社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (13)
2006／12／21

F0パターンの自動推定と目標点の抽出

倪 晋富†‡ 中村 哲†‡

†独立行政法人情報通信研究機構知識創成コミュニケーション研究センター音声言語グループ
‡ATR音声言語コミュニケーション研究所 619-0288「けいはんな学研都市」光台2-2-2
E-mail: {jinfu.ni, satoshi.nakamura}@nict.go.jp

Automatic F0 Contour Fitting and Tonal Target Extraction

Jinfu NI†‡ and Satoshi NAKAMURA†‡

† National Institute of Information and Communications Technology, Kyoto, Japan
‡ ATR Spoken Language Communication Research Labs, Kyoto, Japan
E-mail: {jinfu.ni, satoshi.nakamura}@nict.go.jp

Abstract The fundamental frequency (F0) contours of voices manifest the acoustic correlation of lexical tones in tonal languages (or pitch accents in non-tonal languages), expressive intonation, and such speaker factors as individual vocal ranges. Tonal target points, focusing particularly on tonal F0 peaks and valleys, are used to sparsely specify the prosodic contributions of lexical tones (or pitch accents) and expressive intonation to the F0 contours. This paper presents an approach to automatic extraction of the underlying tonal target points with optimal approximations of the observed F0 contours using a functional F0 contour model. This consists of (1) finding potential target candidates from the observed F0 contours, (2) suppressing the micro-prosodic effects involved in the F0 contours, and (3) selecting appropriate target points from these candidates. This approach is currently evaluated by investigating the errors between the observed and reproduced F0 contours in the spirit of analysis-by-synthesis, conducted on thousands of Chinese speech samples and a few of Japanese and English speech samples. It shows a straightforward way to studying dynamic intonation variations, thanks to its ability to parametric representation of the F0 contours in term of the same way as specifying the prosodic contribution of tone and intonation to the F0 contours.

Keyword Prosody modeling, Poisson distributions, Intonation, F0 contour model, Communicative speech synthesis

1. Introduction

Significant progress during the past decades has been made in advancing speech synthesis technology providing a natural modality for human-machine interaction [1]. Despite this progress, some fundamental limitations in the technology have hindered its widespread use. In the current human-machine inactive speech dialog system, for example, the text-to-speech module is running in a mode with its prosody control almost independent of its dialog background. As a consequence, the synthetic speech fails to deliver intended communication functions. In order to make the machine partners more human-like, or make the synthetic speech more communicative, appropriate techniques must be developed to enable the synthetic speech to deliver the intended communication functions, which are basically conveyed by appropriate intonation patterns. It is widely recognized that the expressive intonation patterns are manifested by the fundamental frequency (F0) contours, in the same time, acoustically coupling with syllabic tones in tonal languages (or pitch accents in non-tonal languages) and speaker factors, such as individual vocal ranges [2][3]. From this view of points, measure of dynamic intonation variations from the observed F0 contours is an important issue in prosody modeling. It is important because dynamic intonation control forms the basis of communicative conversational speech synthesis. While the problem is difficult due to its complexity, however, a constrained tone transformation technique [2] has proven to be promising for solving the acoustic correlation up on the tonal targets based sparser specifications of the prosodic contributions to the F0 contours [4][5]. This leads us to extraction of the underlying tonal target points from observed F0 contours.

There exist several methods for modeling of observed F0 contours. Typically, one tendency is to formulate the physical mechanisms of the F0 generation processes [3]. Another one is to straightforwardly analyze the prosodic contributions of tone (or accent) and intonation to the F0 contours as a series of tonal target points [4][5]. To predict the shape of F0 contours, the transitions through these target points are interpolated by the spline functions. In the previous work [6], we have proposed a functional

model for structural modeling of the F0 contours to consider tone modulations in Mandarin. This model has key features of the two tendencies in that it basically places tonal target points as used in [4] and [5], and that the transitions between the target points are modeled by using the response functions of critically-damped second order linear systems originally suggested in [3]. In this paper, we will use an extended version of this functional model for representation of the observed F0 contours while extracting the underlying tonal target points.

The remainder of this paper is organized as follows. Section 2 describes the proposed approach, including an extended functional F0 contour model and the model based automatic F0 contour fitting for extraction of the underlying tonal target points. Section 3 presents the experimental results. Remarks are given in Section 4.

2. Description of the approach

This approach consists of making use of a functional model to represent the observed F0 contours and extracting the underlying tonal target points through automatic approximation of them by the use of a so-called analysis-by-synthesis technique. The basic framework of this functional model has been presented in [6]. Recently, we manage to extend the functional model by substituting the second order linear systems based filters with the Poisson-process-induced filters [7]. The former is a special case of the latter; the new filters can delay target points in time. That is, a target point can act as a target level, if necessary [8]. This extension makes it possible to simultaneously use the tonal target points as the model parameters to represent the F0 contours. This forms the basis of automatically extracting the tonal target points by optimal approximations of the F0 contours.

2.1. Extension of a functional F0 contour model

In the previous work [6], the F0 contours of voices are represented as a form of non-linear acoustic correlation of syllabic, phrasal, and speaker factors in the spirit of the resonance principle. Figure 1 illustrates the framework of this functional model. In this figure, $F_0(t)$ denotes the F0 contours as a function of time t, which belong to the vocal range $[f_{0b}, f_{0t}]$. Its prosodic contribution is assumed to have two components: $\zeta(t)$ a latent scale and $\Lambda(t)$ the stylized contours in the normalized range $[\lambda_b, \lambda_t]$. The latent scale is invoked to explain observed co-variation in tonal behaviors (such as the phenomena of declination, F0 range expansion and narrowing, relative level shifting upward and downward), thus expressing the intonation components as used in [2]. In a mathematical term,

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t),\zeta(t)) - A(\lambda_b,\zeta(t))}{A(\lambda_t,\zeta(t)) - A(\lambda_b,\zeta(t))},$$

where

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

which indicates a form of resonance curves (amplitude amplifying coefficients of a forced vibrating system); λ indicates the square frequency ratio of the forced vibrating system; and ζ the damping ratio of the system. Because of the symmetry $A(\lambda, \zeta) = A(2-\lambda, \zeta)$ and the peak of $A(\lambda, \zeta)$ occurring at $\lambda = 1$, only the right arm of resonance curves ($\lambda \geqq 1$) is employed here. Note that $\zeta^2 \leqq 0.5$. Accordingly, an F0 contour $F_0(t)$ is represented as a scale transformation of the stylized contours $\Lambda(t)$ scaled by $\zeta(t)$, corresponding to the syllabic tones fitting themselves with the sentence intonation within the vocal ranges.

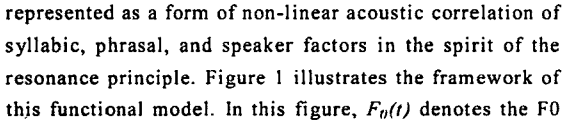

Fig. 1. The framework of a functional F0 contour model.

The prosodic contributions to the F0 contours are here analyzed into a series of tonal target points as [4] and [5]. Accordingly, the stylized contours shall be modeled through these tonal target points, while $\zeta(t)$ can be fixed at ζ_0 [6]. Assume there need n target points, denoted by (t_i, f_{0i}), $i = 1, \ldots, n$, to specify the prosodic contributions to the observed F0 contours, where t_i indicates timing and f_{0i} hertz. In the normalized range $[\lambda_b, \lambda_t]$, let (t_i, λ_i) denote the corresponding target points. Furthermore, a default target (t_0, λ_1) is assumed at $t_0 = 0$, provided $t_1 > 0$. Then the stylized contours are expressed as a simple concatenation of these local transitions:

$$\Lambda(t) = \sum_{i=0}^{n} \Lambda_i(t, t_i, \lambda_i, t_{i+1} - t_i, \frac{\lambda_{i+1} - \lambda_i}{1-\delta}, k_i),$$

where $\Lambda_i(t)$ denotes the transitions between the ith target point and the next and is expressed as follows.

$$\Lambda_i(t, t_i, \lambda_i, t_{i+1} - t_i, \Delta\lambda_i, k_i) = \begin{cases} \lambda_i + \Delta\lambda_i[1 - D(t - t_i, t_{i+1} - t_i, k_i)], & t_i \leq t < t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

$D(t-t_i, t_{i+1}-t_i, k_i)$ indicates the Poisson-process-induced functions [8]:

$$D(t,\Delta t,k)=\sum_{j=0}^{k}\frac{\left[\frac{c(k)t}{\Delta t}\right]^{j}}{j!}e^{-\frac{c(k)t}{\Delta t}},\ t\geq 0,$$

where k is regarded as a model parameter and $c(k)$ is determinative when given k. More specifically, $c(k)$ is determined by the following equation.

$$\sum_{j=0}^{k}\frac{[c(k)]^{j}}{j!}e^{-c(k)}=\delta,$$

where δ is a constant. In the previous experiments as described in [7], δ could be commonly set at 0.1. Table 1 shows part of $c(k)$ with respect to $\delta = 0.1$.

Table 1. Part of coefficients $c(k)$ with respect to $\delta = 0.1$.

k	0	1	2	3	4	5	6	7	8	9
$c(k)$	2.30	3.89	5.32	6.68	7.99	9.27	10.53	11.77	12.99	14.21
k	10	11	12	13	14	15	16	17	18	19
$c(k)$	15.41	16.60	17.78	18.96	20.13	21.29	22.45	23.61	24.76	25.90

For clarity, the notation used in this paper is summarized as follows.

$F_0(t)$: reproduced F0 contours.

$F'_0(t)$: observed F0 contours.

f_{0b}: bottom frequency of a vocal range in hertz.

f_{0t}: top frequency of the vocal range in hertz.

λ_b: bottom value of the normalized ranges; $\lambda_b = 2$.

λ_t: top value of the normalized ranges; $\lambda_t = 1$.

$\zeta(t)$: a latent sale; $\zeta(t) = \zeta_0$ and $\zeta_0 = 0.156$ is commonly used for transformations of observed F0 contours to the corresponding ones in the normalized ranges [6].

$\Lambda(t)$: the stylized contours with respect to $\zeta(t) = \zeta_0$.

δ : a constant; $\delta = 0.1$.

The essential model parameters are

(t_i, f_{0i}): the ith tonal target point, $i = 1, \ldots, n$.

n: number of tonal target points.

k_i:the ith accompanying parameter to delay target in time.

For simple citation, let us denote the transformations of f_{0i} to λ_i as $\lambda_i = T_\lambda(f_{0i}, \zeta_0)$, given the vocal range $[f_{0b}, f_{0t}]$. Note that a tonal target point (t_i, f_{0i}) in the frequency-time space is equal to (t_i, λ_i) in the λ -time space.

2.2. Extraction of tonal target points

Extraction of tonal target points is treated as the same process as estimation of the model parameters for optimal approximations of the observed F0 contours. The prosodic contributions of accentual and phrasal factors to observed F0 contours can be sparsely specified by a series of tonal target points (t_i, f_{0i}), $i = 1, \ldots, n$. In the framework of this extended functional model, these tonal target points (t_i, f_{0i}) can also represent the observed F0 contours with another accompanying parameter k_i. In Mandarin, for example, the number of tonal target points necessary for specifying the prosodic contributions can be predicted from the number of syllabic tones in the utterances [2]; a syllabic tone basically needs two target points. Therefore, we may assume that the number of tonal target points necessary for specifying the prosodic contributions to the F0 contours is given at the input of extraction algorithms.

Figure 2 shows the block-diagram of an algorithm for automatic extraction of tonal target points from the input of (a) observed F0 contours and (b) assumed target number n. As shown in Fig. 2, this algorithm consists of three logical steps. First, search all the possible target point candidates, from which n tonal target points will be selected according to the criterion of minimizing the weighted errors. The target point candidates are such voiced frames that satisfy one of the following constraints, say the jth (voiced) frame whose timing is denoted by t_j.

(1) The j-2 th or j+2 th frame is unvoiced. That is, the jth F0 point may be adjacent to unvoiced parts or pauses.

(2) The jth F0 point is a *turn* of local F0 movements as
rise-level: $F'_0(t_{j-1}) < F'_0(t_j)$ and $F'_0(t_j) = F'_0(t_{j+1})$;
rise-fall: $F'_0(t_{j-1}) < F'_0(t_j)$ and $F'_0(t_j) > F'_0(t_{j+1})$;
level-rise: $F'_0(t_{j-1}) = F'_0(t_j)$ and $F'_0(t_j) < F'_0(t_{j+1})$;

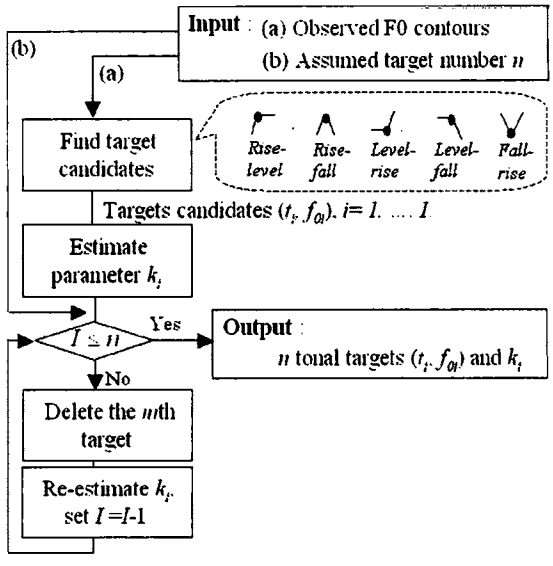

Fig. 2 Schematic block-diagram showing the procedure of automatic extraction of the underlying tonal target points from the observed F0 contours when given the tonal target number n in the input.

leve-fall: $F'_0(t_{j-1}) = F'_0(t_j)$ and $F'_0(t_j) > F'_0(t_{j+1})$;

fall-rise: $F'_0(t_{j-1}) > F'_0(t_j)$ and $F'_0(t_j) < F'_0(t_{j+1})$.

The initial number of target point candidates closely depends on individual observed F0 contours and thus is quite sensitive to possible micro-prosodic effects, local F0 fluctuations, and potential F0 extraction errors. To suppress such effects on the target extraction process, certain measures have been built in the extraction algorithm. A technique introduced here is to incorporate the local F0 fluctuations into the process of evaluating individual F0 contributions to the mismatching errors between the observed and reproduced F0 contours while estimating the underlying tonal target points. This is based on an assumption that tonal target points lie at the relative stable portions of an F0 contour. We define the degree of the jth F0 fluctuation as follows.

$$R(j) = [R(j-1, j)+R(j, j+1)]/2,$$

where

$$R(j,j+1)=\begin{cases}\dfrac{\max\left[F'_0(t_j),F'_0(t_{j+1})\right]}{\min\left[F'_0(t_j),F'_0(t_{j+1})\right]}, & \text{for voiced cases.}\\ \infty, & \text{otherwise}\end{cases}$$

Let N indicate the total number of frames in an utterance. Then the weighted errors between observed and reproduced F0 contours are defined as follows.

$$E\left(F_0(t),F'_0(t)\right)=\sum_{j=1}^{N}\left(F_0(t_j)-F'_0(t_j)\right)^2\times w(j),$$

$$\text{where } w(j)=\exp\left(-\alpha\times\ln\frac{R(j-1,j)+R(j,j+1)}{2}\right).$$

Generally, $w(j)$ = 1 if and only if $F'_0(t_{j-1}) = F'_0(t_j)$ and $F'_0(t_j) = F'_0(t_{j+1})$; $w(j)$ = 0 when any adjacent frame is unvoiced. Furthermore, the weights can be adjusted by selecting appropriate values to α, a decaying factor. A measure of weighted errors is intended to emphasize relatively stable F0 portions. As shown in Fig. 3, voiced frames with high degree of local F0 fluctuations R(j) will lower their contributions to the *mismatching* errors.

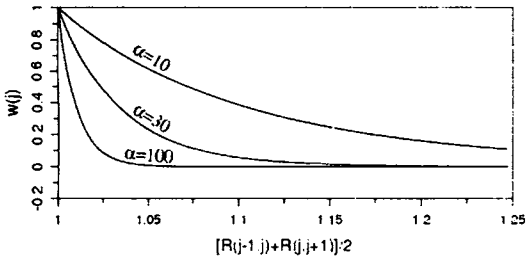

Fig. 3. Weight $w(j)$ as functions of degree of local F0 fluctuations R(j) with varied decaying factor α values.

Second, estimate parameters k_i when given a set of target point candidates, i = 1, ..., I. This is basically to minimize the weighted errors in an iteration process; k_i can be limited to a range from 2 to 15 [7].

Third, select n target points from the candidate point set whose size is I by deleting I-n target candidates step by step. In each cycle or step, a target point candidate is deleted, saying the mth tonal target point candidate. More particularly, this target point candidate is selected in the following way. For each candidate point in the current candidate set, it is taken out from the set in turn, saying the ith candidate. Then the two fragments [from (t_{i-1}, f_{0i-1}) to (t_i, f_{0i}) and (t_i, f_{0i}) to (t_{i+1}, f_{0i+1})] would merge into a fragment. Thus the accompanying parameter k_{i-1} related to the candidate point (t_{i-1}, f_{0i-1}) is re-estimated for the merged fragment. Consequently, a newly reproduced contour results, denoted by $F_0(t)_{I-i}$. The mth target candidate is then determined according to the following criterion.

$$m=\arg\min_{1\le i\le I} E\{F_0(t)_{I-i}, F'_0(t)\}$$

That is, the mth candidate point gives the minimum contribution to the approximations of the F0 contours in the criterion of minimizing the weighted errors. This step is repeated until the number of the candidate points I is equal to (or less than) n, the assumed number of tonal target points given in the input. The final output is the n tonal target points (t_i, f_{0i}) and the accompanying parameter k_i. Of course, they can be in turn used to reproduce the F0 contours using the functional model.

3. Experimental results

The validity of an approach can be tested by its ability to analyze observed samples. For this purpose, we adopt 5,779 speech samples from existing speech corpora. Part of Mandarin samples is taken from CoSS-1, and the others and Japanese and English samples from a multilingual speech corpus built at ATR. The approach is currently evaluated in two experiments. Experiment 1 is aimed at automatic extraction of specified number of tonal target points from the observed F0 contours in Mandarin. The evaluation is based on an investigation of the errors between observed and reproduced F0 contours using these target points. Experiment 2 tests the approach applied to the other languages. In the experiments, measured F0 contours are interpolated at an interval of 5 ms. Vocal ranges $[f_{0b}, f_{0t}]$ are simply fixed at [100 Hz, 450 Hz] for female speakers and [50 Hz, 250 Hz] for male speakers. No correction of F0 extraction errors, if any, is made.

3.1. Experiment 1

In this experiment, Mandarin speech samples include

(i) 2,000 tri-syllabic and 2,049 quadri-syllabic isolated words produced by a native male from CoSS-1, and

(ii) 1,680 dialog sentences uttered by a native female.

In the sample set, all the tone combinations were well balanced at the phase of the prompt design. The experiment procedure is roughly described as follows.

(i) Predict target number n for each utterance by doubling the number of syllabic tones in the utterance (no count of light tones). Note that the number of target points is rather limited for representation of the F0 contours.

(ii) Extract target points using the proposed approach.

(iii) Reproduce the F0 contours based on the target points.

Figure 4 shows percentages of the voiced frames as a function of voiced frame F0 errors: $F'_0(t_j)$-$F_0(t_j)$ for the jth voiced frame. For the 5,729 utterances, more than 76% of the voiced frames drop into an error interval [-5 Hz, 5 Hz], namely, 86.3% for the male's samples and 76.9% for the female's samples; as for [-10 Hz, 10 Hz], 93.8% and 89.0% for each. A good symmetry of the *curves* shown in Fig. 4 also indicates that the reproduced F0 contours can trace the observed ones well. Figure 5 shows an example of observed F0 contours and the optimal approximations with three target points, (0.16 s, 176 Hz), (0.51 s, 89 Hz), and (0.73 s, 133 Hz). Table 2 lists the full parameters used for the optimal approximations. Note that $\lambda_i = T_\lambda(f_{0i}, \zeta_{0i})$, i = *1, 2, 3*. Though, some examples also impart that the algorithm's limitation at the simple strategy of selecting tonal target points from the potential candidates.

3.2. Experiment 2

Experiment 2 is conducted on 30 Japanese and 20 English utterances. In short, given appropriate input of target number n individually, this approach could also find the optimal approximations of these observed F0 contours by the same way as used in Mandarin. Examples are shown in Fig. 6 in Japanese and Fig. 7 in English. The tonal target points and the accompanying parameter k_i are listed in Table 3. Let us take Fig. 7 as an example to show the ability of the weighted error measure to suppress the micro-prosodic effects. The sagging between the 3rd and 4th targets, for example, is micro-prosodic and was suppressed by the measure of weighted errors. Of course, the sagging portions could be assigned with target points when n is larger than 9. This may pose a question: how to measure the confidence of extracted tonal target points? Further work is needed in this aspect.

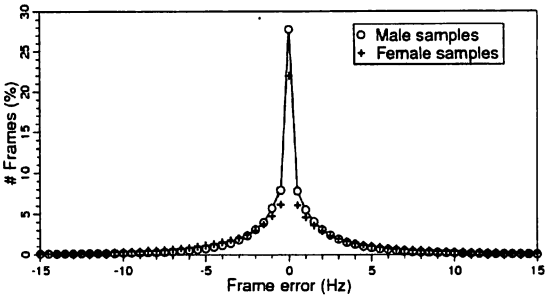

Fig. 4. Percentages of the voiced frames as a function of F0 errors caused from the sparser representations.

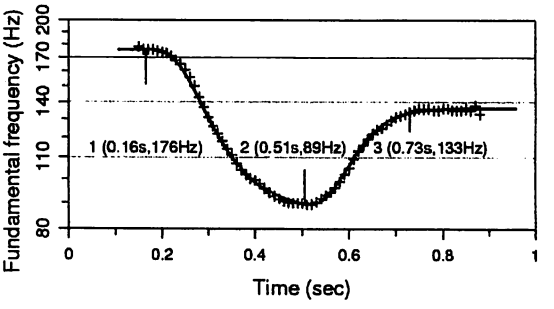

Fig. 5 Example of observed F0 contours ("+" sequences) and the optimal approximations (solid lines) through the three target points indicated by the short vertical lines.

Table 2 Parameters for the approximations shown in Fig.5

λ_0	λ_2	δ	$\zeta(t)$	ζ_0	f_l (Hz)	f_t (Hz)	i	t_i (s)	f_{0i} (Hz)	λ_i	k_i
2	1	0.1	0.156		50	250	1	0.16	176	1.20	3
							2	0.51	89	1.48	3
							3	0.72	133	1.30	3

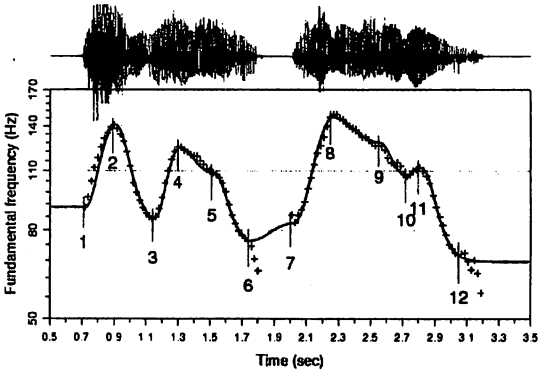

Fig. 6 Approximations (solid lines) of an observed F0 contour (''+" sequences) for the Japanese sentence "*aoiaoinoeha yamanouenoieniaru.*" uttered by a native male. The short vertical bars indicate the position of target points (t_i, f_{0i}), i = 1, ..., 12, whose values are listed in Table 3 (J row).

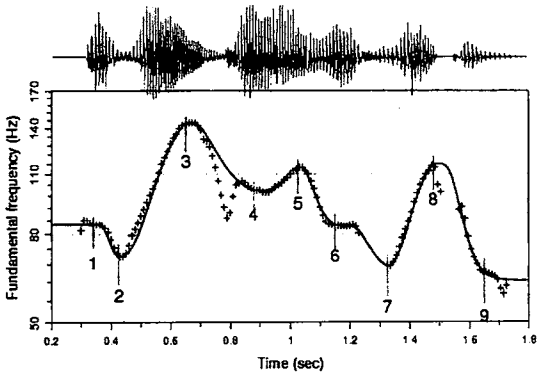

Fig. 7 Approximations (solid lines) of an observed F0 contour (``+" sequences) for the English sentence "*A whole joy was reaping.*" uttered by an American. The short vertical bars indicate the position of target points (t_i, f_{0i}), i = 1, ..., 9, which are listed in Table 3 (E row).

Table 3 Model parameters for optimal approximations of the observed F0 contours shown in Figs. 6 and 7.

	i	1	2	3	4	5	6	7	8	9	10	11	12
J	t_i (s)	0.71	0.89	1.14	1.30	1.51	1.74	2.00	2.25	2.55	2.72	2.80	3.05
	f (Hz)	90	141	85	125	109	75	83	147	128	107	112	70
	k_i	3	3	3	3	3	3	3	3	3	3	3	3
E	t_i (s)	0.34	0.43	0.65	0.88	1.02	1.15	1.32	1.48	1.65			
	f (Hz)	84	71	144	101	114	84	67	116	65			
	k_i	9	2	3	13	3	10	2	7	2			

4. Remarks and future work

This paper presented an approach to automatic fitting of observed F0 contours and extraction of the underlying tonal target points in a unified framework. This is based on an extension of the functional F0 contour model [6]. It is this extension that makes it possible to represent the observed F0 contours in terms of the same observed tonal F0 peaks and valleys (so-called tonal target points) as usually used for sparser specifications of the prosodic contributions to the F0 contours [4][5][2]. The evaluation experiments were conducted on 5,779 speech samples, most in Mandarin and a few in Japanese and English. The results indicated that the proposed method could properly find the underlying tonal target points, which are believed to capture dynamic intonation variations. This is evidenced by the experimental results for accurately predicting the shape of the F0 contours when given the sparser specifications of the prosodic contributions of tone and intonation to them. Also, it appears to us that the use of the Poisson process to generate F0 contours is appropriate regardless of tonal or non-tonal languages, provided that the prosodic contributions to the observed F0 contours can be analyzed into a series of tonal target points.

The proposed approach is currently evaluated with investigation of the voiced frame F0 errors between the observed and reproduced F0 contours in read speech, undertaken in the spirit of analysis by synthesis. It remains to see how the extracted tonal target points are related to the communication functions conveyed by the speech. This will be interesting and very important in the context of communicative conversational speech synthesis. Considering the algorithm for extracting the tonal target points, structural constraints as used in [3] and phonetic labeling of the signal shall be used in the future work.

References

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multi-lingual speech-to-speech translation system," *IEEE Trans. on Speech and Audio Processing*, Vol. 14, No. 2, pp. 365-376, 2006.

[2] 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.

[3] H. Fujisaki, K. Hirose, " Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn.* (E), 5 (4), pp. 233-242, 1984.

[4] J. B. Pierrehumbert, "Synthesizing intonation, " *J. Acoust. Soc. Amer.*, 70 (4), pp. 985-995, 1981.

[5] D. Hirst, A. D. Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and Experiment*, edited by M. Horne (Kluwer Academic Publishers), pp. 51-87, 2000.

[6] J. Ni, K. Hirose, " Quantitative and structural modeling of voice fundamental frequency contours of speech in Mandarin, " *Speech Communication*, 48 (8), pp. 989-1008, 2006.

[7] J. Ni, S. Nakamua, "Modeling local *F0* patterns upon Poisson processes," ***日本音響学会講演論文集***, pp. 203-204, Sept. 2006.

[8] J. Ni, S. Nakamura, "Use of Poisson processes to generate fundamental frequency contours," (forthcoming).