SINGING VOICE CORRECTION USING CANONICAL TIME WARPING

Yin-Jyun Luo[1], Ming-Tso Chen[2], Tai-Shih Chi[2], Li Su[1]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan

ABSTRACT

Expressive singing voice correction is an appealing but challenging problem. A robust time-warping algorithm which synchronizes two singing recordings can provide a promising solution. We thereby propose to address the problem by canonical time warping (CTW) which aligns amateur singing recordings to professional ones. A new pitch contour is generated given the alignment information, and a pitch-corrected singing is synthesized back through the vocoder. The objective evaluation shows that CTW is robust against pitch-shifting and time-stretching effects, and the subjective test demonstrates that CTW prevails the other methods including DTW and the commercial auto-tuning software. Finally, we demonstrate the applicability of the proposed method in a practical, real-world scenario.

Index Terms— Singing voice correction, singing voice synthesis, canonical time warping

1. INTRODUCTION

Synchronizing a music signal with another sequence of different modality, such as audio, [1] MIDI [2], sheet music [3], lyrics [4], or video [5], provides great possibilities in automatizing music annotation [6], interactive performance [7], education [8], and many other applications. In this paper, we focus on the use of audio synchronization algorithms to solve the *singing voice correction* problem. Our main focus of singing voice correction is to automatically modify the unsatisfactory singing of an amateur singer with wrong key and intonation into correct one, while preserve original timbre as much as possible. Here the 'unsatisfactory' singing, namely the *source*, is an audio signal, while the 'correct' singing, namely the *target*, can be a predefined music scale [9], a MIDI score of the same song [10], a phonetic sequence of the lyric [11], or a recording of the same song performed by a professional singer. Previous studies on this topic mostly propose rules or time-warping algorithms to shift the source pitch contour to fit a musical scale or a MIDI score, then use a vocoder to resynthesize the signal with the new pitch contour and the spectral envelope of the source [10]. For instance, intonation is adjusted locally using rule-based algorithms such as the dynamic pitch warping (DPW) [12]. Since the ground truth given here are deadpan notes, the outcomes of these techniques tend to lose the voice articulation such as the sliding and vibrato, which are important for human expression.

One way to achieve expressive singing voice correction is to align the source recording to a professional recording of the same song as the target, then modify the source pitch contour according to the alignment path. In this way, an amateur singer can get the professional singer's key, intonation, and even expression all at once. Seemingly a straightforward idea, such a task is however quite challenging because of not only the wrong pitch in the source, but also the timbre among various sources and targets. Furthermore, since a clean professional monophonic singing is seldom available, in practice one would need to consider the alignment from the user's input (typically monophonic singing) to a polyphonic target containing singing as well as accompaniment. To the best of our knowledge, no study has investigated singing voice correction in such a practical scenario.

Other studies on the alignment of singing voice, such as those for automatic singing voice assessment [13], melody similarity measurement [14], speech-to-singing synthesis [15], and an adaptive karaoke system [16] also face similar challenges caused by high-modality data, in which the conventional dynamic time warping (DTW) can not handle well. They usually focus on pertaining alignment accuracy using DTW on low-level features, but ignoring the expressive parts of the features. For example, the targets for alignment are dull singings without vibrato in [13]. In [14], the alignment is performed on automatically transcribed notes where the vibrato and ornamentation are also suppressed. Also in [16], it is reported that DTW is inferior in aligning clean singings to source-separated professional ones.

These issues motivate us to investigate the use of advanced time-warping algorithms, such as canonical time warping (CTW) [17], to solve the singing voice correction problem. CTW combines the canonical correlation analysis (CCA) and DTW, to offer more robustness and flexibility to cross-domain features than DTW. Although CTW has been used in the lyrics-to-audio alignment task [4], its application in aligning an amateur singing to a professional singing with or without accompaniment, is never investigated. To verify the potential of CTW in a more accurate way, we compiled a professional-singing dataset parallel to the amateur-singing
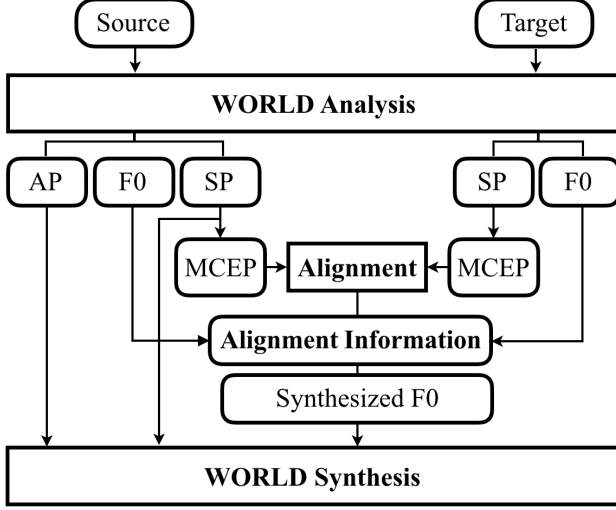
Fig. 1: The proposed system flowchart

MIR-1k dataset [18]. This simplifies the problem a bit to allow us to extract the target pitch contour from monophonic pitch tracking algorithm, while avoid melody tracking in polyphonic music.

The remaining part of this paper is organized as follows. The method and the proposed system are presented in Sec. 2. The new dataset is presented in Sec. 3. The experiments and results are detailed in 4. We conclude the paper in Sec. 5.

2. METHODOLOGY

Fig. 1 illustrates the system diagram with major building blocks including feature extraction, alignment, and synthesis.

2.1. Alignment

Given two time sequences, $\mathbf{X} \in \mathbb{R}^{d \times n_x}$ and $\mathbf{Y} \in \mathbb{R}^{d \times n_y}$, where d and n denote the feature and temporal dimension, respectively. A DTW algorithm aligns the samples in $\mathbf{X}$ and $\mathbf{Y}$ by optimizing the following cost:

$$\{\mathbf{W}_x, \mathbf{W}_y\} = \underset{\{\mathbf{W}_x, \mathbf{W}_y\}}{\arg\min} \|\mathbf{X}\mathbf{W}_x^T - \mathbf{Y}\mathbf{W}_y^T\|_F^2, \quad (1)$$

where $\mathbf{W_x} \in \{0,1\}^{m \times n_x}$ and $\mathbf{W_y} \in \{0,1\}^{m \times n_y}$ are binary-valued matrices representing the aligned indexes of the two sequences, and m is the length of the warping path. Despite the efficiency of the algorithm, DTW usually falls short of aligning two sequences with nonlinear distortion between them, e.g., an in-tune singing voice and an out-of-tune one. Though being drastically different, a pair of source-target singing could share common characteristics. We thereby investigate CTW, an advanced technique that combines CCA [19] and DTW, to offer feature selection and alignment at the same time [17]. The objective function of CTW therefore has

additional linear transformation compared to (1):

$$\{\mathbf{W}_x, \mathbf{W}_y\} = \underset{\{\mathbf{W}_x, \mathbf{W}_y\}}{\arg\min} \|\mathbf{V}_x^T\mathbf{X}\mathbf{W}_x^T - \mathbf{V}_y^T\mathbf{Y}\mathbf{W}_y^T\|_F^2, \quad (2)$$

where $\mathbf{V}_x \in \mathbb{R}^{d \times b}$ and $\mathbf{V}_y \in \mathbb{R}^{d \times b}$ are the projection matrices for $\mathbf{X}$ and $\mathbf{Y}$, respectively, and b is the dimension of the low-dimensional embedded space sought by CCA, where $b \leq d$. Inherited from CCA, multiple constraints are imposed for CTW, and the optimization rule can be found in [17]. In order to solve Eq. (2), we consider initializing $\mathbf{W}_x$ and $\mathbf{W}_y$ in two ways. One is with DTW by solving Eq. (1), and the other is by linearly scaling, as in [4], the temporal dimensions of the input sequences $\mathbf{X}$ and $\mathbf{Y}$ such that the two sequences share a common length $L = \max(n_x, n_y)$. In this paper, we denote CTW initialized with the two methods as CTW-dtw and CTW-uniform, respectively, which will be compared in Sec. 4.

2.2. Singing voice analysis and synthesis

In this work, we leverage alignment information to achieve expressive, yet efficient, singing voice pitch correction. To analyze audio and synthesize ones from manipulated pitch contours, we use WORLD [20] (D4C edition [21]) which is a vocoder-based speech analysis/synthesis system. It is shown to reduce computational cost without performance deterioration compared to STRAIGHT [22] which is widely used in the literature. WORLD extracts parameters including spectral envelope (SP), fundamental frequency (F0) and aperiodicity (AP), from input audio. In this study, 1024-point FFT and 5-ms frame shift are used to extract the parameters. As the representative feature for alignment, 24-ordered mel-cepstral coefficients (MCEP) are derived for every frame from SP. Such a setting is utilized because it is common in aligning pairs of speech and for synthesis purpose in voice conversion [23].

Note that in the subjective test, an optimal warping path is derived from either DTW or CTW given source and target audio features. Our goal is to synthesize a pitch-corrected singing with source SP, AP and modified pitch contour. We keep same the length of the modified and unmodified source pitch contour, by duplicating or averaging the F0 values in target pitch contour according to a derived warping path. SP is left untouched to preserve the timbre of source singing.

3. DATASET

For the purpose of this study, parallel songs that are sung by amateur/below-average and professional/above-average singers are necessary. Moreover, monophonic singing tracks are favored for better performance of an F0 extractor, on which synthesis quality depends. Such a parallel dataset is not within reach to our best knowledge. We therefore com-
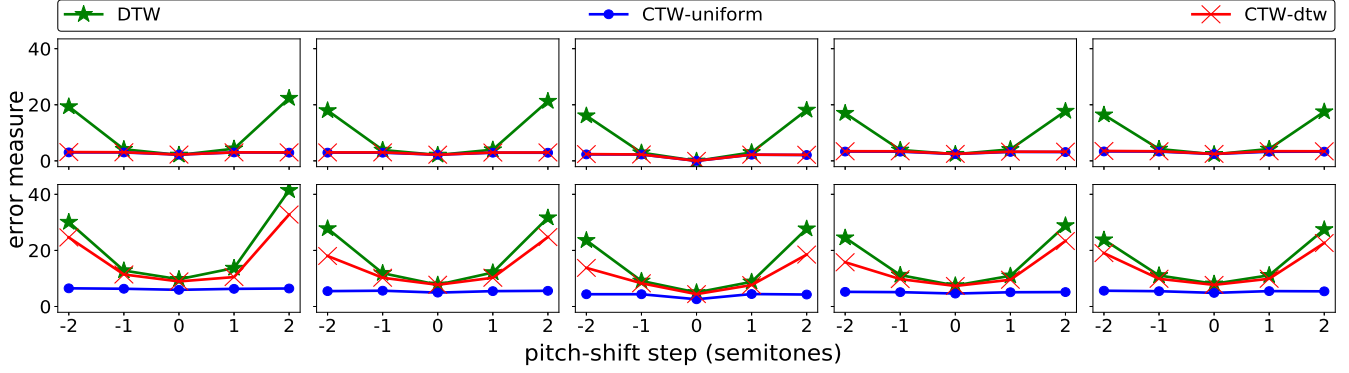
Fig. 2: The error measure e of linear time-stretch. Top panel: *mono-to-mono*; bottom panel: *mono-to-poly*. From left to right: time-stretch rate $r = 0.8, 0.9, 1.0, 1.1, 1.2$.

pile a new dataset based on the MIR-1k[1] dataset covered by amateurs as the source data, then recruit singers to build the target data. Although a more reasonable way is simply using the high-quality professional singings available online as our target data, we did not adopt this way because most of the state-of-the-art melody extraction algorithm from polyphonic music are still not that satisfactory for a subjective test.

Four female and five male singers were recruited to record the target singings of the dataset. Though not up to the professional level, they are with better singing skills than the ones in MIR-1k, and are therefore feasible to record the data to verify the performance of a singing voice correction algorithm. We picked 104 songs in MIR-1k and assigned around 10 to 30 songs to each singer depending on his or her acquaintance with the songs. The recruited singers were demanded to sing in-tune and temporally correct. The signing voice was then recorded independently as a mono-channel track. Though there were 216 recordings in total, we only selected three songs for the subjective test (see Sec. 4.2) to demonstrate the proposed system. On the other hand, 104 songs in the original MIR-1k are used in the objective test. For each song, only a segment of the first 30 seconds is used (see Sec. 4.1).

4. EXPERIMENTS AND RESULTS

4.1. Objective Test

The three time-warping algorithms, namely DTW, CTW-uniform and CTW-dtw, are tested under two tasks. The first one, named as the *mono-to-mono* task, is to align two different monophonic singing voice tracks, while the second one, named *mono-to-poly*, is a more practical and challenging task that aims to align a singing voice track to a mixture having both singing voice and accompaniment. Since there is no ground-truth annotation for the alignment path between a source and a target, we manage to generate sample pairs with ground truth in a synthetic way. Specifically, we employ time-
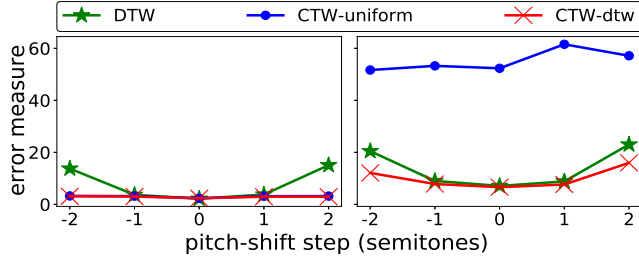
[1] the MIR-1k dataset is in https://sites.google.com/site/unvoicedsoundseparation/mir-1k.
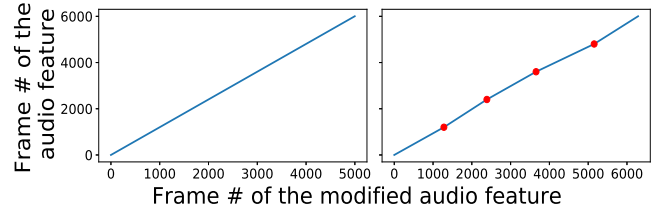
Fig. 3: Examples of ground-truth warping paths. Left: linear, $r = 1.2$; right: nonlinear, $r = 0.93, 1.08, 0.94, 0.80, 1.06$. The dots denote the chunk boundaries.

stretching and pitch-shifting effects with predetermined parameters on the 104 song in the original MIR-1k, and thereby obtain the sample pairs and known alignment path. Two scenarios are considered: in the *linear* time-stretching scenario, five time-stretching rates $r \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$ and five pitch-shifting steps $s \in \{-2, -1, 0, 1, 2\}$ are considered, resulting in 25 distinct parameter sets. In the *non-linear* time-stretching scenario, the same five pitch-shifting steps are also used, while every song is divided into 5 equal-length chunks, each assigned with a randomly sampled r from a uniform distribution $\mathcal{U}(0.8, 1.2)$ in order to provide irregular speed variation. Fig. 3 illustrates examples of ground-truth warping paths of the linear and nonlinear scenario. The time-stretching and pitch-shifting effects are implemented by `librosa` [24]. We evaluate the performance by an error measure e comparing the discrepancy between the estimated warping path and the ground-truth warping path; the definition of such an error measure can be found in [25]. The lower the error measure is, the better the system performs.

Fig. 2 and 4 shows the median of the error measure e over all songs in the testing dataset under each parameter set for the linear and nonlinear time-stretching scenario, respectively. Overall, CTW-based methods are more robust than DTW to the pitch-shifting effect, a nonlinear distortion that causes instability in the distance measure between features. It is worth noting that, in the mono-to-poly task, CTW-uniform outperforms the other two algorithms in the linear time-stretching scenario, while CTW-dtw outperforms the others

Fig. 4: The error measure e of nonlinear time-stretch. Left: *mono-to-mono*; right: *mono-to-poly*.

	Song	Source	DTW	CTW-u	Melo.
Pitch accuracy	1	1.23	2.16	2.58	2.06
	2	1.26	2.35	2.90	1.42
	3	1.34	2.64	2.45	1.80
	Avg.	1.28	2.38	2.65	1.76
Fluency	1	2.08	2.50	2.78	2.43
	2	2.23	2.37	2.72	2.01
	3	2.18	2.32	2.34	2.19
	Avg.	2.16	2.40	2.61	2.21

Table 1: Subjective test results with DTW, CTW-uniform (CTW-u) and Melodyne (Melo.) methods.

in the nonlinear ones. Moreover, CTW-uniform performs the worst under the nonlinear scenario. These results indicates the importance of initialization when performing CTW. Our results suggest that, when there is severe temporal discrepancy between two sequences, initialization with DTW can help stabilize the performance, while uniform initialization may force the CCA to project the features into a confounding subspace that simply makes the resulting features be aligned uniformly.

4.2. Subjective Test

For subjective test, we conduct an online questionnaire[2] containing three clips of amateur singing voice, each being modified with three pitch-corrected algorithms. The first two are the DTW and CTW-uniform with mono-to-mono setting, while the third one is the auto-tuning function of the commercial software Melodyne [9]. In Melodyne, we firstly make comparable the pitch level of source and target, and then each individual note is snapped to the musical scale of the original CD version. To prevent the participants from fatigue, the length of each testing clip is only 15 seconds. For each song, participants are required to listen to the amateur voice and its three corresponding pitch-corrected versions, and rate their pitch accuracy and singing fluency in a range from 1 to 5.

Results collected from 143 valid responses, including the average rating of pitch accuracy and singing fluency, are listed in Table 1. In average, DTW and CTW-uniform improve the pitch accuracy of the source singing voice by 1.10 and 1.37 points, respectively, while Melodyne improves it by 0.48 points. Moreover, CTW-uniform outperforms DTW significantly under the two-tailed t-test ($p < 0.01\%, d.f. = 283$). Both alignment methods outperform Melodyne, implying that the behaviors of an unsatisfactory singing, such as time-varying tonal drift and intonation drift, are hard to be corrected without an associated music score or original singing. Meanwhile, CTW-uniform also prevails in terms of singing fluency ($p < 0.1\%, d.f. = 283$), which implies that CTW-uniform is more likely to preserve human singing expression and is more natural. However, the results still show relatively

low ratings overall (i.e., even the best rating is lower than 3 points), probably because the targets recordings are also not as professional as the in-market ones, and are likely to leave an unsatisfactory impression to the participants. Besides, unlike the result of the objective test, DTW in the subjective test seems to be relatively insensitive to the significant pitch level difference between the source and target singings. This might be due to that the artificial pitch-shifting algorithm tends to introduce more distortion effects than the out-of-tune but natural singing.

Last but not least, in order to demonstrate the proposed method in an even more general case, we also consider the short-time Fourier transform (STFT) being the feature for alignment, as one might exploit the alignment information for various applications other than singing voice correction, and STFT is one of the most broadly used data representations in the field. Specifically, STFT is extracted from a monophonic source and its parallel target recording mixed with accompaniment music. The FFT size and hop length are kept the same with that used for SP. The dimension of STFT is reduced from 513 to 25 by principal component analysis (PCA) before alignment. We demonstrate the results on the website.[3]

5. CONCLUSION AND FUTURE WORK

We demonstrate the effectiveness of utilizing CTW-based algorithms with the aid of parallel data on the singing voice correction task. With the high stability in the alignment between a singing track and a mixture, and with improved pitch accuracy and fluency in the listening test, such a method is promising for real-world application. As future work, we will extend the study by incorporating other techniques such as melody extraction from polyphonic signals and query-by-humming algorithms, to offer facilities for retrieving melody contours from polyphonic sources which are accessible on the web. Furthermore, since the relation between a source-target pair is highly nonlinear, more advanced algorithms such as the deep CTW [26] might lead to further improvement.

6. REFERENCES

[1] N. Montecchio and A. Cont, "A unified approach to real time audio-to-score and audio-to-audio alignment using sequential montecarlo inference techniques," in *ICASSP*, 2011, pp. 193–196.

[2] C. Raffel and D. P.W. Ellis, "Optimizing DTW-based audio-to-midi alignment and matching," in *ICASSP*, 2016, pp. 81–85.

[3] S. Ewert, M. Muller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *ICASSP*, 2009, pp. 1869–1872.

[4] Sungkyun C. and Kyogu L., "Lyrics-to-audio alignment by unsupervised discovery of repetitive patterns in vowel acoustics," *IEEE Access*, vol. 5, pp. 16635–16648, 2017.

[5] L. Bochen, D. Karthik, D. Zhiyao, and S. Gaurav, "See and listenl score-informed association of sound tracks to players in chamber music performance videos," in *ICASSP*, 2017, pp. 2906–2910.

[6] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning features of music from scratch," *arXiv preprint arXiv:1611.09827*, 2016.

[7] M. Grachten, M. Gasser, A. Arzt, and G. Widmer, "Automatic alignment of music performances with structural differences," in *ISMIR*, 2013.

[8] T. Fukuda, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A score-informed piano tutoring system with mistake detection and score simplification," in *SMC*, 2015.

[9] "Melodyne by celemony," `http://www.celemony.com/en/melodyne/what-is-melodyne`, [Online; accessed 24-October-2017].

[10] E. Azarov, M. Vashkevich, and A. Petrovsky, "Guslar: a framework for automated singing voice correction," in *ICASSP*, 2014, pp. 7919–7923.

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

[12] O. Perrotin and C. Dalessandro, "Target acquisition vs. expressive motion: Dynamic pitch warping for intonation correction," *ACM Trans. Computer-Human Interaction (TOCHI)*, vol. 23, no. 3, pp. 17, 2016.

[13] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," in *ICASSP*, 2013, pp. 744–748.

[14] N. Kroher, E. Gómez, C. Guastavino, F. Gómez, and J. Bonada, "Computational models for perceived melodic similarity in a cappella flamenco singing," in *ISMIR*, 2014, pp. 65–70.

[15] L. Cen, M. Dong, and P. Y. Chan, "Template-based personalized singing voice synthesis," in *ICASSP*, 2012, pp. 4509–4512.

[16] Y. Wada, Y. Bando, E. Nakamura, K. Itoyama, and K Yoshii, "An adaptive karaoke system that plays accompaniment parts of music audio signals synchronously with users' singing voices," in *SMC*, 2017, pp. 110–116.

[17] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," in *Advances in neural information processing systems*, 2009, pp. 2286–2294.

[18] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *TASLP*, vol. 18, no. 2, pp. 310–319, 2010.

[19] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics. Wiley, 2003.

[20] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[21] Masanori M., "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, no. Supplement C, pp. 57 – 65, 2016.

[22] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *ICASSP*, 2008, pp. 3933–3936.

[23] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, 2017.

[24] B. McFee, M. McVicar, O. Nieto, S. Balke, C. Thome, D. Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, and et al., "librosa 0.5.0," Feb. 2017.

[25] F. Zhou and F. De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *CVPR*, 2012, pp. 1282–1289.

[26] G. Trigeorgis, M. A Nicolaou, S. Zafeiriou, and B. W. Schuller, "Deep canonical time warping," in *CVPR*, 2016, pp. 5110–5118.