IEICE TRANS. INF. & SYST., VOL.E87–D, NO.5 MAY 2004

PAPER Special Section on Speech Dynamics by Ear, Eye, Mouth and Machine

F_0 Dynamics in Singing: Evidence from the Data of a Baritone Singer

Hiroki MORI[†a)], ***Member***, **Wakana ODAGIRI**[†*], ***Nonmember***, ***and*** **Hideki KASUYA**[†], ***Fellow***

SUMMARY Transitional fundamental frequency (F_0) characteristics comprise a crucial part of F_0 dynamics in singing. This paper examines the F_0 characteristics during the note transition period. An analysis of the singing voice of a professional baritone strongly suggests that asymmetries exist in the mechanisms used for controlling rising and falling. Specifically, the F_0 contour in rising transitions can be modeled as a step response from a critically-damped second-order linear system with fixed average/maximum speed of change, whereas that in falling transitions can be modeled as a step response from an underdamped second-order linear system with fixed transition time. The validity of the model is examined through auditory experiments using synthesized singing voice.

key words: *singing voice, overshoot, transition time, asymmetry, voice synthesizer*

1. Introduction

Because the fundamental frequency (F_0) for each note is specified by the musical score for songs, F_0 generation for voice synthesizers seems to be simpler than that for text-to-speech systems. Using the specified value as a guideline, however, the actual F_0 contour produced is influenced by the physical limitations of the vocal organs, and sometimes by the demands of artistic expression. For the sake of creating musical melodies, almost all linguistic information contained in F_0 is omitted, and in Western classical music in particular, periodic fluctuation called vibrato is appended instead.

Among studies on the F_0 characteristics of singing voice, vibrato has been studied [1]–[8] intensively because of its importance for artistic expression. In contrast, there have been relatively few studies on the transitional aspects of F_0 dynamics. When a singer is about to move to the next note, a rapid F_0 change is thought to occur, which is a consequence of an F_0 control system whose target value is changing. From this point of view, Ohala et al. [9] and Sundberg [10] measured the maximum speed of F_0 changes in tone alternations. In the latter work, Sundberg suggested that the response times in F_0 rising were significantly longer than those in F_0 falling. Fujisaki et al. [11] also showed similar results. However, although these works implied an asymmetry in control mechanisms for the different directions of F_0 change, they did not mention the so-called "overshoot," in which the transitory F_0 exceeds the target value, which is often observed in singing voice. de Krom et al. [12] measured the extent of overshoot/undershoot for F_0 contours of an artificial song, and showed that overshoot was observed more consistently in F_0 falling than in rising as a whole. A similar study conducted by the authors' group supported their results [13], and revealed that the extent of overshoot affected the likelihood of synthesized singing voice being perceived as a human's voice.

However, these studies did not provide sufficient data to derive actual generative models of F_0 that properly reflect the nature of the asymmetry. This study therefore set out to establish a mathematical model to describe F_0 dynamics in singing, which could then be used to create a high-quality singing voice synthesizer. To accomplish this, transitional characteristics of F_0 were investigated in detail based on the recordings of passages sung by a professional vocalist in Western classical style. Although the data set is therefore merely a typical sample of singing voice, quantitative analysis of the data should help to derive a standard model for generating the F_0 contour for singing voice synthesizers.

This paper is organized as follows. Section 2 describes the method of measuring transitional characteristics of singing voice, and reveals two kinds of asymmetries in transition directions. In Sect. 3, an F_0 model of singing voice is proposed based on the findings described in Sect. 2, and the distribution of model parameters is investigated. In Sect. 4, an auditory experiment to confirm the validity of the hypothesized F_0 model is described, and the results are discussed. Section 5 concludes the paper.

2. Analysis

2.1 Experimental Procedure

A professional baritone (age 49) was presented with 24 independent musical scores, which were the combination of 2 patterns and 12 intervals. The patterns, which are shown in Fig. 1, consisted of "Rise→Fall (D3#)" and "Fall→Rise (C4)," in which the keynote was D3# (155.6 Hz) and C4 (261.6 Hz), respectively. Each score consisted of three notes, where the second note was surrounded by one of the keynotes. Therefore, each score contained one rising and one falling. The interval between the keynote and the second note ranged from 1 to 12 semitones (100–1200 cents). In addition, an extra set of "Fall→Rise (E3)" was presented with keynote E3 (164.8 Hz) and intervals ranging from 1 to 5 semitones. The singer sang each of the 29 scores three

Manuscript received August 29, 2003.
Manuscript revised November 17, 2003.

[†]The authors are with the Faculty of Engineering, Utsunomiya University, Utsunomiya-shi, 321–8585 Japan.
[*]Presently, with Casio Computer Co., Ltd.
a) E-mail: hiroki@klab.jp

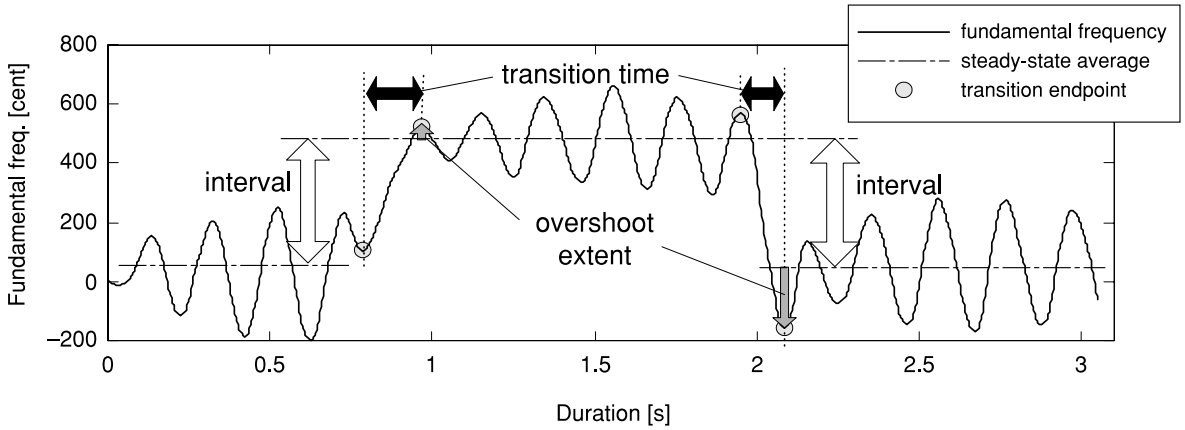

Fig. 2 An example of F_0 contours (Rise→Fall).

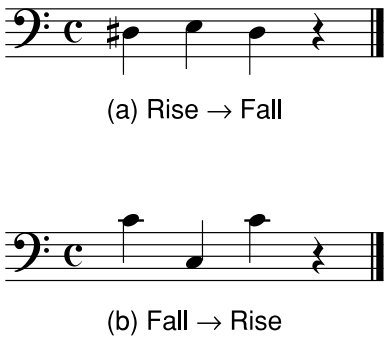

Fig. 1 Examples of presented musical scores.

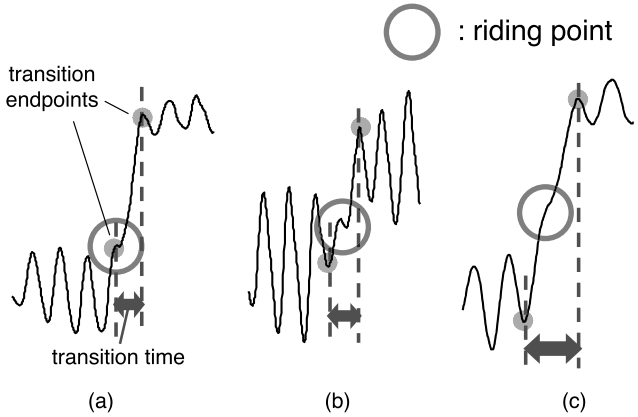

Fig. 3 Description of riding.

times. The singing style was legato, that is, transitions were smooth and natural. The singing was digitally sampled at 44.1 kHz and recorded using a DAT. All scores were sung with the vowel /a/.

The extraction of F_0 was performed for every 2.5 ms with the multiple window length method [14].

2.2 Parameters and Measurement

Figure 2 shows an example F_0 contour extracted from a Rise→Fall sample. In this example, the prescribed interval was 5 semitones (= 500 cents). The vertical axis indicates the value of sung $\log F_0$ relative to the prescribed keynote. This figure shows that sinusoidal modulation (i.e. vibrato) overlays the stationary part of each note. Note that the "overshoot" phenomenon certainly existed, at least in the falling transition in this case. Although it is difficult to distinguish overshoots from the fluctuations caused by vibrato, one of the authors had shown that overshoots were observed even for an untrained singer who could not produce vibrato at all [13].

From the extracted F_0 contours, the following parameters were measured:

steady-state average The averaged value of F_0 over the stationary part of each note. The stationary part was identified by visually checking the F_0 contour.

interval The difference of the steady-state average between consecutive notes. It is not necessarily the same as the interval prescribed in the score.

transition time The duration of transition. The endpoints of a transition were defined as local extrema of F_0 just before (or after) the transition.

overshoot extent The F_0 difference between the transition endpoint just after the transition and the steady-state average after the transition.

It is not easy to identify the transition endpoint because the direction of F_0 change demanded by the musical score and the one to which the vibrato is going to move may conflict. Myers et al. [3] pointed out that in such conflicting cases vibrato could "ride" on the rapid F_0 change rather than being interrupted. Based on the analysis, we defined "riding" as one of the three cases illustrated in Fig. 3:

(a) The local extremum is immediately followed by a global F_0 jump.
(b) Small vibrato overlies the global F_0 jump.
(c) The vibrato and the global F_0 jump are fused.

Transition endpoints in the riding cases were marked according to the scheme shown in Fig. 3.

2.3 Results

2.3.1 Transition Time

Figure 4 shows the measured transition time for each transition. A positive value of interval corresponds to rising, and a negative value to falling. The result reveals that the transition time in the falling cases was virtually constant, whereas in the rising cases, the transition time had a strong positive correlation with interval. There was no difference between the distribution of Rise→Fall pattern and Fall→Rise pattern. Also, no difference was observed in transition time between the Fall→Rise (C4) group and Fall→Rise (E3) group.

For most cases, the transition time for falling was not longer than that for rising. This overall tendency does not conflict with prior studies by Sundberg [10] or by Fujisaki et al. [11]. The rich variations of interval in our result support their observation, and clarify the difference between rising and falling.

Figure 4 also suggests that almost all "riding" cases occurred in rising transitions. Myers et al. [3] did not suggest this kind of asymmetry. Investigation of the cause of this phenomenon is beyond the scope of this paper, but the interaction between global F_0 control and vibrato certainly should be considered in conjunction with the asymmetry in transition directions.

2.3.2 Overshoot Extent

Figure 5 shows the extent of overshoot measured for each transition. In most cases for rising, the overshoot extent was quite small (< 1 semitone), whereas there was a strong positive correlation between intervals and overshoot extents in the falling cases. A slight difference was observed in the distributions among the patterns (Rise→Fall (D3#), Fall→Rise (C4) and Fall→Rise (E3)), but the overall tendencies were the same.

Previous studies([12], [13]) showed that overshoot extent is larger in falling transitions than in rising transitions. Our data shown in Fig. 5 strongly suggest that different mechanisms are at work in global F_0 control for different transition directions.

The sample average of vibrato depth (frequency deviation from steady-state average) was ±135.1 cents. In the falling cases, most (absolute) overshoot extent was significantly higher than the vibrato depth. Thus, overshoot is almost certainly a separate phenomenon from vibrato in falling transitions.

In rising transitions, on the other hand, the overshoot extent is so small that it is difficult to distinguish it from the deviation of vibrato, and so overshoot can be considered to be nonexistent, or at least ignored, in rising transitions.

The asymmetry of overshoot in transition directions contradicts the hypothesis of de Krom et al. that: "singers exaggerate the prescribed pitch transitions [12]." If overshoot were intentional, it would appear also in rising transitions. The causes of this asymmetry, which might include physiological mechanisms, should be investigated further.

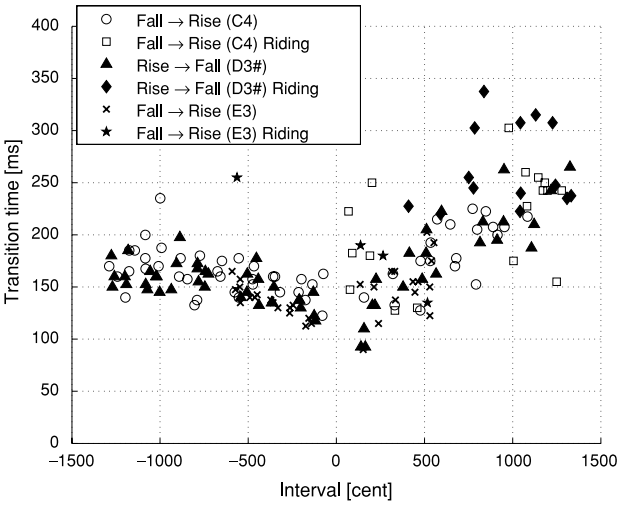

Fig. 4 Scatter plot of transition time versus interval.

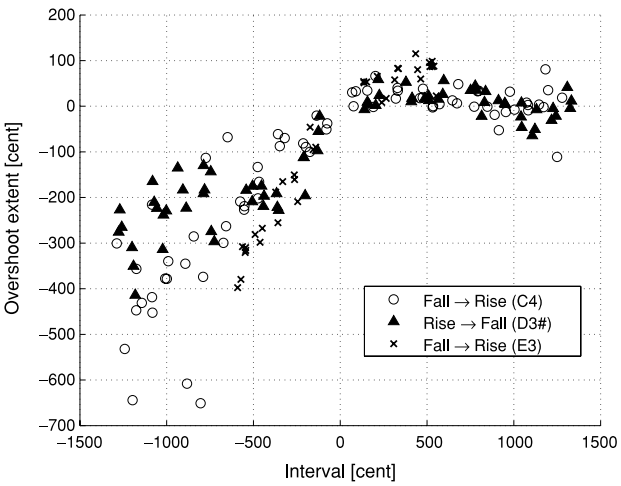

Fig. 5 Scatter plot of overshoot extent versus interval.

3. Modeling of F_0 Contour in Singing

3.1 Formulation

In this paper, the F_0 contour of singing voice is separately modeled as global F_0 and vibrato. Vibrato is assumed to exist only in the steady portion of global F_0. The final F_0 contour is represented as the sum (in the semitone domain) of them, as illustrated in Fig. 6. In this section, the modeling of global F_0 is described.

Transition portions are modeled as a step response from a second-order linear system in which transfer function $G(s)$ is represented as follows:

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{1}$$

where ζ is the damping factor and ω_n the natural frequency.

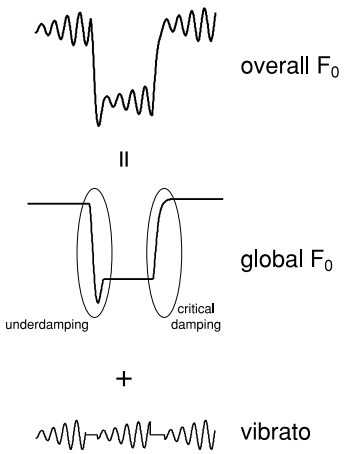

Fig. 6 Schematic representation of F_0 of singing voice.

As shown in Sect. 2.3, overshoots were observed in falling transitions but not in rising transitions. A rising transition is therefore modeled as the response from a critically-damped system, whereas a falling transition is modeled as that from an underdamped system. Other portions are assumed to be constant.

3.2 Parameter Estimation

The model parameters were estimated against the recorded F_0 contours described in Sect. 2. Each transition portion was segmented from the data by hand, then downsampled into 12.5–25 ms shift (depending on samples) to smooth the contour. The samples with "riding" were discarded because they violated the assumption of the previous formulation.

The Analysis-by-Synthesis method was used to estimate the parameters, which were: damping factor ζ (for falling transition only), natural frequency ω_n, and instant of step input t_{in}. For rising transitions, ζ was fixed to 1.

Figure 7 shows the distribution of damping factor ζ estimated against the falling transitions. There is no strong correlation between interval and ζ in this figure. Figure 8 shows the distribution of natural frequency ω_n estimated against the falling transitions. Again, no correlation between interval and ω_n was found.

The result shown in Fig. 4 suggests a somewhat proportional relationship between interval and transition time in rising transitions. For a critically-damped second-order linear system, the settling time is inversely proportional to ω_n. Thus, Fig. 9 is plotted against interval and ω_n^{-1} estimated against the rising transitions. The figure indicates a proportional relationship between interval and ω_n^{-1}. Assuming linearity, the following least-squares regression was obtained:

$$\omega_n^{-1} = 4.106 \times 10^{-5}|d|, \quad (2)$$

where d denotes interval (in cents). This analysis can be interpreted as follows. In our F_0 model for rising transitions, F_0 change is given as:

$$F_0(t) = d\{1 - e^{-\omega_n t}(1 + \omega_n t)\} \quad (t > 0), \quad (3)$$

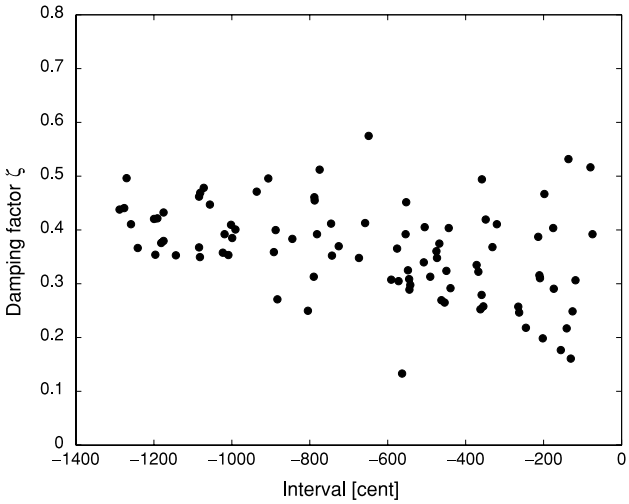

Fig. 7 Distribution of estimated ζ in falling transitions.

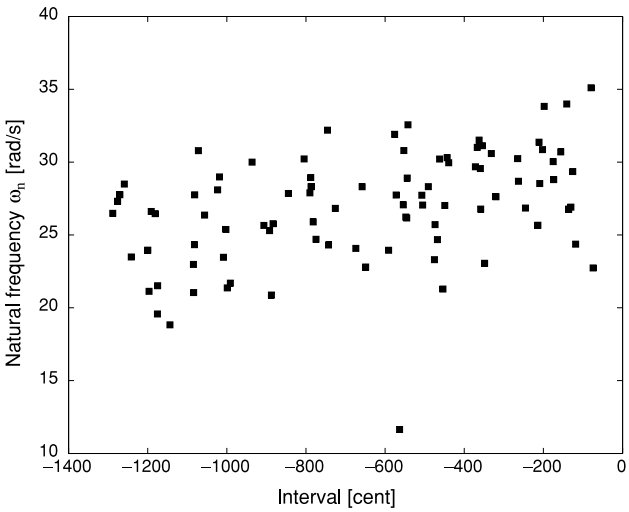

Fig. 8 Distribution of estimated ω_n in falling transitions.

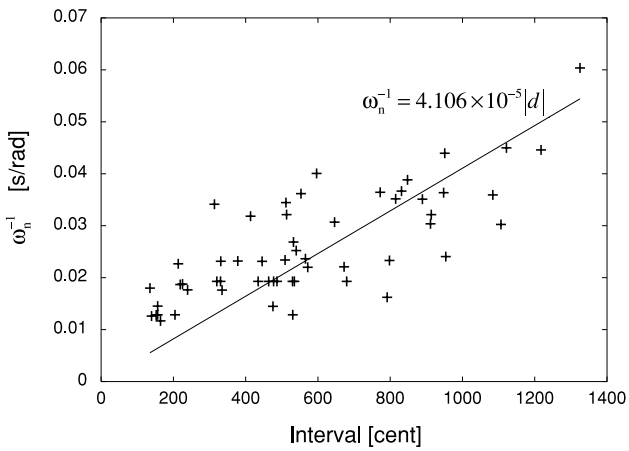

Fig. 9 Distribution of estimated ω_n^{-1} in rising transitions.

where $t_{\text{in}} = 0$ for simplicity. The speed of F_0 change is then given by:

$$\frac{d}{dt}F_0(t) = d\omega_{\text{n}}e^{-\omega_{\text{n}}t}(\omega_{\text{n}}t) \quad (t > 0). \tag{4}$$

From Eqs. (2) and (4), the following equation is derived:

$$\frac{d}{dt}F_0(t) = Ce^{-\frac{C}{d}t}\left(\frac{C}{d}t\right) \quad (t > 0). \tag{5}$$

In Eq. (5), manipulation of the interval d only affects the scaling of t. This implies that the average/maximum speed of F_0 change is not affected by interval under the assumption of linearity.

4. Auditory Experiment

4.1 Synthesis of Singing Voice

To evaluate the F_0 model of singing voice proposed in Sect. 3, several versions of synthesized singing voice were generated using the ARX speech analysis-synthesis system [15]. The parameters required for the synthesis were obtained beforehand by analyzing a sample from the materials.

To generate the global component of F_0 contours, two independent transition models were applied. The first one is the CS model (critically-damped, fixed average/maximum speed of change), which corresponds to rising transitions. Parameter ω_{n} was determined according to Eq. (2).

The second one is the UT model (underdamped, fixed transition time), which corresponds to falling transitions. The parameters were determined as $\zeta = 0.36$ and $\omega_{\text{n}} = 27.08$, according to Figs. 7 and 8.

The vibrato component was generated according to a sinusoidal model given by the following formula:

$$v(t) = \begin{cases} e^{\alpha\left(t-\frac{N}{f}\right)}E\,\sin(2\pi ft+\theta) & 0 \le t \le \frac{N}{f} \\ E\,\sin(2\pi ft+\theta) & \frac{N}{f} < t < T \end{cases}, \tag{6}$$

where f, E, N, and T denote the vibrato rate, vibrato depth, a constant, and the duration of vibrato, respectively. α denotes a constant to set the initial vibrato depth, which enables the vibrato component to be gradually increased. The parameters were determined in practice as $f = 5$ [Hz], $E = 100$ [cent], $N = 3$, and $\alpha = 2$. Before combining the two components, the timings of transition were adjusted so as not to cause discontinuities in resultant F_0 contours. An example of generated F_0 contours is shown in Fig. 10.

Two melodies (Melody 1 and 2) were used to synthesize test samples. Both melodies are two bars long, picked from a famous Japanese song "Kojo no tsuki (Moon over the Ruined Castle)." Melody 1 is the beginning of the song and contains a rising transition of 5 semitones. Melody 2 is the ending of the song and contains a falling transition of 7 semitones.

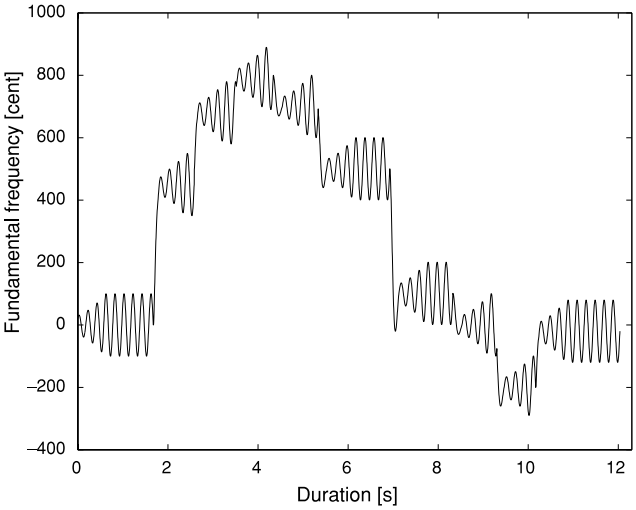

Fig. 10 An example of generated F_0 contours.

4.2 Design and Results

The experiments were designed to clarify whether singing voice synthesized with the proposed F_0 model (asymmetric in transition directions) sounds more natural than that synthesized with a symmetric model.

Experiment 1 Two versions of singing voice were synthesized for Melody 1.

a) Symmetric: the global F_0 contour was generated according to the UT model for both rising and falling transitions.
b) Asymmetric: the global F_0 contour was generated according to the CS model for rising transitions and the UT model for falling transitions.

Experiment 2 Two versions of singing voice were synthesized for Melody 2.

a) Symmetric: the global F_0 contour was generated according to the CS model for both rising and falling transitions.
b) Asymmetric: the global F_0 contour was generated according to the CS model for rising transitions and the UT model for falling transitions.

The subjects included 12 undergraduate and 8 graduate students whose musical experience was not controlled. In each experiment, a paired comparison test was performed, where 40 identical pairs of stimuli composed of versions a) and b) were presented. The order of the stimuli was randomized. The interstimulus interval was 0.8 sec. The subjects were asked to choose one of the two stimuli as more natural within 3 sec.

Table 1 shows the percentages and numbers of subjects who showed a significant difference in the judgment (binomial test, $p < 0.05$). Although most subjects were not sensitive to the difference of F_0 contour, not a few subjects

Table 1 Results of the auditory experiments. The values indicate the percentages and numbers of subjects who showed a significant difference in the judgment.

	a) symmetric	b) asymmetric	not significant
Exp. 1	0% (0/20)	35% (7/20)	65% (13/20)
Exp. 2	0% (0/20)	10% (2/20)	90% (18/20)

consistently chose the asymmetric construction as more natural in Experiment 1. This tendency was even less clear in Experiment 2, however, no subject preferred the symmetric version consistently.

The singer himself (excluded from the subjects) chose the asymmetric version for both Experiments 1 and 2. He also commented that stimulus a) in Experiment 1 and stimulus a) in Experiment 2 seemed rather artificial.

The experimental results did not perfectly coincide with the analyses described in Sect. 2. It is possible that a person who has less experience in performing music is less sensitive to acoustical properties of singing voice. To prove the effectiveness of the proposed model for singing voice synthesis, more musically-experienced subjects might be needed.

5. Conclusion

Transitional characteristics of F_0 in singing were investigated. There were clear asymmetries in transition directions. The analysis showed that the F_0 contour in rising transitions can be modeled as a step response from a critically-damped second-order linear system with fixed average/maximum speed of change, whereas that in falling transition can be modeled as a step response from an underdamped second-order linear system with fixed transition time. The auditory experiment showed that the singing voice of a melody synthesized with the proposed model sounded more natural than that synthesized with a symmetric model for some listeners.

In this paper, interaction between rapid F_0 change and vibrato was not studied in depth. Although the adjustment mechanism is still unclear [3], proper modeling of the interaction can affect the perceived naturalness of synthesized singing voice.

Besides F_0, there are several acoustic parameters (e.g. voice source amplitude) that dominate the quality of synthesized singing voice. Transitional aspects of such parameters should be studied in future investigations.

Acknowledgment

The authors would like to thank Prof. Kenji Ishino at Utsunomiya University for providing the singing voice data, and Prof. Johan Sundberg at Royal Institute of Technology, Sweden, for providing articles related to the subject of this paper.

References

[1] C. Seashore, ed., Studies in the Psychology of Music Vol.1: The vibrato, University of Iowa City, 1938.

[2] H.B. Rothman and A.A. Arroyo, "Acoustic variability in vibrato and its perceptual significance," J. Voice, vol.1, no.2, pp.123–141, June 1987.

[3] D. Myers and J. Michel, "Vibrato and pitch transitions," J. Voice, vol.1, no.2, pp.157–161, June 1987.

[4] L.A. Ramig and T. Shipp, "Comparative measures of vocal tremor and vocal vibrato," J. Voice, vol.1, no.2, pp.162–167, June 1987.

[5] Y. Horii, "Acoustic analysis of vocal vibrato: A theoretical interpretation of data," J. Voice, vol.3, no.1, pp.36–43, March 1989.

[6] C. d'Allessandro and M. Castellongo, "The pitch of short-duration vibrato tones," J. Acoust. Soc. Am., vol.95, no.3, pp.1617–1630, March 1994.

[7] E. Prame, "Measurements of the vibrato rate of ten singers," J. Acoust. Soc. Am., vol.96, no.4, pp.1979–1984, Oct. 1994.

[8] E. Prame, "Vibrato extent and intonation in professional Western lyric singing," J. Acoust. Soc. Am., vol.102, no.1, pp.616–621, July 1997.

[9] J. Ohala and W. Ewan, "Speed of pitch change," J. Acoust. Soc. Am., vol.53, p.345, Jan. 1973.

[10] J. Sundberg, "Maximum speed of pitch changes in singers and untrained subjects," J. Phonetics, vol.7, no.2, pp.71–79, April 1979.

[11] H. Fujisaki, M. Tatsumi, and N. Higuchi, "Analysis of fundamental frequency control in singing," Transactions of the Committee on Speech Research S79-80, The Acoustical Society of Japan, March 1980.

[12] G. de Krom and G. Bloothooft, "Timing and accuracy of fundamental frequency changes in singing," Proc. ICPhS '95, Stockholm, Sweden, vol.1, pp.206–209, Aug. 1995.

[13] M. Yatabe, Y. Endo, and H. Kasuya, "Dynamic characteristics of fundamental frequency in singing," Proc. Autumn Meeting Acoust. Soc. Japan, pp.383–384, Sept. 1998.

[14] T. Takagi, N. Seiyama, and E. Miyasaka, "A method for pitch extraction of speech signals using autocorrelation functions through multiple window-lengths," IEICE Trans. Fundamentals (Japanese Edition), vol.J80-A, no.9, pp.1341–1350, Sept. 1997.

[15] T. Ohtsuka and H. Kasuya, "Robust ARX-based speech analysis method taking voice source pulse train into account," J. Acoust. Soc. Jpn., vol.58, no.7, pp.386–397, July 2002.

Hiroki Mori received B.E., M.E. and Ph.D. degrees from Tohoku University, in 1993, 1995 and 1998, respectively. He was with the Graduate School of Engineering, Tohoku University in 1998. He is now a Research Associate of Utsunomiya University. His research interests include speech recognition, speech synthesis, spoken dialogue systems, and natural language processing. He is a member of the Acoustical Society of Japan and IPSJ.

Wakana Odagiri received B.E. and M.E. degrees from Utsunomiya University, in 1999 and 2001, respectively. Since 2001, she has been working for the Casio Computer Co., Ltd.

Hideki Kasuya received B.S., M.S., and Ph.D. degrees in Electrical Communication Engineering all from Tohoku University, Sendai, Japan, in 1963, 1965, and 1970, respectively. In 1968, he joined the Research Institute of Electrical Communication as a research associate at Tohoku University, where he was primarily engaged in speech analysis, perception and recognition. From 1974 to 1977 he was a visiting researcher at the Speech Communications Research Laboratory, Inc., California, U.S.A., working on speech recognition. Since 1978 he has been with the Faculty of Engineering, Utsunomiya University, where he is now a professor in the Department of Electrical and Electronic Engineering. His research interests include various areas of speech science and technology, digital signal processing, and image processing.