, upward, rightward, and diagonal direction in our setting, we need to smooth the path such that the stretching ratio is within a reasonable range. To this end, we use a Savitzky-Golay filter, an approximation method that fits a subset of sequence values with low-order polynomials in a convolutional manner [11]. Specifically, we applied 3rd-

[4]https://github.com/MLSpeech/AutoPhonemeClassifier

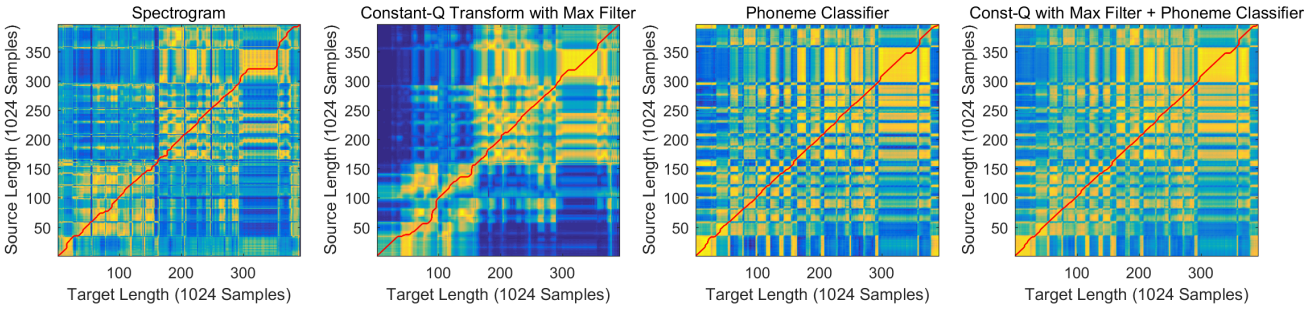

Fig. 2. Examples of similarity matrices between two singing voices with the same song but different audio features.

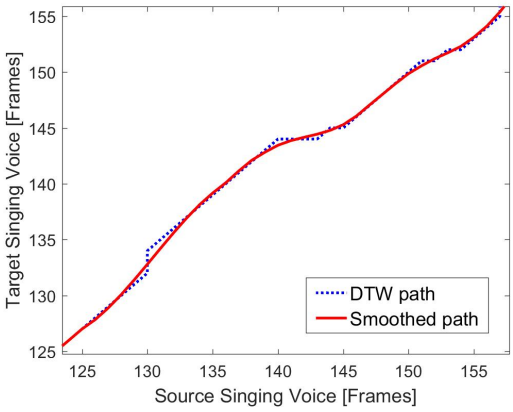

Fig. 3. A magnified view of the alignment path (blue) and filtered path by the Savitzky-Golay filter (red).

order Savitzky-Golay filter to the piece-wise linear alignment path using the function in MATLAB Signal Processing Toolbox[5]. The result after the smoothing is compared to the alignment path in Figure 3. To calculate the time-stretching rate α, we simply used the slope of the filtered path. Once we obtain the time-stretching ratio that varies each frame, we apply it to Time-Scale Modification (TSM) algorithm in order to temporally align the voices. Specifically, we used Waveform-Similarity based OverLap and Add (WSOLA) from the TSM Toolbox [12].

2.3. Pitch and Dynamics Alignment

Once the two signals are temporally aligned, the next step is transferring the pitch expressions from the target to the source. In order to modify pitch without timbre change, we used Pitch-Synchronous OverLap-Add (PSOLA) [13]. This algorithm requires pitch ratio, that is, the relative pitch change amount between the target and the source. In our case, we compute the pitch ratio β as follows:

$$\beta(i) = \begin{cases} f0_T(i)/f0_{S_T}(i) & \text{if } a_{S_T}(i) < 0.2 \\ 1 & otherwise \end{cases} \quad (1)$$

[5]https://www.mathworks.com/help/signal/ref/sgolayfilt.html

Table 1. The list of songs used for the experiment.

	Song1	**Song2**	**Song3**	**Song4**
Gender	female	male	male	male
No. of source	3	3	3	3
Remarks	high pitch	low pitch	swing rhythm	swing rhythm

where $f0_T(i)$ and $f0_{S_T}(i)$ denote the frame-level pitch sequence of the target and the source, respectively. $a_{S_T}(i)$ is the aperiodicity obtained from the source after temporal alignment. As Equation 1 indicates, we apply the pitch modification only for the segments that have strong periodicity. We used YIN algorithm [14] to extract the pitch of each voice. The algorithm returns the aperiodicity as a by-product. We also used harmonic-percussive source separation (HPSS) with median filter [15] to separate the harmonic signals from each of the voices before applying them to the pitch detector.

The final step is transferring dynamics from the target to the source. We conduct this by computing the frame-level amplitude gain between the two voices and multiplying it to the source voice. We used root-mean-square (RMS) value to extract the amplitude envelope from each voice and obtain the amplitude gain from the ratio of two amplitude envelopes.

3. EVALUATION

3.1. Datasets

We collected four recordings for each of four songs (total 16 recordings from different singers) for the experiment. One of the four recordings is a target singing voice from professional or those with proficient singing skills, and the rest are from ordinary singers. Since the ordinary singing voices are modified by taking musical expressions from the target, we have 12 pairs of singing voices (3 pairs for each song). They sang the songs while looking at the screen where the lyrics are displayed. The length of each song was about 10 seconds to 20 seconds and they were taken from the chorus part of the original songs. We chose the four songs so that they have different styles. Table 1 summarizes the dataset.

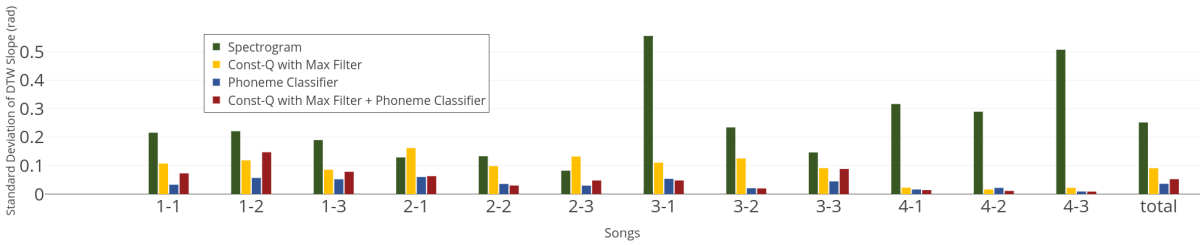

Fig. 4. Temporal alignment results: standard deviation of local slope from the diagonal via different audio features. The x-axis label (m-n) denotes the song number (m) and the pair number between the source and target voices (n).

3.2. Evaluation of Temporal Alignment

To evaluate the performance of the temporal alignment, we aligned the modified source voice, S_{TPE}, in Figure 1 with the target voice using DTW with spectrogram, and computed the standard deviation of local slope on the DTW path from one (the diagonal slope when they are perfectly aligned). Also, instead of using the local slope directly in calculating the standard deviation, we converted the slope with the arctangent function, $\theta = arctan(s)$, where s is the local slope from the path, so that the value (from 0 to infinity) is mapped to a finite range (from 0 to $\pi/2$ in radian).

Figure 4 compares the standard deviations of the local slope via different audio features. In general, the lyrical feature using the phoneme classifier is most reliable over all examples. This might be because the singers performed the songs with lyrics and so the phonetic features are quite accurate. The melodic feature using constant-Q transform with maximum filter also helped improving the alignment but it sometimes failed for the songs with low pitch (e.g. song 2-1 to 2-3). This might be because the pitch resolution in the low pitch range is not sufficiently high in the constant-Q transform. Combining the two features does not necessarily improve the results. It achieved best results for half of the examples but it yielded even worse results than the lyrical feature only for the other half.

3.3. Evaluation of Pitch and Dynamics Alignment

To evaluate the pitch and dynamics alignment, we computed the average of difference in pitch and dynamics. For pitch, the average pitch difference between the source and target is compared before and after the pitch alignment. We measured the pitch with YIN algorithm and counted only the segments that have strong periodicity (i.e. when the aperiodicity is less than 0.2). Figure 5 shows that the average pitch difference is reduced by 78.8% for total after the pitch alignment. For dynamics alignment, we computed the average of difference in the amplitude envelope. Specifically, we used Root-Mean-Square (RMS) value. Figure 6 shows that the average dynamics difference is reduced by 86.4% for total after the dynamics alignment.

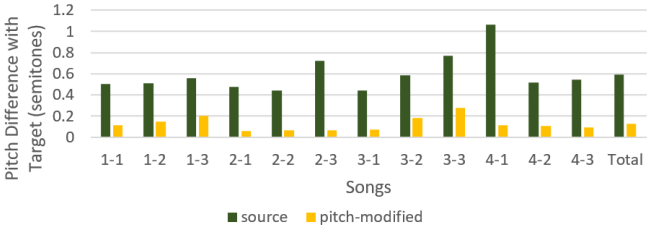

Fig. 5. Average differences in pitch between the source and target voices

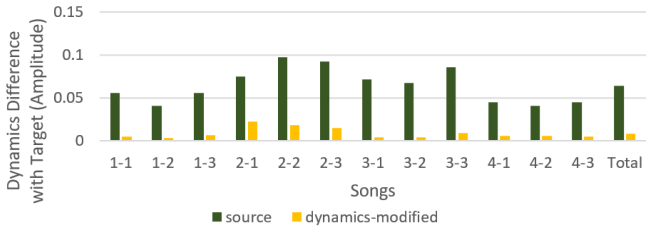

Fig. 6. Average difference in dynamics (in RMS) between the source and target voices.

While all of the alignment errors above provide some indications of how well the singing expressions are transferred, they are not perfect measures of how natural the modifications are. Actually, there are some audible artifacts by the limitations of time-scale modification and pitch shifting algorithms that we used. Examples of the results in this experiment are found at the link.[6]

4. CONCLUSION

We proposed a method to transfer vocal expressions from one voice to another in terms of tempo, pitch and dynamics. We suggested to use max-filtered constant-Q transform and the prediction distribution of phoneme classifier as melodic and lyrical features, respectively, for the temporal alignment. Once the voices are aligned, we modified pitch and dynamics according to the differences in pitch and amplitude envelope. From the experiment, we showed the proposed method effectively transformed the source voices so that they mimic singing skills from the target voice.

[6]https://seyong92.github.io/ICASSP2018

5. REFERENCES

[1] Nicholas J. Bryan, Jorge Herrera, and Ge Wang, "User-guided variable-rate time-stretching via stiffness control," in *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx)*, 2012.

[2] Sebastian Böck and Gerhard Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx)*, 2013.

[3] Pedro Cano, Alex Loscos, Jordi Bonada, Maarten de Boer, and Xavier Serra, "Voice morphing system for impersonating in karaoke applications," in *Proceedings of the International Computer Music Conference (ICMC)*, 2000.

[4] Tomoyasu Nakano and Masataka Goto, "Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation," in *Proceedings of the Sound and Music Computing Conference*, 2009, pp. 343–348.

[5] Takeshi Saitou, Masataka Goto, Masashi Unoki, and Masato Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[6] Shimpei Aso, Takeshi Saitou, Masataka Goto, Katsutoshi Itoyama, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Speakbysinging: Converting singing voices to speaking voices while retaining voice timbre," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, 2010.

[7] Matthew Roddy and Jacqueline Walker, "A method of morphing spectral envelopes of the singing voice for use with backing vocals," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx)*, 2014.

[8] Meinard Müller, "Fundamentals of music processing: Audio, analysis, algorithms, applications," in *Springer*, 2015.

[9] Roger B. Dannenberg, "An on-line algorithm for real-time accompaniment," in *International Computer Music Conference*, 1984, vol. 84.

[10] Meinard Müller and Sebastian Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011, to appear.

[11] Sophocles J. Orfanidis, *Introduction to Signal Processing*, Prentice-Hall, 1996.

[12] Jonathan Driedger and Menard Müller, "TSM toolbox: MATLAB implementations of time-scale modification algorithms," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx)*, 2014.

[13] F.J. Charpentier and M.G. Stella, "Diphone synthesis using an overlap-add technique for speech waveform concatenation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86*, 1986, vol. 11, pp. 2015–2018.

[14] Alain De Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," in *The Journal of the Acoustical Society of America*, 2002, vol. 111.4, pp. 1917–1930.

[15] Jonathan Driedger, Meinard Müller, and Sebastian Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," in *IEEE Signal Processing Letters*, 2014, vol. 21(1), pp. 105–109.