Prosodic Modeling of Nagauta Singing and Its Evaluation

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 one of the classical styles of Japanese singing characterized by original and unique prosodic patterns. Abrupt and sharp changes of F_0 are often observed and they induce simultaneous abrupt changes of power. It is very interesting that the F_0 *increases* are synchronized with the power *decreases*. In our previous study, we proposed two models to synthesize the unique prosodic patterns from standard scores. Both of the F_0 and power patterns are acoustically modeled as damping oscillations realized by second-order systems. In this work, the proposed models are evaluated from two viewpoints. The first evaluation is rather elementary, where three very simple musical scores are used to test the F_0 and power models. In the second evaluation, a standard score of a real song is adopted to synthesize its Nagauta prosodic patterns. Here, a well-known Japanese song of "Furusato" (my old hometown) is used and the results indicate high validity and effectiveness of the proposed models.

1. Introduction

Recent advances of computation have realized remarkable progress 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, researchers' attention has shifted to more challenging tasks, one of which is singing due to its own specific difficulties with regard to prosody and voice quality. Different cultures have different styles of singing. Many of previous studies, however, are thought to focus on the styles of rather restricted regions, which are European styles and Bel Canto is their representative. In our previous work, we analyzed Nagauta singing acoustically and proposed two models for its unique and original F_0 and power patterns[1]. The F_0 modeling was based on another previous work[2, 3, 4], which proposed a second-order system to synthesize F_0 patterns for chorus style singing from standard scores. Since the abrupt changes of F_0 are considered as grace notes, by adding another component of the grace note to the previous model, the F_0 patterns of Nagauta were characterized in our previous study[1]. The power modeling, on the other hand, was also based on a previous work[5], which proposed a similar mechanism to produce power patterns for read speech. By adding a grace note component, the power model was modified to be able to synthesize Nagauta's patterns.

2. Nagauta singing

2.1. What is Nagauta?

Nagauta literally means a long (長) song (唄). It 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 three-string guitars. These days, Nagauta is often played not accompanied by the dance, and therefore, Nagauta may remind young Japanese not of Kabuki music but of shamisens. Unlike Bel Canto, which is performed according to a given musical score, Nagauta respects a character's feelings on the stage and his interpretation of the mood of the play. Then, its actual melody often depends on singers. This paper uses Nagauta singing samples by an experienced female singer but 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. F_0 and power modeling of Nagauta singing

An abrupt and sharp F_0 change is called *Furi* when it happens at a note transition with a change of tone. It is called *Atari* at a note transition without a change of tone. Figure 1 shows two F_0 patterns observed when a singer sings the same musical score with the two different styles, chorus and Nagauta. Furis are observed at note transitions in the Nagauta singing.

F_0 patterns in chorus singing are often modeled as responses of a second-order system, where input step-wise commands are generated from a given standard score[2, 3, 4].

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

The response draws a smoothed curve, which is considered the baseline melody, and a vibrato component is added to the baseline pattern to produce the final F_0 pattern for chorus. Another component, *a grace note component*, is added to the chorus pattern to give the Nagauta F_0 pattern[1]. Figure 2 shows a diagram of the proposed F_0 model. Nagauta power patterns are

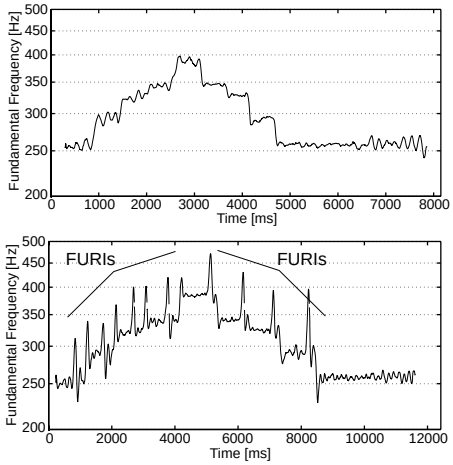

Figure 1: F_0 *patterns observed in chorus and Nagauta singing*

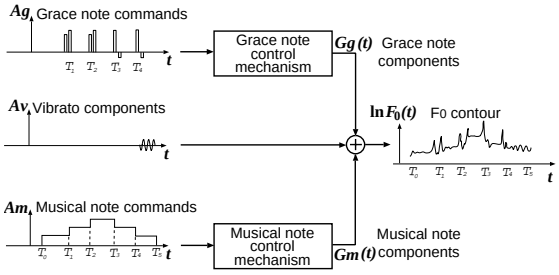

Figure 2: *Proposed model for F_0 pattern synthesis*

similarly generated, where a grace note component is added to the power pattern produced for read speech[5]. Figure 3 shows an example of the synthesized F_0 and power patterns. Synchronicity of F_0 increases and power decreases is well modeled. Detailed discussion of the models is found in [1].

3. Elementary evaluation of the models

3.1. Evaluation of the proposed F_0 model

Three scores in Figure 4 were prepared and sung by the female Nagauta singer several times. A speech sample of vowel /a/ was prepared by a male speaker, which was long enough to generate a type-i singing sample by modifying the /a/'s F_0 pattern. The F_0 modification was done on STRAIGHT[6]. Two kinds of F_0-modified samples were generated; (a) the F_0 pattern was modified to be one-octave lower than the Nagauta singer's pattern and (b) the F_0 pattern was modified by the proposed model. 10 students joined this experiment who knew Nagauta but not so familiar. Before the test, we had the subjects listen to several samples of the singer's performance for the subjects to ascertain

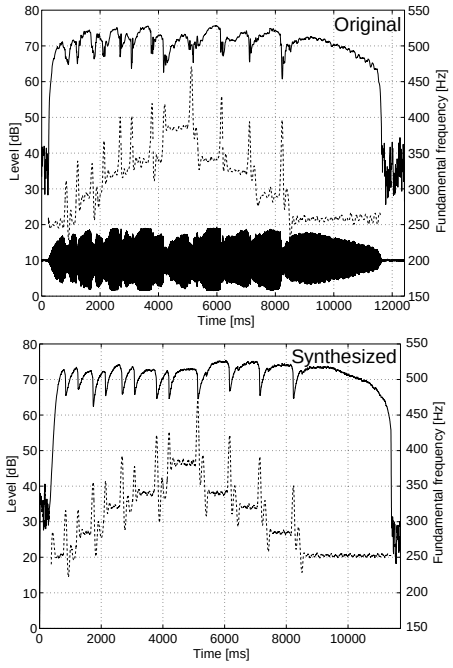

Figure 3: *Original and synthesized patterns of F_0 and power*

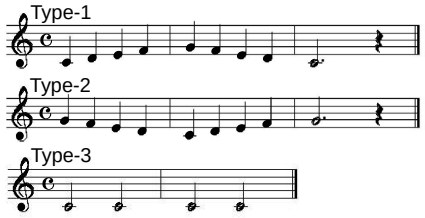

Figure 4: *Three scores prepared for the elementary evaluation*

how Nagauta sounds. Then, three pairs of samples, 3 types $\times$ 2 F_0 modifications, were presented through headphones. After listening, the subjects were asked to judge which one characterized Nagauta best in terms of melody. The judgment was done on a 5-degree scale where 5 and 4 meant that (a) could characterize Nagauta very well and rather well, respectively, and 1 and 2 meant that (b) could very well and rather well, respectively.

Table 1 shows the results. Although many of the F_0 pairs were judged to have no clear differences, it was surprising to us that the modeled F_0 patterns were judged to be better than the original F_0 ones. After the experiment, we asked the singer to listen to the synthetic samples. Her comment was as follows. "Grace notes of Nagauta are originally designed to represent a character's feelings and it is natural that acoustic realizations of the notes are often changed. But when singing with no feelings, regular patterns may be perceived better." The results imply that people without deep knowledge on Nagauta tend to prefer regular and stable realizations of the grace notes.

3.2. Evaluation of the proposed power model

The male /a/ sample was converted to have an F_0 pattern one-octave lower than the F_0 pattern of the original type-i sample. Three kinds of power modifications were done, (a) the power pattern was flat, (b) the original power pattern was used for resynthesis, and (c) the power pattern was generated by the proposed model. After listening, the subjects were asked to select the sample characterizing Nagauta best in terms of melody.

Results are shown in Table 2. Although the original power pattern was judged to characterize Nagauta best, about 40 % of the judgments supported the model-based patterns. We asked the singer again to listen to the samples and judge which characterized Nagauta best. It was 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 realizes very stable singing." The *ideal* Nagauta singing may exist only as a singer's image, not as actual acoustics.

Table 1: *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 2: *Results of evaluating the power model*

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

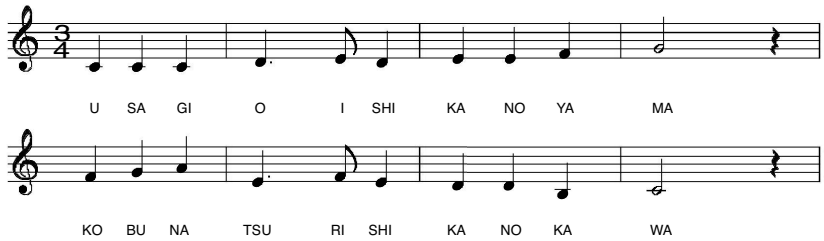

Figure 5: *The first eight bars of "Furusato" (my old hometown)*

4. Applied evaluation of the F_0 model

4.1. Selection of a real musical score

A well-known old Japanese song, "Furusato", which means my old hometown, was used as real musical score. Figure 5 shows the first eight bars of the song. Although "Furusato" is not a Nagauta song originally, the female singer told us that it was very easy to sing "Furusato" in Nagauta style. Seven recordings were done in the following manner.

1. Sing "Furusato" in Nagauta style only as /a/.
2. Sing it in Nagauta style with original lyrics.
3. Sing it in Nagauta style only as /a/ (but faithfully duplicating the melody realized in Recording #2).
4. Sing it in chorus style only as /a/.
5. Sing it in chorus style with original lyrics.
6. Sing it in Nagauta style with original lyrics.
7. Sing it in Nagauta style only as /a/.

The authors asked her to sing the song only as /a/, with original lyrics, and then, as /a/ again in Recordings #1 to #3. Nagauta is sometimes regarded as talking, not singing. If semantic aspects of the lyrics have some effects on the prosodic pattern, it was expected that the singer would find some prosodic differences between Recordings #1 and #2. Recording #3 was intended so that the singer could realize exactly the same prosodic pattern that was generated in #2. Recordings #4 and #5 were done to obtain chorus singing samples to estimate model parameters of the grace notes for chorus. The proposed methods require not only Nagauta samples but also chorus samples of the same score. The remaining two recordings were conducted just to increase the number of training data for parameter estimation.

F_0 and power patterns observed in bars 5 to 8 of "Furusato" are shown in Figure 6. The upper figure is from Recording #3 and the lower is from #2. Some grace notes disappear when the original lyrics are used. This is due to voiceless sounds and considered inevitable. Gray areas in the lower figure indicate where F_0 is not observed for that reason.

4.2. Parameter modifications for a real musical score

In Nagauta, all the note transitions, even with changes of tone, are not always accompanied by grace notes. It is assumed that every singer has his/her own strategy to add grace notes on the baseline melody. Visual inspection of the prosodic patterns will indicate whether additional modification is required or not for the proposed models to be applied to a real musical score. Here, only F_0 patterns were focused and the authors obtained the following findings from the "Furusato" singing samples

1. After breath breaks, an F_0 contour always becomes a curve smoother than that generated by the model.

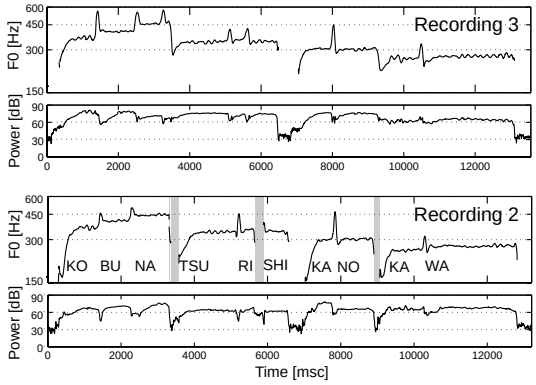

Figure 6: *Prosodic differences between recordings 2 and 3*

2. The proposed model generates a Furi as a sequence of two responses. But it is found that every Furi at a rising note transition is composed of a single response and every Furi at a falling transition with only a short interval to the following or previous note transition is also composed of a single response.

The first finding is seen in Figure 6. The lyrics are KO BU NA TSU RI SHI KA NO KA WA and they are divided into two phrases, KO BU NA TSU RI SHI and KA NO KA WA. The singer had a breath break before each phrase and the beginning of each phrase shows a rather smoothed F_0 curve. The second finding can be interpreted as follows. In our previous study, both of a Furi at a rising transition and that at a falling transition were modeled as two short step responses (two abrupt F_0 changes). But if a Furi is followed or preceded quickly by a note transition, due to some mechanical constraints of the articulators, it would be very difficult to realize two F_0 changes for a Furi. However, the authors could not find any strong reasons why all the rising Furis were composed of only single responses irrespective of an interval to the following note transition or the previous one.

As is described, all the note transitions are not accompanied by grace notes. Can the singer's strategy be modeled by some rules? "Furusato" has 45 notes and the authors focused on 37 notes by removing the 8 ones located at the beginning of phrases. This is because these 8 notes showed rather smoothed F_0 curves. Visual inspection of the singing samples of Recordings #1, #3, and #7 implied that a note transition with a large change of F_0 and that with a long duration tend not to be accompanied by a grace note. However, in this analysis, the authors could not build rules adequate enough to determine which note transition should be accompanied by a grace note. In the following section, the above two findings about modifications of the proposed F_0 model will be examined through listening experiments with young adult Japanese.

Table 3: *Parameter estimation for Furis in "Furusato"*

	ζ	ω	pos.[ms]	mag.[cent]
UP	0.46	0.044	41	293
DOWN	0.35	0.043	-98	212
DOWN1	0.33	0.041	-242	259
DOWN2	0.46	0.057	42	-144

Table 4: *Parameter estimation for Ataris in "Furusato"*

ζ	ω	pos.[ms]	mag.[cent]
0.37	0.035	19	375

Table 5: *Parameter estimation for Furis in* Section 3

	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

Table 6: *Parameter estimation for Ataris in* Section 3

ζ	ω	pos.[ms]	mag.[cent]
0.35	0.037	7	348

Table 7: *Averaged score for each segment in "Furusato"*

segment	1	2	3	4	5	6	7	8
score	3.7	3.2	4.5	3.3	4.3	4.3	3.3	3.2

4.3. Evaluation of the modifications through listening

Before the experiments, parameter estimation was done for Furis and Ataris using the "Furusato" singing samples. Results are shown in Tables 3 and 4. In the estimation, all the rising Furis and some falling ones were modeled as single responses. Parameters for the rising Furis are listed as UP and those for the falling Furis implemented as single responses are listed as DOWN. As for the other falling Furis, each implemented as two responses, DOWN1 and DOWN2 show their parameters. Pos. and mag. are position and magnitude of a command. Furi parameters of DOWN1 and DOWN2 in Table 3 and Atari parameters in Table 4 are very close to the parameters obtained in the experiments in Section 3, shown in Tables 5 and 6. This result indicates high robustness of the parameter estimation. Using a long male /a/ sample, "Furusato" was re-synthesized in two ways, without and with the modifications. In the former case, Tables 5 and 6 were used and, in the latter case, Tables 3 and 4 were used. A smoothed F_0 curve at the beginning of each phrase was realized with ζ and ω being 1.5 and 0.030, respectively. Figure 7 shows two prosodic patterns of bars 5 to 8 of "Furusato", generated without and with the modifications.

After dividing "Furusato" into 8 segments (2 bars for each), two synthetic singing samples, without and with the modifications, of each segment were presented to 10 university students. They were asked to judge which of the two characterized Nagauta singing best. The judgment was done on a 5-degree scale.

4.4. Results and discussions

Table 7 shows results of the experiments, which are averaged scores for each segment. Here, 1.0 means that prosodic patterns without the modifications are much better and 5.0 means that those with the modifications are much better. In segments 3, 5, and 6, the modifications are highly evaluated. For the other segments, the scores are around 3.0, which means that the modifications caused no improvement in terms of naturalness. What is commonly found in segments 3, 5, and 6 is that some falling note transitions are preceded quickly by other transitions, and therefore, two responses are difficult and simplified into a single response. To sum up, while the modifications with regard to falling Furis very close to neighboring transitions improve naturalness, those for rising Furis seem to be difficult for naive young Japanese to perceive.

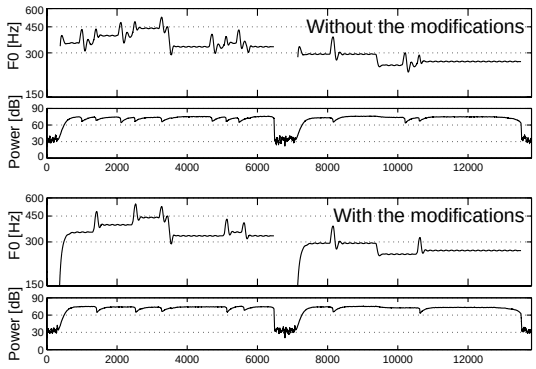

Figure 7: *Prosodic differences between without and with the modifications*

5. Conclusions

This paper proposed two models to generate F_0 and power patterns of Nagauta from standard scores. Both of the models were composed of a combination of a global component, a local one, and another one for grace notes. Two kinds of evaluation experiments were carried out. The first evaluation was rather elementary, where F_0 and power patterns corresponding to simple and short note sequences were acoustically realized in Nagauta style. Results of the evaluation showed that, although the synthesized F_0 and power patterns were relatively regular, they were preferred even by the Nagauta singer. In the second evaluation, the models were applied to a real musical score. Here, some additional modifications of the F_0 model were examined by considering articulatory constraints. Smoothing of the F_0 curves and simplification of acoustic realization of the grace notes were shown to be effective to improve naturalness of the synthesized F_0 patterns. Evaluation of the models by Nagauta experts will be done as a future work.

6. References

[1] N. Minematsu, B. Matsuoka, and K. Hirose, 2003. Prosodic analysis and modeling of the Nagauta singing to synthesize its prosodic patterns from the standard notation, *Proc. EUROSPEECH*, 385–388

[2] H. Fujisaki, M. Tatsumi, and N. Higuchi, 1981. Analysis of pitch control in singing, in *Vocal Fold Physiology*, 347–363, University of Tokyo Press

[3] K. Kashino and H. Murase, 1998. A Karaoke creator from original music signals using part score information, *Proc. Spring Meeting Acoust. Soc. Jpn.*, 625–626 (Japanese)

[4] T. Saito, M. Unoki, and M. Akagi, 2001. Extraction of F0 dynamic characteristics and development of F_0 control model in singing voice, *Tech. report of ASJ*, H-2001-93, 683–690 (Japanese)

[5] S. Ohno and H. Fujisaki, 2001. Quantitative analysis of the effects of emphasis upon prosodic features of speech, *Proc. EUROSPEECH*, 661–664

[6] H. Kawahara, 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited, *Proc. ICASSP*, 1303–1306