MiruSinger: A Singing Skill Visualization Interface Using Real-Time Feedback and Music CD Recordings as Referential Data

Tomoyasu Nakano[†] Masataka Goto[‡] Yuzuru Hiraga[†]

[†]Graduate School of Library, Information
and Media Studies, University of Tsukuba
Tsukuba, Ibaraki 305-8550, Japan
{nakano, hiraga}@slis.tsukuba.ac.jp

[‡]National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

Abstract

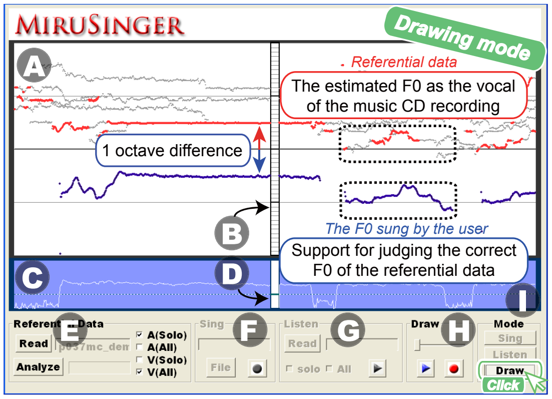
MiruSinger is a singing skill visualization interface that analyzes and visualizes vocal singing in reference to the vocal-part of music CD recordings. The system focuses on visualizing the characteristics of singing skills with real-time feedback. Although there are previous systems for singing training assistance that provide real-time visual feedback of the singing voice, none had utilized real-world (commercial) recordings as referential data. MiruSinger has the capability of visualizing F_0 (fundamental frequency) and vibrato sections of the user's singing voice in real-time, showing comparison with the estimated F_0 trajectory of the vocal-part in music CD recordings. The extracted F_0 trajectory can be hand-corrected to improve referential quality. A trial usage of the system shows that it would be a useful tool for average users, and that the system itself is entertaining and fun for the users.

1. Introduction

The aim of this study is to develop an interface for improving singing skill. The criteria for judging singing skills has been derived from the authors' previous study on singing evaluation by human subjects [1]. From both the objective analysis of the results and introspective comments reported by the subjects, the features identified to be significant for judging singing skills include: tonal stability (pitch accuracy), rhythmical stability, pronunciation quality, singing technique (in particular, the use of *vibrato*), vocal expression and quality, and personal preference.

Previous systems that assist singing training have focused on visualizing the characteristics of singing voice in real-time [2, 3]. The results of these works suggest that real-time visualization of sung fundamental frequency (F_0) shown in comparison with the referential pitch are useful functions for improving pitch accuracy [2, 3]. In these works, the referential pitch were synthesized sound, generated from music score data.

MiruSinger presented in this paper focuses on the visualization of two key features — F_0 (for pitch accuracy improvement) and vibrato sections (for singing technique improvement). Unlike previous systems, real-world music CD recordings are used as referential data. The F_0 of the vocal-part is estimated automatically from music CD recordings, which can further be hand-corrected interactively using a graphical interface on the MiruSinger screen. Using such referential data and correction interface is far more friendly and easy for the average user, disposing the need for separately preparing music score data. A snapshot of an example MiruSinger session is shown in Figure 1.



The estimated F0 as the vocal of the music CD recording

Support for judging the correct F0 of the referential data

Figure 1. An example MiruSinger screen.

2. Overview of MiruSinger

MiruSinger operates in three modes – *singing mode*, *listening mode*, and *drawing mode*, described in Sections **2.2–2.4**. These operating modes are selected by clicking the corresponding buttons in the lower right portion of the MiruSinger screen (Fig.1: Ⓘ). The upper part of the screen consists of the following two windows.

- *"Score" window* (Ⓐ) shows a graphical score of the vocal-part, sung by the user and/or extracted from CD recordings. The horizontal axis shows the timeline, with data moving from right to left as a real-time flow. The vertical axis shows the estimated F_0 frequency, in a window range of four octaves. The vertical bar at the center (Ⓑ) is the "present time", with the left half showing past, and the right half, future events.

0-7695-3084-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ISM.Workshops.2007.19

- *"Power" window* (Ⓒ) shows the input power level in coordination with the score window. The colored indicator at the center of this window (Ⓓ) corresponds to the "present" power level.

2.1. Analyzing music CD recordings

A music CD recording is analyzed by clicking the "Analyze" button (Fig.1: Ⓔ). The system estimates the relative dominance of possible F_0 candidates, and selects the most dominant F_0 in middle- and high-frequency regions as the vocal-part. The vocal signal is resynthesized and recorded from the estimated the F_0 and its harmonic structure.

The obtained data is used for both auditory and visual feedback. The auditory feedback is either the resynthesized vocal signal or the original CD recording, and the visual feedback is either the single F_0 trajectory of the vocal-part, or multiple trajectories of salient F_0 candidates. The feedback mode is selected by checking one of the following.

A (Solo) auditory feedback of resynthesized vocal signal
A (All) auditory feedback of the original CD recording
V (Solo) visual feedback of the F_0 of the vocal-part
V (All) visual feedback of the salient F_0 candidates

2.2. Singing mode

Recording starts by clicking the "record button (red circle)" (Fig. 1: Ⓕ). In this mode, vocal input sung by the user is recorded and analyzed in real-time. The F_0 trajectory is estimated and displayed, and the detected vibrato sections are highlighted by colored rectangular background.

The singing can be run simultaneously with the feedback of the referential data, in the auditory modes of A (Solo) or A (All), and/or visual mode of V (Solo).

2.3. Listening mode

Playback starts by clicking the "play button (blue triangle)" (Fig. 1: Ⓖ). In this mode, the user can listen to the recorded vocal input sung by the user, or to the referential data (resynthesized vocal signal or original CD recording), while visual feedback displays the F_0 and vibrato sections.

The feedback modes of A (Solo) or A (All), and V (Solo) can be selected in this mode.

2.4. Drawing mode

Estimating the F_0 trajectory of a particular part of an ensemble (either vocal or instrumental) is in general a difficult task, and judging the existence of a vocal part (as opposed to only instrumental parts) only adds on further difficulty. Our current system is not complete in this respect. The human listener, on the other hand, can easily recognize the existence of vocal parts, and can make simple corrections to a certain extent (*e.g.* correcting an octave error is relatively easy). So a practical solution is to provide facility for hand-correction of the obtained results. In the drawing mode, the user can correct the F_0 trajectory of the vocal-part working interactively with a graphical interface.

When the feedback mode V (All) is selected, the "Score" window shows the F_0 of the vocal-part with red dots, and the salient F_0 candidates with gray dots. Only the vocal-part F_0 is shown in the V (Solo) mode. The auditory feedback of A (Solo) or A (All) can be played for checking the results.

By using the slider (Ⓗ), the user can display portions of the data where the F_0 estimation may be incorrect. An irrelevant vocal-part F_0 can be removed by drag-and-drop using the right button of the mouse. The vocal-part F_0 can be changed to another candidate by drawing a rough trajectory using the drag operation (left button) of the mouse. The system re-selects the most dominant F_0 nearest to the user's drawing. The F_0 trajectory sung by the user is shown in blue dots, as further support for judging the correct trajectory of the referential data.

3. Internal mechanism of MiruSinger

The implementation of MiruSinger requires the following three main functions: (1) The F_0 estimation of the sung input in real-time, (2) F_0 extraction and estimation of the vocal-part in CD recordings, and (3) detection of vibrato sections in real-time.

Although the details cannot be described here, the basic methods used for the above three functions are as follows. (1) is realized by finding the most dominant harmonic structure of the sung input [4]. (2) uses *PreFEst* for estimating the predominant F_0 trajectory audio signals [5]. Experimental results with a set of ten music excerpts from CD recordings showed that PreFEst was able to detect melody lines with the accuracy of 88.4% [5]. (3) is realized by calculating vibrato likeliness using short-term Fourier transform (STFT) of the first order finite differential of F_0 [6].

4. Conclusion

This demonstration presented MiruSinger, a singing skill visualization interface with real-time visual and auditory feedback using music CD recordings as referential data. Trial usage of the system suggests that it would be a useful tool for average users, and that the system itself is entertaining and fun for the users. Remaining issues such as the visualization of other features related to singing skills (other than F_0 trajectory and vibrato) are topics of future work.

References

[1] T. Nakano et al., "Subjective evaluation of common singing skills using the rank ordering method," in *Proc. of ICMPC2006*, pp. 1507–1512, 2006.

[2] D. Hoppe et al., "Development of real-time visual feedback assistance in singing training: a review," *Journal of computer assisted learning*, Vol. 22, pp. 308–316, 2006.

[3] S. Hirai et al., "Clinical Support System for Poor Pitch Singers," *Trans. IEICE D-II*, Vol.J84-D-II, No.9, pp.1933–1941, 2001. (in Japanese)

[4] M. Goto et al., "A Real-time Filled Pause Detection System for Spontaneous Speech Recognition," in *Proc. of Eurospeech' 99*, pp.227–230, 1999.

[5] M. Goto, "A Real-time Music Scene Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals," *Speech Communication (ISCA Journal)*, Vol.43, No.4, pp.311–329, 2004.

[6] T. Nakano et al., "An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features," in *Proc. of Interspeech2006*, pp. 1706–1709, 2006.