EXTRACTION OF PITCH REGISTER FROM EXPRESSIVE SPEECH IN JAPANESE

Jinfu Ni, Yoshinori Shiga, and Chiori Hori

Spoken Language Communication Lab., Universal Communication Research Institute,
National Institute of Information and Communications Technology, Kyoto, Japan

ABSTRACT

Human uses intonation to make focal prominence to give emphasis that highlights the focus of speech. Automatic extraction of proper intonation features from a speech corpus is desirous for processing speech prosody, especially in the context of speech synthesis. This paper presents a method to extract pitch register from observed F_0 contours for this purpose. The method utilizes a constrained tone transformation technique under an assumption that lexical accents are confined to parallel high and low tone lines with a limited constant span. Consequently, the extracted pitch register captures dynamic range variation of the pitch accents of an utterance. The method is evaluated by objective tests upon a large-scale expressive speech corpus. A finding is that proper intonation manifested in pitch register in Japanese is very comparable with English intonation in the sense of structural form.

Index Terms— Fundamental frequency analysis, intonation proper, pitch register, pitch decomposition, and speech prosody

1. INTRODUCTION

All vocal languages use pitch (or fundamental frequency (F_0)) to convey linguistic and paralinguistic meaning [1][2]. Japanese is a pitch accent language where both lexical accent and proper intonation are manifested in F_0 contours [3]. Automatic extraction of proper intonation from a speech corpus is desirous for processing speech prosody [4], especially in the context of speech synthesis. According to the Fujisaki model [5], the F_0 contour of an utterance in Japanese is superposition of accent and phrase components. The principle of superposition is attractive as it is intuitive to model different components or functions of separately [6] [7]. However, automatic pitch decomposition into its constituent part turns out to not be a trivial task [8][9][10] [11][12][13]. The difficulty is that, unless certain assumptions are made, there is no unique solution to pitch decomposition because the contour components can trade to produce the same F_0 contours [14].

In the previous work [15], another superpositional model is developed to decompose the F_0 contour of an utterance into pitch movements exclusively owing to lexical accents and the rest contributing to changes in intonation proper, named pitch register, a landscape that involves raising or lowering of both high (H) and low (L) tones of lexical accents.

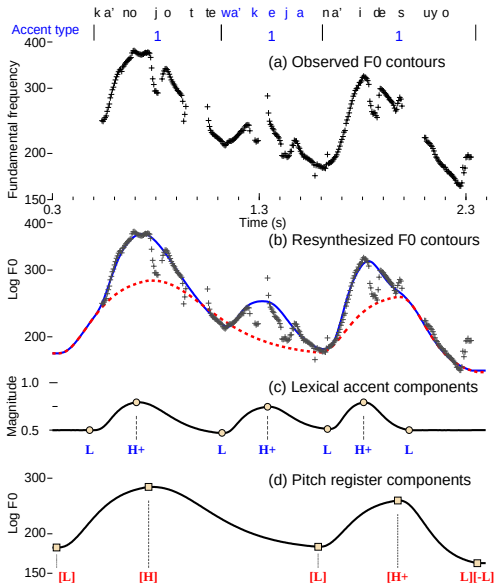

Fig. 1. *Example of decomposing F_0 contours into lexical accent and pitch register components and corresponding labels.*

This paper presents a method of extracting pitch register from observed F_0 contours to highlight the expressive aspect of intonation. The method is based on a constrained tone transformation technique [16] with an assumption that lexical accent components are confined to a limited constant range so as to achieve unique pitch decomposition. The basic motivation is to automatically extract proper intonation features from a large-scale speech corpus toward improving the expressiveness of HMM-based speech synthesis. Our investigation also reveals that new structural forms of expressive intonation patterns exist in Japanese, although they are well known in intonation languages like English.

The rest of this paper is organized as follows. Section 2 presents the methodology including a functional F_0 model to formulate F_0 contours, two work assumptions, and an algorithm of extracting pitch register from observed F_0 contours. Section 3 describes experimental results and discussions. Section 4 concludes this paper.

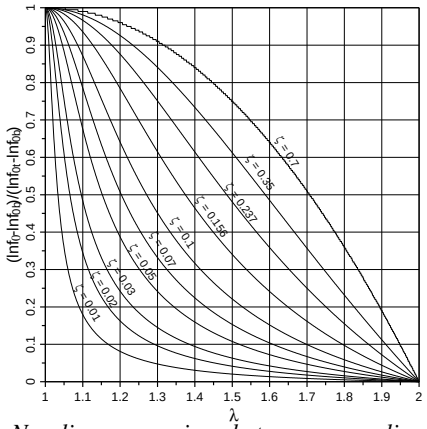

Fig. 2. *Non-linear mappings between normalized F_0 and λ ($1 \leq \lambda \leq 2$) when giving ζ value ($\zeta^2 < 0.5$).*

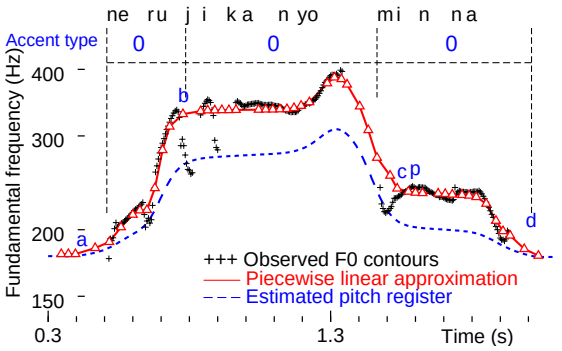

Fig. 3. *Illustration of pitch register estimation. The vertical lines indicate accentual phrase boundaries. The triangle sequences show piecewise linear approximation of F_0 contours and the dashed curves indicate the estimated pitch register.*

2. METHODOLOGY

2.1. Superpositional representation of F_0 contours

In the light of Fujisaki model [5], the F_0 contour of an utterance is regarded as superposition of lexical accent component on pitch register component [15]. Both components are parameterized by two sets of tonal targets, called accent targets and register targets, respectively, and connections between two adjacent accent/register targets are interpolated by the Poisson process [17]. The F_0 contour of an utterance over time t at the logarithmic scale, $\ln F_0(t)$, is expressed by

$$\ln F_0(t) = (C_a(t) - 0.5) + C_r(t), \quad t \geq 0,$$

$$C_a(t) = \bigsqcup_{i=1}^{I_a} (\gamma_{a_{i-1}} + (\gamma_{a_i} - \gamma_{a_{i-1}}) P(t, t_{a_{i-1}}, t_{a_i})), \quad (1)$$

$$C_r(t) = \bigsqcup_{i=1}^{I_r} (\gamma_{r_{i-1}} + (\gamma_{r_i} - \gamma_{r_{i-1}}) P(t, t_{r_{i-1}}, t_{r_i})), \quad (2)$$

$$P(t, t_{x_{i-1}}, t_{x_i}) = 1 - \sum_{j=0}^{2} \frac{(\frac{6.3 \times (t - t_{x_{i-1}})}{t_{x_i} - t_{x_{i-1}}})^j}{j!} e^{-\frac{6.3 \times (t - t_{x_{i-1}})}{t_{x_i} - t_{x_{i-1}}}}, \quad (3)$$

which is Poisson process-based interpolation. $C_a(t)$ and $C_r(t)$ indicate accent and register components, respectively.

The model parameters are listed as follows.

$I_r + 1$: Number of register targets.
(t_{r_i}, γ_{r_i}): ith register target; t_{r_i} is time and γ_{r_i} magnitude.
$I_a + 1$: Number of accent targets.
(t_{a_i}, γ_{a_i}): ith accent target; t_{a_i} is time and γ_{a_i} magnitude.

Japanese is a pitch accent language. Each accentual phrase has in lexicon a lexical accent (including non-accent), which is indicated by accent type 0 (non-accent), 1, 2, ... The observed F_0 contour of an utterance is the phonetic implementation of structures of lexical accents and proper intonation underlying the utterance. To achieve unique decomposition of $C_a(t)$ and $C_r(t)$, certain assumptions are necessary.

2.2. Work assumptions

Assumption 1: Lexical accent components are confined to high (H) and low (L) tonal grid lines and the span between the two lines is constant (limited to 4 semitones).

The method of using tonal grid lines is rooted in the Garding model [18], where the original grid lines are two parallel lines fixed in the voice range of a speaker. In this work, both H and L lines are floating onto the pitch register of an utterance.

Four tones are used to label the accent targets, as used in the literature [3][4].

L(low), H(high), H+L(nuclear fall), H% (boundary rise).

Assumption 2: Pitch register is the remainder of observed F_0 contours minus the underlying lexical accent components.

Given the strict constraint on the lexical accent components that characterize the lexical accents, pitch register turns to represent intonation proper. Particularly, pitch register manifests the intonation structures. In the light of intonation modeling in English [21], a set of register tones is defined below to label pitch register components.

- Accent register tones: [L], [H], [H+L], and [H+H].
- Phrase register tones: [-L] and [-H].
- Boundary register tones: [L%] and [H%].

[H+L] labels sharp fall at the mid-back part of an accentual phase and [H+H] high plateau to consider the high-level features of Japanese accents (except for types 1 and 2).

Figure 1 shows an example of pitch decomposition with the two assumptions. The circles and squares indicate target points to anchor the lexical accent and pitch register components, respectively. The tonal targets of the lexical accent components are labeled by L H+L H+L H+L and those of the pitch register components by [L][H][-L][H+L][-L]. There are three lexical accents with the same accent type 1. It clearly demonstrates that the expressive intonation of the utterance is manifested in the pitch register components.

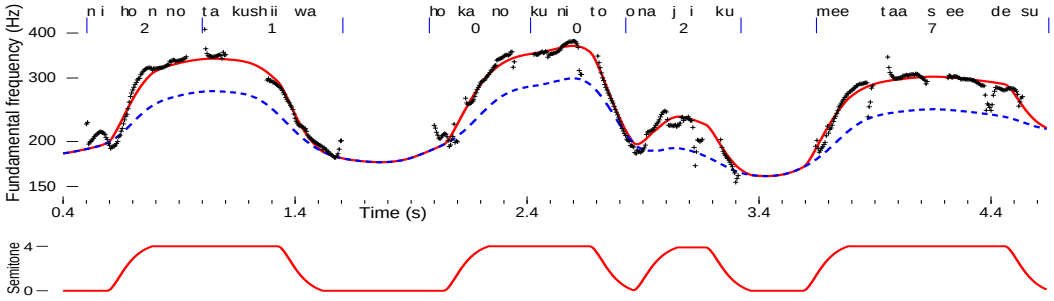

Fig. 4. *Example of automatically estimating pitch register (the dashed curves) and lexical accent components (the solid curves shown on the bottom). Superposition of the pitch register and lexical accent components is superimposed on the observed F_0 contours (the cross sequence). The phonemes, accent types, and accentual phrase boundaries are displayed on the top.*

2.3. Pitch register estimation

A constrained tone transformation technique [16] is used to automatically estimate pitch register from the observed F_0 contours under the assumptions mentioned above. The technique consists of an affine transformation (Eq. (4)) and a resonance transformation controlled by damping ratio ζ (Eq. (5)).

$$\frac{\ln f_0 - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\lambda,\zeta) - A(2,\zeta)}{A(1,\zeta) - A(2,\zeta)}, \quad (4)$$

$$A(\lambda,\zeta) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \quad (5)$$

f_0, λ, and ζ: Three variables, $1 \le \lambda \le 2$ and $\zeta^2 < 0.5$.
$[f_{0_b}, f_{0_t}]$: Given parameters to specify a F_0 range span.

A system of non-linear mappings between normalized F_0 (the left hand at Eq. (4)) and λ is well defined with parameter ζ, as shown in Fig. 2. Actually, when giving any one parameter, mapping between the other two is well defined, too. For convenience, three symbols are used to indicate these mappings.
$f(f_0 \to \lambda|\zeta)$: map f_0 to λ given ζ subject to Eqs. (4) and (5).
$f(\lambda \to f_0|\zeta)$: map λ to f_0 given ζ subject to Eqs. (4) and (5).
$f(\lambda \to \zeta|f_0)$: map λ to ζ given f_0 subject to Eqs. (4) and (5).

An algorithm is developed to estimate pitch register from observed F_0 contours using the mappings and assumptions.
Step-0 Input data.

- The observed F_0 contours of an utterance.
- The phone labels of accentual phrases.
- The lexical accent types of the accentual phrases.

Step-1 Preprocessing the F_0 contours.

- Smooth the observed F_0 contours after interpolating the unvoiced regions [19][20] and generate piecewise linear approximation of the F_0 contours. An example is shown in Fig 3 where the triangle sequence indicates the piecewise linear approximation.
- Partition the piecewise linear approximation into segments using the boundary times of accentual phrases, and adjust the boundary points to the nearest local valleys. In the example shown in Fig. 3, the boundary points marked by "a" and "c" are re-adjusted.

Step-2 Pitch register estimation.

- For a segment, set maxinum F_0 to f_{0_p} (e.g., "p" in Fig 3) and set minimum F_0 to f_{0_b},$f_{0_t} = f_{0_b} + 3 \times (f_{0_p} - f_{0_b})$.
- Set $\hat{f}_{0_p} = f_{0_p} - \Delta f_0$ (taking 4 semitones in the paper).
- Calculate λ_p by using $f(f_{0_p} \to \lambda_p|\zeta = 0.7)$.
- Calculate ζ_p by using $f(\lambda_p \to \zeta_p|\hat{f}_{0_p})$.
- Use ζ_p to convert all F_0 in the segment, f_{0_i}, $i = 0, ..., n$, thus obtaining raw register curves $\hat{f}_{0_i}$.
  - Compute λ_i by $f(f_{0_i} \to \lambda_i|\zeta = 0.7)$.
  - Compute $\hat{f}_{0_i}$ by $f(\lambda_i \to \hat{f}_{0_i}|\zeta_p)$.
- Smooth the raw register curves to obtain the pitch register of F_0 contours (the dashed curves in Fig. 3).

Step-3 Lexical accent component estimation.

- Estimate accent components from the remainder of the F_0 contours minus the estimated pitch register curves.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Experimental evaluation is conducted on a Japanese expressive speech corpus amounting to 5.5 hours of speech by a female speaker. It includes two steps. The first step extracts the pitch register and lexical accent components from the speech corpus using the proposed method; forced phone alignment is used at the input. The second step parameterizes the extracted components by analysis-by-synthesis to detect tonal targets using the functional F_0 model. The metrics used in the first step include root-mean-square error (RMSE) in Hz and Pearson's correlation coefficients (hereafter, correlation). Informal perception is carried out at the second step by analysis-by-synthesis method, thus confirming the effectiveness of using the pitch register to represent intonation proper instead of the original F_0 contours via re-synthesizing speech.

Table 1 shows the objective test results, where "approximation" indicates the piecewise linear approximation of F_0 contours and "superposition" the superposition of extracted lexical accent and pitch register components. The test samples are 6,402 utterances (the whole speech corpus used here).

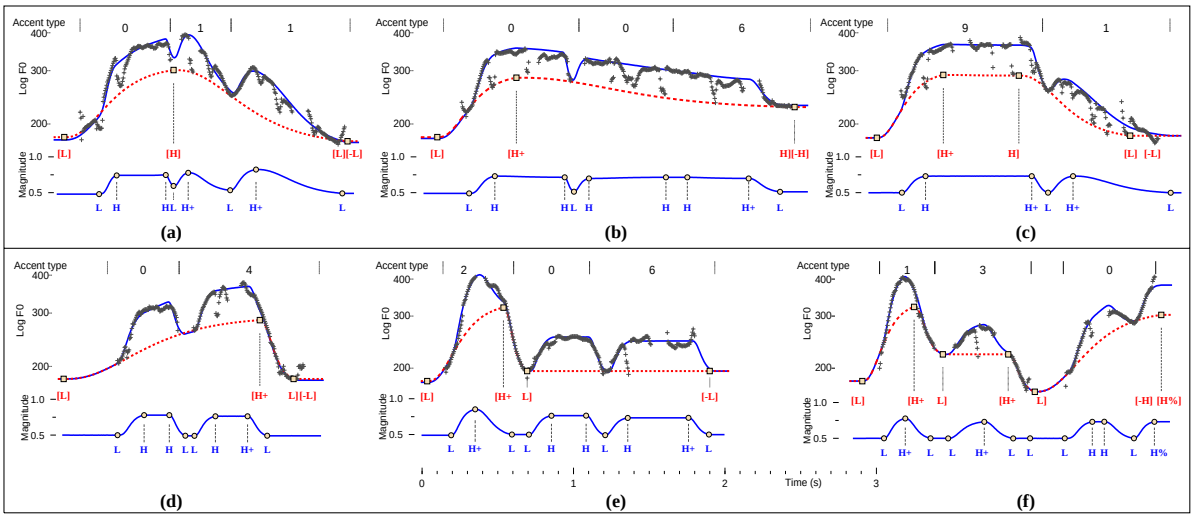

Fig. 5. *Typical structural forms of pitch register curves and corresponding accent and register tone labels. The dashed lines indicate the pitch register curves and the solid lines display the lexical accent components shown at the bottom of each panel.*

Table 1. *Objective test results.*

Metrics	RMSE	Correlation
Approximation vs. observed F_0's	13.6 Hz	0.973
Superposition vs. observed F_0's	16.4 Hz	0.961
Superposition vs. approximation	8.2 Hz	0.991

The number of the extracted lexical accent components is 16,970. Their mean magnitude is 3.84 semitones with standard deviation 0.229. Figure 4 shows an example of the extracted pitch register and lexical accent components. By confining the lexical accent components to a constant range (4 semitones), the dynamic pitch range variation is clearly captured by the pitch register as demonstrated in this example.

It is expected that the extracted pitch register curves are useful for detecting such prosodic events as intonation phrase boundary, relative prominences of accentual phrases, and the focus of an utterance. Most importantly, the lexical accent and pitch register curves themselves could be directly used to train separated component HMMs in speech synthesis using the superpositional HMM-based intonation synthesis [15]. The aspects of work will be done in the future.

An investigation is performed on the model-based representation of F_0 contours, focusing particularly on the effectiveness of the pitch register capturing the features of intonation proper. We select 300 utterances from the speech corpus and label the accent and register targets with visual inspection. One finding is that there exist more structural forms of "phrase component" in expressive speech than the one modeled by the Fujisaki model [5]. Figure 5 shows typical forms of pitch register configuration and corresponding accent and register tone labels. Generally speaking, the speaker appears to intentionally control the slope of pitch movements globally and locally as well as the pitch range. These distinct configurations fit words into prosodic phrases but do not change the word identity indicated by lexical accents.

We tentatively classify the configurations of pitch register curves into a limited number of structural forms. In Fig. 5 there exist three forms: form A (initial rise plus gradual decaying in Fig. 5 (a)), form B (high plateau in (b)(c)), and form C (gradual rise plus sharp fall in (d) (e) (f)). A count of structural forms among the 300 utterances shows that form A takes 16.0%, form B 17.3%, form C 55.2%, and others 11.5%. Form C is an essential way of making focal prominence regardless of the underlying accent types. This observation is confirmed by the perception of speech re-synthesized by using the pitch register curves. Note that in Japanese only form A is widely known as modeled by the Fujisaki model.

The structural forms of pitch register curves as shown in Fig. 5 have corresponding forms of intonation patterns in English [21]. Phrase register tone [-L] in Fig. 5(e), for example, works as phrase tone -L used in autosegmental metrical (AM) model [21][22]; it fills in the rest portion of a phrase after the last pitch accent. It should be noted that AM model does not assume the superposition principle in tone sequence and the phonetic tone implementation is under-specified [21].

4. CONCLUSION

This paper presents a method to automatically extract pitch register from observed F_0 contours for highlighting the expressive aspect of intonation. By the way of confining the lexical accent components to a limited constant range, the extracted pitch register curves can capture intonation proper. Particularly, the pitch register reveals several new structural forms of intonation configurations in Japanese that are very comparable with English intonation. Future work will apply the proposed method to improve expressive speech synthesis.

5. REFERENCES

[1] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in P. Cohen, J. Morgan, and M. Pollack, (eds). *Intentions in communication*, MIT Press, Cambridge MA, pp. 271-311, 1990.

[2] P. Taylor, *Text-to-speech synthesis*, Cambridge University Press, 2009.

[3] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook 3*, pp. 255–309, 1986.

[4] J. Venditti, K. Maekawa. and M. Beckman, "Prominence marking in the Japanese intonation system," in *The Oxford Handbook of Japanese Linguistics*, Oxford University Press. 2008.

[5] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn.*, No. 5, pp. 233-242, 1984

[6] J. 'tHart, R. Collier and A. Cohen, *A perceptual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge University Press, 1990.

[7] J. Santen and J. Hirschberg, "Segmental effect on timing and height of pitch contours," *Proc. of ICSLP1994*.

[8] A. Sakurai and K. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," in *Proc. of ICSLP1996*, pp. 817-820, 1996.

[9] H. Mixdorff, "A novel approach to the fully automatic extraction of fujisaki model parameters," in *Proc. of ICASSP 2000*, Vol.3, pp. 1281–1284, 2000.

[10] S. Narusawa, N. Minematsu, K. Hirose and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. of ICASSP 2002*, I-509 – I-512, 2002.

[11] J. van Santen, T. Mashira and E. Klabbers, "Estimating phrase curves in the general superpositional intonation model," in *Proc. of the 5th Speech Synthesis Workshop*, pp. 61–66, 2004.

[12] T. Mishira, "Decomposition of fundamental frequency contours in the general superpositional intonation model," PhD. thesis, the Oregon Health & Science University, 2008.

[13] H. Kameoka, K. Yoshizato, T. Ishihara, Y. Ohishi, K. Kashino, and S. Sagayama, "Generative modeling of speech F0 contours," in *Proc. of INTERSPEECH2013*, pp. 1826–1830. 2013.

[14] M. Langarani, E. Klabbers, and J. Santen, "A novel pitch decomposition method for the generalized linear alignment model," *Proc. ICASSP2014*, pp. 2603-2607, 2014.

[15] J. Ni, Y. Shiga, and C. Hori, "Superpositional HMM-based intonation synthesis using a functional F0 model," *Proc. of ISCSLP2014*, pp. 270-274, 2014.

[16] J. Ni, H. Kawai, and K. Hirose, "Constrained tone transformation technique for separation and combination of Mandarin tone and intonation", *J. Acoust. Soc. Amer.*, 119 (3), pp. 1764–1782, 2006.

[17] J. Ni and S. Nakamura, "Use of Poisson processes to generate fundamental frequency contours", in *Proc. of ICASSP2007*, pp. 825–828, 2007.

[18] E. Garding, "On parameters and principles in intonation analysis," Lund University, Dept. of Linguistics, Working Papers 40 (1993), pp. 25-47, 1993.

[19] H. Hashimoto, K. Hirose, and N. Minematsu, "Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis," in *Proc. of INTERSPEECH2012*. 2012.

[20] J. Ni, Y. Shiga, C. Hori and Y. Kidawara, "A targets-based superpositional model of fundamental frequency contours applied to HMM-based speech synthesis," in *Proc. of INTERSPEECH2013*, pp. 1052–1056, 2013.

[21] J. Pierrehumbert, *The phonology and phonetics of English intonation*, MIT Ph.D. dissertation, 1980.

[22] K.E.A. Silverman *et al* , "TOBI: A Standard for Labeling English Prosody," *Proc. of ICSLP92*, pp.867-870, 1992.