

Prosodic Analysis and Modeling of the NAGAUTA Singing to Synthesize Its Prosodic Patterns from the Standard Notation

Nobuaki MINEMATSU*, Bungo MATSUOKA*, and Keikichi HIROSE**

*Graduate School of Information Science and Technology, University of Tokyo

**Graduate School of Frontier Sciences, University of Tokyo

{mine,matsuoka,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

NAGAUTA (長唄) is a classical style of the Japanese singing. It has very original and unique prosodic patterns in its singing, where an abrupt and sharp change of F_0 is always observed at a transition from a note to another. This F_0 change is often found even where the transition is not accompanied by a change of tone. In this paper, we propose a model to synthesize this unique F_0 pattern from the standard notation. Further, this paper shows an interesting phenomenon about power movements at the F_0 changes. Acoustic analysis of NAGAUTA singing samples reveals that sharp increases of F_0 and sharp decreases of power are observed synchronously. Although no discussion on physical mechanisms of this phenomenon is done in this paper, another model to generate this unique power pattern is also proposed. Evaluation experiments are done through listening and their results indicate high validity of the two proposed models.

1. Introduction

Recent advances of computation have realized remarkable progresses of speech technologies, such as large vocabulary continuous speech recognition, concatenation-based speech synthesis, high-resolution speech analysis, and so on. With these technical advances, much of researchers' attention is paid to more challenging tasks, one of which is singing speech. Different cultures have different styles of singing. Many of previous studies on the singing speech, however, are thought to focus upon the styles of rather restricted regions, which are Western styles and BEL CANTO is one of their representatives. Singing styles have high relevance to acoustic properties of their *native* languages. Reference [1] reports that BEL CANTO emphasizes resonance of singing and that singing in Japanese with BEL CANTO comes to reduce naturalness as the Japanese language in the singing.

This paper focuses on a Japanese classical style of singing, NAGAUTA, and proposes two methods to model their unique F_0 and power patterns. The F_0 modeling was based upon a previous work[2], which proposed a second-order transfer function to synthesize F_0 patterns in the chorus-style singing from the standard notation. The abrupt and sharp changes of F_0 are considered as grace notes and, by adding another component for the grace note to the previous model, the F_0 patterns of NAGAUTA are characterized. Power patterns are also generated based upon a similar model but with some modifications.

2. The NAGAUTA singing

2.1. What is NAGAUTA ?

NAGAUTA literally means a long (長) song (唄). Its old and long history started in the 17th century and its main characteristics were developed in Edo Period. As NAGAUTA is often

regarded as “the heart of Kabuki music”, it was originally performed as Kabuki dance music with shamisens, Japanese long-necked and three-string guitars, and its unique singing style. These days, NAGAUTA is often played independently of the dance and it may remind young Japanese not of Kabuki but of shamisens. Unlike BEL CANTO, NAGAUTA respects a character's feelings on the stage and mood of the play. Then, its actual melody often depend upon singers and players. This paper acoustically analyzes NAGAUTA singings of an experienced singer but the authors have to admit that the above characteristics of NAGAUTA imply that the obtained results may be specific to the singer. However, the authors perceived the so-called NAGAUTA singing quite well in the singer's performance.

2.2. Recording of NAGAUTA singing samples

One of the original and unique characteristics of NAGAUTA is an abrupt and sharp change of F_0 observed at every note transition. This F_0 change is not represented explicitly in the notation and, based upon singers' experiences, they add it as a grace note to the baseline melody according to mood of the play. There are two kinds of the F_0 changes. One is added when the note transition is accompanied by a change of tone, called “FURI”, and the other is not, called “ATARI”. This paper focuses upon these two kinds of the F_0 changes and singing speech material including many of these changes were recorded.

A female semi-professional singer with 15-year experience of NAGAUTA and 6-year experience as chorus member joined the recording. NAGAUTA has its original notation, which is different from the standard one. But she indicated no difficulty in singing NAGAUTA with looking at a musical score represented by the standard notation. Then, the recording was done by singing a single vowel /a/ both in the chorus style and in the NAGAUTA style with the following three musical scores, listed in Figure 1. This recording strategy facilitated the subsequent acoustic analysis and modeling because differences between the two singing styles were directly observed and modification of the F_0 model of the chorus style singing[2] was expected to be able to characterize the NAGAUTA style singing. Type-1 and 2 scores were prepared to observe F_0 patterns in FURIs and type-3 score was for F_0 patterns in ATARIs. Each musical score was repeated several times and the recorded material was digitized at 16 bit / 16 kHz sampling to be used in the acoustic analysis.

2.3. F_0 patterns observed in the NAGAUTA singing

F_0 was extracted from all the singing material by using a high-resolution speech analyzer STRAIGHT[4] with 1,024 pt frame length and 1 ms frame shift. Figure 2 shows two F_0 patterns of type-1, chorus and NAGAUTA. Clearly shown, many abrupt and sharp F_0 changes, called FURIs, are observed in the NA-

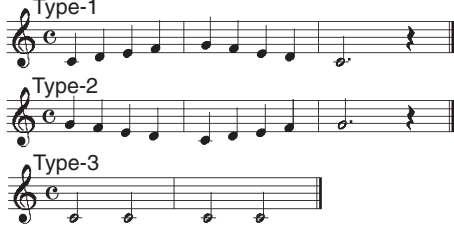


Figure 1: Three musical scores used in the recording

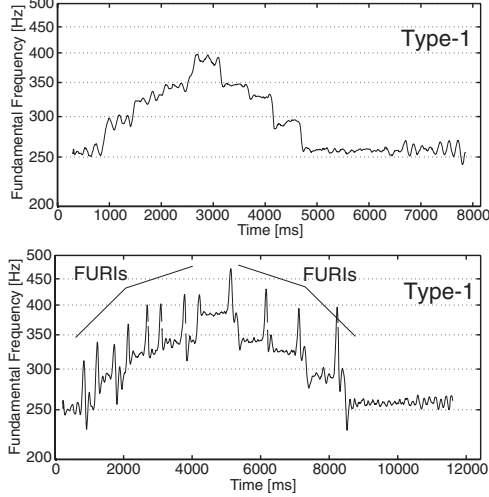


Figure 2: F_0 patterns observed in the Type-1 singing

GAUTA singing. Although F_0 patterns of the other types are not shown due to limit of space, the abrupt F_0 changes were regularly observed at a transition from a note to another irrespective of whether the transition was accompanied by a change of tone.

2.4. Power patterns observed in the NAGAUTA singing

Power was also calculated by STRAIGHT. Figure 3 shows the pattern observed in the NAGAUTA singing of type-1 with its F_0 pattern. A previous study of acoustic analysis of read speech showed positive correlation between F_0 and power[5]. Considering that speech segments with higher F_0 have pitch waveforms closer to each other, the above correlation is very normal. It was surprising that Figure 3 clearly indicates that the abrupt and sharp F_0 increases in NAGAUTA are accompanied with power decreases[6]. It is very interesting to clarify the physical mechanisms to induce this somewhat abnormal phenomenon. But in this study, main focus was put on modeling of the F_0 and power patterns observed in the NAGAUTA singing.

3. F_0 modeling in the NAGAUTA singing

3.1. Outline of the proposed model

In some of previous works, F_0 patterns of singing speech were often modeled as responses of the following second-order transfer function which took as input step-wise commands generated automatically from the standard notation[2].

$$H(s) = \frac{\omega^2}{s^2 + 2\zeta\omega s + \omega^2} \quad (1)$$

The response draws a rather smoothed curve, which is considered as the baseline melody, and a vibrato component is added

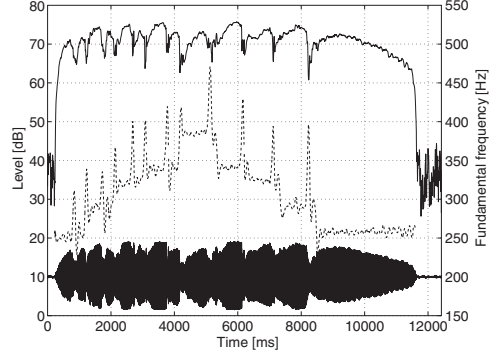


Figure 3: NAGAUTA power pattern in the Type-1 singing

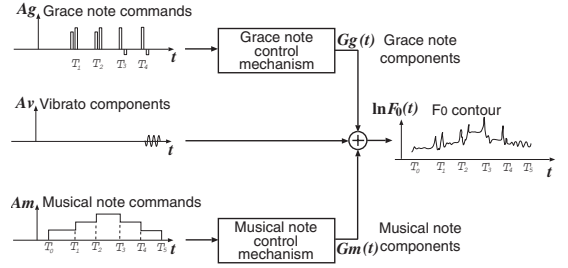


Figure 4: Proposed model for the F_0 pattern synthesis

to the baseline to finalize the F_0 pattern for singing. Figure 4 shows the F_0 model proposed for the NAGAUTA singing in this paper, where a new command, *grace note command*, is additionally considered and its output, *grace note component*, is added to the baseline melody with the vibrato component.

3.2. Parameter estimation for the new command

Since each of an ATARI and a FURI is modeled as a transfer function with its step-wise input command, ζ , ω , onset and offset of an input step, and its magnitude have to be estimated from the material. The recording gave us chorus and NAGAUTA singings with the same musical score and comparison between the two enabled the estimation. In the following, the parameter estimation for FURIs is explained in detail and that for ATARIs can be done only with minor modifications. It should be noted that F_0 in the explanation always denotes $\log(F_0)$.

Firstly, for each note transition of the two F_0 patterns, its onset of the F_0 change was detected by visual inspection and the two onset values were aligned to 0.0 on time axis. Then, subtraction of the chorus F_0 pattern from the NAGAUTA one was done to give us a differential pattern between the two, which is the grace note component in Figure 4. The upper figure of Figure 5 shows a differential pattern observed at a rising note transition. Two abrupt and sharp F_0 changes are seen and they were modeled as two short-time step responses. Two changes observed at a rising transition were always modeled as two positive responses but two changes at a falling transition were done as a positive and a negative responses. Parameters of the input command and the transfer function were estimated by a greedy search, where offset of the step command was automatically decided as the peak position of the first dumping of the response.

1. Initial values of ζ , ω , and magnitude of the grace note command were determined by visual inspection of the differential F_0 pattern.
2. Onset of the command as well as the above three param-

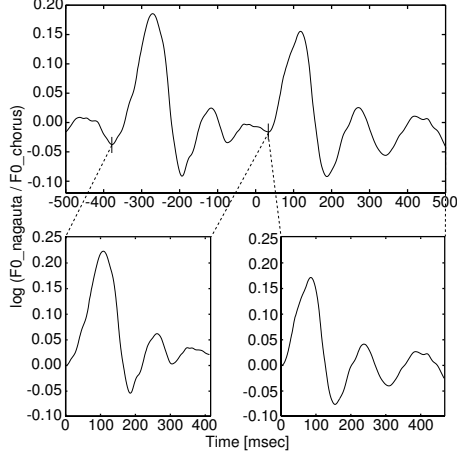


Figure 5: A differential F_0 pattern between the two singings

Table 1: Final parameter values for FURIs

	ζ	ω	pos.[ms]	mag.[cent]
UP1	0.23	0.036	-389	229
UP2	0.32	0.052	37	128
DOWN1	0.31	0.039	-250	287
DOWN2	0.45	0.056	46	-112

eters were searched by a greedy method to minimize an error function defined as normalized difference between the F_0 pattern observed as the grace note component and its synthetic pattern generated by the model.

The above procedure gave us a set of values of the four parameters for each FURI and their averaged values were adopted as the final values of the parameters, listed in Table 1. UP i shows the parameter values of the i -th FURI observed at a rising note transition and DOWN i is for the i -th one at a falling transition. Pos. in the table represents onset values of the command.

3.3. F_0 pattern synthesis with the proposed model

Using the parameter values listed in Table 1 for the grace note command and its transfer function, F_0 patterns were generated. For the musical note component, their parameter values were determined by referring to [2], which suggested $\omega=0.035$ and $\zeta=0.55$ for rising note transitions and $\omega=0.030$ and $\zeta=0.55$ for falling transitions. As for the vibrato component, according to [7], it was generated by adding a sinusoidal waveform of 6.6 Hz to white noise, which was followed by low-pass filtering. The resulting vibrato component waveform came to have peak-to-peak magnitude of about 96 [cent]. Figure 6 shows the F_0 pattern synthesized for the type-1 singing.

4. Power modeling in the NAGAUTA singing

4.1. Outline of the proposed model

As for the power modeling, a similar approach was taken. Literature [3] proposed a model to generate a power pattern for read speech, which is also formed as addition of a global component and a local one. In this paper, by adding another command for the abrupt power decrease, a power pattern model was built. Figure 7 shows the proposed model, where the global power pattern $G_I(t)$ is formulated as the following equations. In the proposed model, step width was set to 20 ms and onset was set so that the F_0 peaks and the power valleys became synchronous.

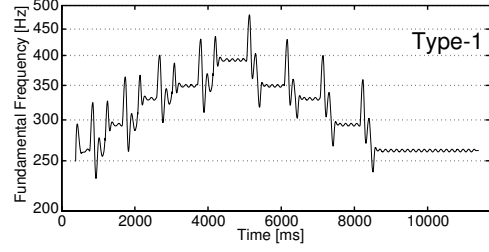


Figure 6: F_0 pattern synthesized for the Type-1 singing

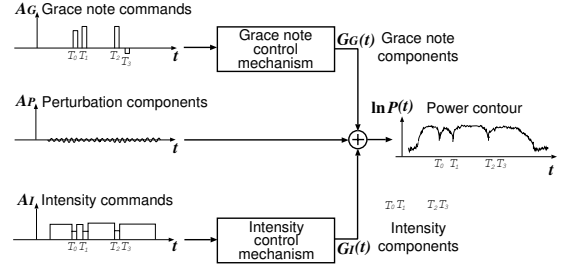


Figure 7: Proposed model for the power pattern synthesis

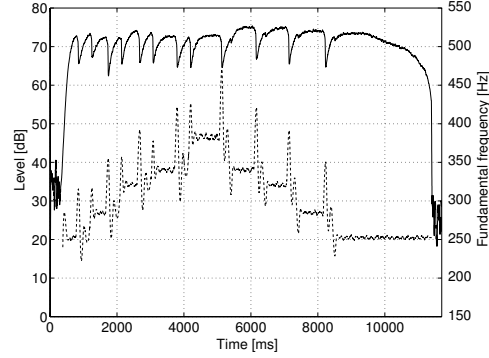


Figure 8: Power pattern synthesized for the Type-1 singing

$$\ln I(t) = \sum_{l=1}^L A_{I_l} \{G_I(t - T_{1l}) - G_I(t - T_{2l})\} \quad (2)$$

$$G_I(t) = \begin{cases} 1 - (1 + \lambda t) \exp(-\lambda t) & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (3)$$

4.2. Power pattern synthesis with the proposed model

As in the case of the F_0 pattern synthesis, estimation of the power model parameters were done, some of which were determined directly by referring to [3]. The resulting power pattern for the type-1 singing is shown in Figure 8. Clearly shown in the figure, the synchronous occurrences of the F_0 increases and the power decreases are adequately characterized.

5. Evaluation experiments

5.1. Evaluation of the proposed F_0 model

5.1.1. Procedures

A singing speech sample of vowel /a/ was prepared, which was long enough to generate a type- i singing sample by modifying the F_0 pattern of the sample. Since the sample was obtained from a male speaker, the target note transition of the type- i was set to that lower by one octave than the type- i shown in Figure 1. Re-synthesis through the F_0 modification was implemented on

STRAIGHT. Two kinds of re-synthesized speech were generated; a) the F_0 pattern was modified to be exactly one-octave lower than that observed in the NAGAUTA singer’s singing and b) the F_0 pattern was modified by the proposed model according to the type- i score. 10 university students with normal hearing joined this experiment as subjects who knew NAGAUTA but were not familiar¹. Firstly, we had the subjects listen to several samples of the NAGAUTA singing performed by the singer to enable the subjects to ascertain how the NAGAUTA singing sounds. After that, three pairs of singings, 3 types \times 2 kinds of F_0 , a) and b), were presented to the subjects through USB headphones, where D/A conversion was done in the headphones to cancel computer noise. After listening to each pair, they were asked to judge which one characterized the NAGAUTA singing better in terms of its melody. The judgment was done on a 5-degree scale where 5 and 4 meant that a) could characterize the NAGAUTA singing very and rather well respectively, and 1 and 2 meant that b) could do that very and rather well respectively.

5.1.2. Results and discussions

Table 2 shows results of the experiment. Although many of the F_0 pairs were judged to have no clear differences, it was surprising that the modeled F_0 patterns were judged to be relatively better than the original F_0 ones. After the listening experiment, we asked the singer to listen to the synthetic singing samples and obtained the interesting comment. “Grace notes of NAGAUTA were originally designed to represent a character’s feelings and mood of the play. Therefore, it is natural that acoustic realizations of the grace notes are often changed. But in this case of singing with no feeling or mood, the singing with monotonous grace notes may be perceived better as the NAGAUTA singing.” Comparison between Figures 2 and 6 certainly shows that the abrupt and sharp F_0 changes are realized monotonously in the latter. However, the results imply that people without deep knowledge on the NAGAUTA singing tend to favor monotonous and stable realizations of the grace notes. In either case, the majority of the judgments indicate no clear differences between the original F_0 pattern and the modeled one, which represents high validity of the proposed F_0 model.

5.2. Evaluation of the proposed power model

5.2.1. Procedures

Evaluation of the power model was also done through listening to re-synthesized speech samples. The long male /a/ speech sample was converted to have an F_0 pattern one-octave lower than the F_0 pattern of the original type- i singing. During the F_0 modification, the power pattern was generated in the following three ways. a) the power pattern was flat all over the singing, b) the power pattern of the type- i singing by the singer was used for the modification, and c) the power pattern was generated by the proposed model. After listening to the three kinds of singing samples with different power patterns, the subjects were asked to select the sample out of the three which best characterized the NAGAUTA singing in terms of its melody. Preliminary experiments without any instructions on the unnatural F_0 and power synchronization showed that the subjects tended to pay no attention to the power pattern. Then, we explicitly explained how F_0 and power were correlated in the NAGAUTA singing by acoustically presenting some singing examples, which was followed by the listening experiment to select the best sample.

¹Their familiarity with the NAGAUTA singing were thought to be on the averaged level as young Japanese.

Table 2: Results of evaluating the F_0 model

	1	2	3	4	5
type-1	1	3	5	1	0
type-2	0	4	5	1	0
type-3	0	3	6	1	0

Table 3: Results of evaluating the power model

	a)	b)	c)
type-1	0	7	3
type-2	0	6	4
type-3	0	6	4

5.2.2. Results and discussions

Results of the experiment are shown in Table 3. Although the original power pattern was judged to best characterize the NAGAUTA singing, about 40 % of the judgments supported the model-based power pattern. As mentioned above, without the advanced instructions, the flat power patterns tended to be favored. We asked again the NAGAUTA singer to listen to the samples and to judge which best characterized the NAGAUTA singing *for her*. It was very surprising to us that her answer was a), which was with flat power patterns. “Power decreases synchronized with F_0 increases are not always perceived as the so-called NAGAUTA singing and my favorite is a), which shows very stable singing.” The *ideal* NAGAUTA singing may exist only as a singer’s image, not as actual observations.

6. Conclusions

This paper proposed two models to generate F_0 and power patterns of the NAGAUTA singing from the standard notation. Both of the models were composed by a combination of a global component, local one, and another one for the grace note. Although the evaluation experiments showed high validity of the proposed models, they also implied inherent difficulty to technically deal with singing speech. We’re planning to discuss NAGAUTA singing with its song because words in a NAGAUTA song are very essential to its original characteristics.

7. References

- [1] I. Nakayama *et al.*, “An attempt to compare vocal expressions in Japanese traditional and western classical-style singing, using a common verse,” Tech. report of IEICE, SP97-114, pp.47–54 (1998, Japanese)
- [2] T. Saito *et al.*, “Extraction of F_0 dynamic characteristics and development of F_0 control model in singing voice,” Tech. report of ASJ, H-2001-93, pp.683–690 (2001, Japanese)
- [3] S. Ohno *et al.*, “Quantitative analysis of the effects of emphasis upon prosodic features of speech,” Proc. EUROSPEECH, pp. 661–664 (2001)
- [4] Hideki KAWAHARA, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,” Proc. ICASSP, pp.1303–1306 (1997)
- [5] Kenzo ITOH, “Correlation analysis between speech power and pitch frequency for twenty spoken languages,” Proc. ICSLP, pp.331–334 (1994)
- [6] N. Kobayashi *et al.*, “Acoustics and physiological characteristics of traditional singing in Japan,” Proc. Int. Conf. on Music Percep. and Cog. pp.171–174 (1989)
- [7] The Psychology of Music, edited by Diana Deutsch and published by Academic Press Inc. (1982)