Acappella synthesis demonstrations using RWC music database

[Application of auditory morphing based on STRAIGHT] *

Hideki Kawahara†
Wakayama University
930 Sakaedani, Wakayama
Wakayama, 640-8510 Japan
kawahara@sys.wakayama-u.ac.jp

Hideki Banno‡
Wakayama University
930 Sakaedani, Wakayama
Wakayama, 640-8510 Japan
banno@sys.wakayama-u.ac.jp

Masanori Morise
Wakayama University
930 Sakaedani, Wakayama
Wakayama, 640-8510 Japan
s055068@sys.wakayama-u.ac.jp

ABSTRACT

A series of demonstrations of synthesized acappella songs based on an auditory morphing using STRAIGHT [5] will be presented. Singing voice data for morphing were extracted from the RWCmusic database of musical instrument sound. Discussions on a new extension of the morphing procedure to deal with vibrato will be introduced based on the statistical analysis of the database and its effect on synthesized acappella will also be demonstrated.

Keywords

Rencon, Acappella, RWCdatabase, STRAIGHT, morphing

1. INTRODUCTION

Human voice is an ultimate musical instrument. Its information bandwidth from performers' intention to actual performance may be the best of all possible musical instruments. However, its bandwidth is taking advantage of timbre dimensions which have not being explored extensively by computer based music synthesis. A high-quality speech analysis, synthesis and modification system STRAIGHT [5] and an auditory morphing procedure [6] based on it have a potential to help explore these new and important domain of musical performance. This set of demonstrations intend to introduce potentials of STRAIGHT and the morphing procedure in such research, and hopefully, performance.

*Demonstrations and additional information can be found at `www.sys.wakayama-u.ac.jp/~kawahara/NIME04/`
Information about STRAIGHT can also be found at `www.sys.wakayama-u.ac.jp/~kawahara/PSSws/`

†Also an invited researcher of ATR Human Information Science Research Laboratories

‡Also a visiting researcher of ATR Spoken Language Translation Research Laboratories

2. STRAIGHT-BASED MORPHING

STRAIGHT decomposes a speech sound into the source information, namely fundamentally frequency (F0) with voiced-unvoiced (V/UV) distinction, and the smoothed time-frequency representation with virtually no interferences due to periodicity [5]. It also extracts band-wise periodicity indices for mixed-mode excitation [3]. In morphing between two speech tokens, firstly, the time-frequency coordinate system is piecewise bilinearly interpolated. The spectral and periodicity values on the time-frequency coordinate system are also linearly interpolated after linearization by appropriate nonlinear transformations and then inversely transformed. The F0 trajectory is also piecewise linearly interpolated in the log-frequency domain based on temporal markers indicating corresponding time-frequency points. Those morphed parameters are fed into parameters of STRAIGHT synthesis module and used to produce synthetic speech. The markers for defining corresponding points of two speech tokens are set manually [6].

3. RWC MUSIC DATABASE

Auditory morphing needs voice examples to start with. A portion[1] of RWCmusic database [2] which consisted of singing sounds by 15 singers, spanning from classical to modern R&B, supplied necessary source. Segmentation based on power and differential power yielded over 16,000 segments depending on thresholds. They were analyzed using STRAIGHT and YIN [1].

4. EXTENSIONS AND DEMONSTRATIONS

Figure 1 shows the F0 trajectory extracted for a sustained vowel /a/ sang at F#3 in *forte* dynamics without vibrato by one of a bass singer. As can be seen in the figure, the F0 trajectory shows a regular frequency modulation that is typical in vibrato. It is the natural behavior of singers and it was frequently found in other singers' data in the RWC

[1]RWC-MDB-I-2001 No. 45–50

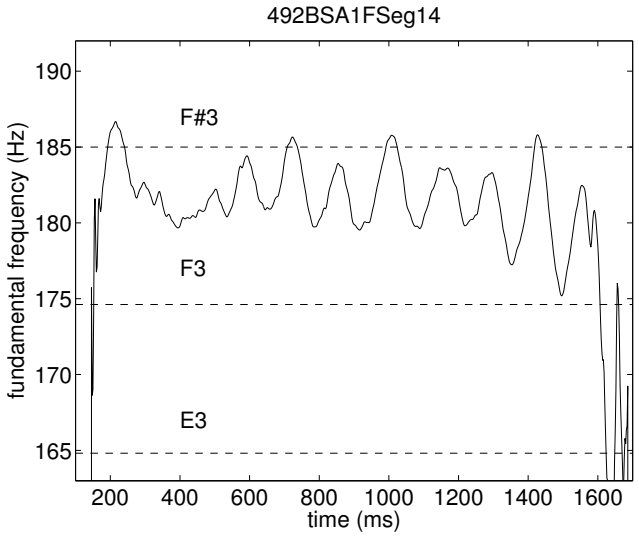

Figure 1: Fundamental frequency trajectory of F#3 /a/ sound sang by a bass singer. Note that there still exists a vibrato-like F0 modulation even though the singer was instructed not to do so.

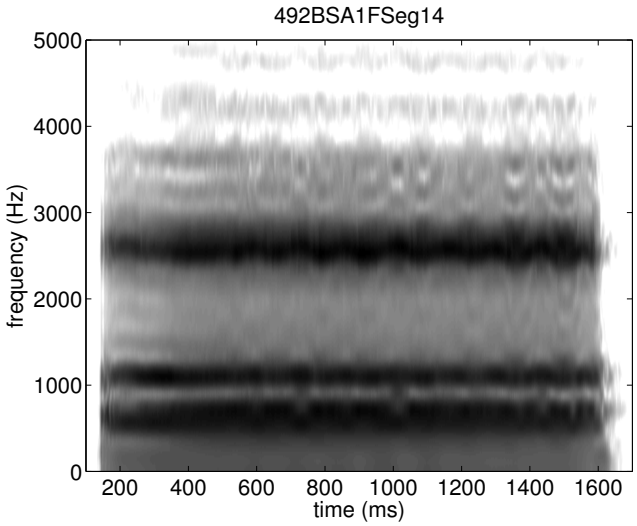

Figure 2: Smoothed time frequency representation for the same sound with Figure 1

database. However, it introduces difficulty in our current morphing procedure [3].

First problem is the interference between frequency modulations. A morphed vibrato sound made from two different examples needs to have a vibrato with intermediate rate and intermediate depth. The current implementation of auditory morphing simply interpolates F0 trajectories on the modified time axis and results into a vibrato with two modulation rates. The second problem is the correlation between F0 modulation and the smoothed time-frequency representation as shown in Figure 2. A systematic change that is synchronized with the F0 modulation is observed in this plot. A instantaneous frequency and instantaneous amplitude analysis on F0 frequency modulation and a decorrelation process based on the multiple regression analysis of the time frequency representation were introduced to solve these problems. Materials for acappella synthesis were processed using the proposed procedure to *de-vibrato* and made as loops that can be endlessly repeated. Synthetic singing examples with and without the proposed preprocessing will be demonstrated using several acappella pieces from different genre.

5. CONCLUSIONS

The demonstration illustrates only a portion of an evolutionary development based on *systematic downgrading* [4] for extracting rules on vocal performance. Even with current primitive stage of development, the proposed system demonstrates potential power and flexibility of morphing based acappella synthesis.

6. ACKNOWLEDGMENTS

This work is supported in part by a Grant in Aid for Scientific Research (B) 14380165 and Wakayama University. It is also supported in part by by the National Institute of Information and Communications Technology of Japan. Prof. Toshio Irino and Dr. Takanobu Nishiura made valuable, sometimes critical comments on our approach and were very helpful. The authors acknowledge Mr. Ryuichiro Yanaga and Ms. Rie Sakai for their assistance.

7. ADDITIONAL AUTHORS

Additional authors: Yumi Hirachi (Wakayama University, email: `s055051@sys.wakayama-u.ac.jp`)

8. REFERENCES

[1] A. de Chevengné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

[2] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Wc music database: Music genre database and musical instrument sound database. In *Proc. ISMIR 2003*, pages 229–230, october 2003.

[3] H. Kawahara. Exemplar-based voice quality analysis and control using a high quality auditory morphing procedure based on STRAIGHT . In *ISCA workshop VOQUAL'03*, pages 109–114, Geneva, August 2003.

[4] H. Kawahara and H. Katayose. Scat generation research program based on STRAIGHT, a high-quality speech analysis, modification and synthesis system. *IPSJ Journal*, 43(2):208–218, 2002. [In Japanese].

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. *Speech Communication*, 27(3-4):187–207, 1999.

[6] H. Kawahara and H. Matsui. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *ICASSP'2003*, volume 1, pages 256–259, Hong Kong, 2003.