USE OF POISSON PROCESSES TO GENERATE FUNDAMENTAL FREQUENCY CONTOURS

Jinfu Ni †,‡ *and Satoshi Nakamura*†,‡

†National Institute of Information and Communications Technology, Japan
‡ATR Spoken Language Communication Research Labs, Japan

ABSTRACT

The prosodic contributions to voice fundamental frequency (F_0) contours can be analyzed into a series of sparser tonal targets (F_0 peaks and valleys). The transitions through these targets are interpolated by spline or filtering functions to predict the shape of F_0 contours. A functional model was proposed in the previous work for this purpose. This paper presents an enhanced version of this model achieved by replacing its decay filter with a Poisson-process-induced filter. It is enhanced because the former is a special case of the latter. The new filter manages to delay the decaying process while interpolations are being uttered. A target point can thus act as target levels, if necessary. The algorithms for estimating parameters, which were implemented on computers, are also presented. Experiments conducted on thousands of observed F_0 contours, including Mandarin, Japanese, and English, indicate that the enhanced version significantly facilitates their automatic parameterization.

Index Terms— Prosody modeling, Poisson distributions, Voice conversion, Speech synthesis, Speech processing

1. INTRODUCTION

One tendency in modeling F_0 contours is to formulate the physical (or physiological) mechanisms for the processes generating F_0 [1][2]. It is also more straightforward to analyze the prosodic contributions (of lexical tones/accents and sentence intonation) to the F_0 contours as a series of target points [3][4] (and others). The target points basically focus on tonal F_0 peaks and valleys, which will be referred to as *tonal targets* after this. The transitions through these target points are interpolated by either filters [3] or spline functions [4] to predict the shape of the F_0 contours. A functional model was proposed in the previous work [5] to structurally model the F_0 contours to take tone modulations into account. This model has various features of the two tendencies in that it basically places tonal target points, following [3] and [4], and that the transitions between targets are interpolated by a filter, i.e., the response functions of critically-damped second order linear systems as suggested in [1]. This filter can be regarded as being associated with Poisson distributions. We developed this model by considering the generation of F_0 contours as a Poisson process. This paper also presents the algorithms we used to estimate parameters and experimental results.

2. ASSUMPTION: A FUNCTIONAL F0 MODEL

In [5], an F_0 contour of $F_0(t)$ as a function of time t is represented as a scale transformation of local components $\Lambda(t)$ from normalized range $[\lambda_b, \lambda_t]$ in λ ($\geq$ 1) to vocal range $[f_{0_b}, f_{0_t}]$ (bottom and top frequencies in hertz). Particularly,

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

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

which physically means the amplitude amplifying coefficients of forced vibration, i.e., λ, squared frequency ratio, and ζ, the damping ratio. While fitting the F_0 contour, ζ is fixed at ζ_0. After this, $\zeta_0 = 0.156$, $\lambda_t = 1$, and $\lambda_b = 2$. Consequently, $F_0(t)$ can be mapped to, one-to-one, $\Lambda(t)$, given $[f_{0_b}, f_{0_t}]$. For simplification, let us define the following terms [5].

$F_0(t) = T_{f_0}(\Lambda(t), \zeta_0)$, computing $F_0(t)$ from $\Lambda(t)$, and
$\lambda = T_\lambda(f_0, \zeta_0)$, computing λ from $f_0 \in [f_{0_b}, f_{0_t}]$.

The prosodic contributions of syllabic tones and sentence intonation to the F_0 contours can be analyzed into a series of sparser tonal target points [6]. Thus, $F_0(t)$, or $\Lambda(t)$, is the most appropriate contour through these target points.

Assume n target points are denoted by (t_i, f_{0_i}), or (t_i, λ_i), $\lambda_i = T_\lambda(f_{0_i}, \zeta_0)$, $i = 1, ..., n$, to specify an observed F_0 contour, where t_i indicates time and f_{0_i} hertz. Let $\Lambda_i(t)$ denote the transitions between the *i*th target point and the next. A default target (t_0, λ_0) is assumed at $t_0 = 0$, provided $t_1 > 0$. Assume $\lambda_0 = \lambda_1$ and $t_{n+1} = \infty$. $\Lambda(t)$ is then expressed as

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

where $\Lambda_i(t, t_i, \lambda_i, \Delta t_i, \Delta\lambda_i)$ [or simply denoted by $\Lambda_i(t, ...)$]

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

$$\text{where } D(t, \Delta t) = \left(1 + \frac{4.8t}{\Delta t}\right) e^{-\frac{4.8t}{\Delta t}},\ t \geq 0. \quad (2)$$

Given coefficient 4.8 in Eq. (2), δ is 0.05 [5]. The core part is a decaying process $D(t, \Delta t)$ changing from 1 to δ, or $1 - D(t, \Delta t)$ from 0 to $1 - \delta$. When given Δt, the decaying process $D(t, \Delta t)$ is deterministic. In the rest of this paper, $D(t, \Delta t)$ is revised from the view of a stochastic process.

3. DESCRIPTION OF APPROACH

3.1. Inter-relation of Poisson process and Fujisaki model

F0 may be regarded as a count of the opening-closing cycles of the glottis per second. Mathematically, this is a point counting process, and the most familiar stochastic point process is perhaps the Poisson process. Consider a Poisson process where in an infinitesimal time interval an event happens with sufficiently small probability β independent of events outside the interval. Let $N(t)$ be the number of events that occur before time t. Then $N(t)$ obeys the Poisson (βt) distribution:

$$P\{N(t) = k\} = \frac{(\beta t)^k}{k!} e^{-\beta t}, \;\; k = 0, 1, 2, ...$$

A superposition of multiple Poisson processes is still a Poisson process. The cumulative distribution function is

$$P\{N(t) \leq k\} = \sum_{j=0}^{k} \frac{(\beta t)^j}{j!} e^{-\beta t}.$$

Here, we do not discuss what an event is or the physiological origin that leads to the Poisson process in F_0 control. Let us consider the case $k = 2$. We have

$$P\{N(t) \leq 2\} = (1 + \beta t)e^{-\beta t} + 0.5t \times \beta^2 t e^{-\beta t}. \qquad (3)$$

The first term on the right hand side of Eq. (3) is the core of the accent control mechanism in the Fujisaki model [1], and the second term may approximate the phrase control mechanism, i.e., $\alpha^2 t e^{-\alpha t}$ in [1]. Recalling the functional model, the first term on the right hand side of Eq. (3) indicates filter $D(t, \Delta t)$, given $\beta = 4.8/\Delta t$. Based on these observations, it is clear that the F_0 generation process described by the Fujisaki model is perhaps a Poisson process with $k = 2$ and the decaying process in Eq. (2) an exact Poisson process with $k = 1$. This implies that the F_0 generation process may be treated as a Poisson process. Consequently, there is a generalized filter with varied k values as well as $D(t, \Delta t)$.

3.2. Poisson-process-induced filter

A Poisson-process-induced filter is expressed as

$$D(t, \Delta t, k) = \sum_{j=0}^{k} \frac{[\frac{c(k)t}{\Delta t}]^j}{j!} e^{-\frac{c(k)t}{\Delta t}}, \;\; t \geq 0, \qquad (4)$$

where $c(k)$ is a k-dependent coefficient and is determined by solving equation $\sum_{j=0}^{k} \frac{[c(k)]^j}{j!} e^{-c(k)} = \delta$. After this, $D(t, \Delta t)$ in Eq. (2) will be substituted with $D(t, \Delta t, k)$ in Eq. (4), and k is a model parameter. Figure 1 shows the dynamics of $D(t, \Delta t, k)$ with varied k values, given Δt = 4 and $\delta = 0.1$. We can see from this figure that:

(i) $D(t, \Delta t, k)$ consistently decays from 1 to δ during the time period [0, Δt] regardless of the k values.

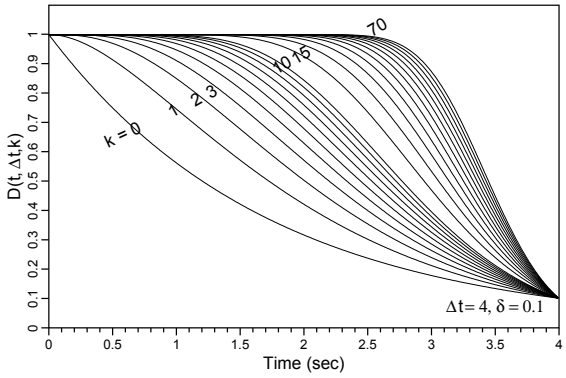

Fig. 1. Schematic view of dynamics of Poisson-process-induced filter with parameter k; $\Delta t = 0.4$, $\delta = 0.1$.

(ii) Poisson-process-induced filter $D(t, \Delta t, k)$ yields many interpolations from 1 to δ as well as $D(t, \Delta t)$ in Eq. (2) (i.e., $k = 1$). More interpolations facilitate approximations of more non-linear F_0 phenomena.

(iii) Parameter k achieves a *time delay* for the decaying process from 1 to δ. There is always a certain *time delay* except for $k = 0$. Note that estimating k values from the *time delay* for speech synthesis will be not discussed in this paper.

3.3. Estimation of parameters

The model parameters estimated from the observed F_0 contours can be found using two algorithms. The first is used to estimate parameter k, given target points. The second is used to detect the target points that yield optimal approximations of the F_0 contours.

Let $\hat{F}_0(t)$ denote an observed F_0 contour, $F_0(t)$ the model-based approximations, N the number of frames in $\hat{F}_0(t)$, and t_j the time for the jth frame. $\hat{F}_0(t_j) = 0$ if the jth frame is unvoiced. The measure of weighting errors between observed and reproduced F_0 contours is defined as

$$E\left(F_0(t), \hat{F}_0(t)\right) = \textstyle\sum_{j=1}^{N} [F_0(t_j) - \hat{F}_0(t_j)]^2 \times w(j),$$

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

where α is a decay factor that took 30 in the experiments. $R(j, j+1) = \max[\hat{F}_0(t_j), \hat{F}_0(t_{j+1})] / \min[\hat{F}_0(t_j), \hat{F}_0(t_{j+1})]$, if both the jth and $j+1$th frames are voiced, and $R(j, j+1) = \infty$ otherwise. Generally, $0 \leq w(j) \leq 1$; $w(j) = 1$ if and only if $\hat{F}_0(t_j)$ = $\hat{F}_0(t_{j-1})$ and $\hat{F}_0(t_j)$ = $\hat{F}_0(t_{j+1})$; $w(j) = 0$ if any of the $j-1$th, jth, or $j+1$th frames is unvoiced. Weighting $w(j)$ is intended to suppress local F_0 fluctuations, such as micro-prosodic effects and isolated F_0 extraction errors.

3.3.1. Estimation of parameter k

An algorithm is described below to estimate parameter k for each target by minimizing errors $E(F_0(t), \hat{F}_0(t))$, given a series of target points for sparsely specifying $\hat{F}_0(t)$.

Algorithm A: Estimation of k_i for target point (t_i, f_{0_i})

Input: • Observed F_0 contours $\hat{F}_0(t)$;
• Vocal range [f_{0_b}, f_{0_t}]; and
• The target points, (t_i, f_{0_i}), $i = 1, ..., n$.

Initial: Compute $\lambda_i = T_\lambda(f_{0_i}, \zeta_0)$, $i = 1, ..., n$. Set $i = 1$.

Loop 1: Set $\dot{k}_i = 1$ and $\epsilon_{min} = \infty$.

Loop 2: • Compute $\Lambda_i(t, ...)$ with $D(t, t_{i+1} - t_i, \dot{k}_i)$.
• Compute $F_{0_i}(t)$=$T_{f_0}(\Lambda_i(t, ...), \zeta_0)$, $t_i \leq t < t_{i+1}$.
• Compute $\epsilon = E\left(F_{0_i}(t), \hat{F}_0(t)\right)$ for $t_i \leq t < t_{i+1}$.
If $\epsilon < \epsilon_{min}$, set $\epsilon_{min} = \epsilon$ and $k_i = \dot{k}_i$.
$\dot{k}_i = \dot{k}_i + 1$. If $\dot{k}_i \leq k_{max}$ (e.g., 15), go to *Loop 2*.
$i = i + 1$. If $i \leq n$, go to *Loop 1*.

Output: k_i is associated with the ith target point, $i = 1, ..., n$.

3.3.2. Detection of potential target points

Given observed F_0 contours, this is still not sufficient for estimating the relevant target points to them if other information is unavailable. As a trade-off with blind estimation, here, we have assumed that the *total* number of target points for specifying an observed F_0 contour has been given; it may be able to be predicted from text information, such as lexical tones. An analysis-by-synthesis (AbS) algorithm with this model is described below for estimating potential targets in position achieved by minimizing errors.

Algorithm B: Detection of potential target points

Input: • Observed F_0 contours $\hat{F}_0(t)$; $\hat{F}_0(t_j), j = 1, ..., N$;
• Vocal range [f_{0_b}, f_{0_t}]; and
• Terminal condition: the number of target points, n.

Initial: Compute weighting $w(j), j = 1, ..., N$, from $\hat{F}_0(t)$.

Step 1: Find target candidates, say the jth voiced frame, if
• The $j - 2$th or $j + 2$th frame is unvoiced; or
• The jth frame is a *turn* of local F_0 contours from (i) *rise* to *level*; (ii) *rise* to *fall*; (iii) *level* to *rise*; (iv) *level* to *fall*; (v) *fall* to *rise*; or (vi) *fall* to *level*, where *rise*: $\hat{F}_0(t_j) < \hat{F}_0(t_{j+1})$, *level*: $\hat{F}_0(t_j) = \hat{F}_0(t_{j+1})$, and *fall*: $\hat{F}_0(t_j) > \hat{F}_0(t_{j+1})$.

Step 2: There are I target candidates (t_i, f_{0_i}), $i = 1, ..., I$.
• Estimate k_i for all the target points using *Algorithm A*.
• Compute error $\epsilon = E\left(F_0(t), \hat{F}_0(t)\right)$.

Step 3: If $I \leq n$, go to *Output*.
• $m = \text{argmin}_{1 \leq i \leq I}\left[E\left(F_0(t)_{I-i}, \hat{F}_0(t)\right) - \epsilon\right]$, where $F_0(t)_{I-i}$ indicates the approximations if the ith target point were deleted and k_{i-1} re-estimated.
• Delete the mth target point, and re-estimate k_{m-1}.
• Compute error $\epsilon = E\left(F_0(t), \hat{F}_0(t)\right)$. Go to *Step 3*.

Output: Target points (t_i, f_{0_i}) and k_i, $i = 1, ..., I$ ($\leq n$).

This AbS algorithm first estimates all the possible target candidates in *Step 1*, then iteratively deletes candidates with the minimum error increment in each loop until the number of target points n is derived, as described in *Step 3*. The target points are simply assumed to lie on the observed F_0 contours.

4. EXPERIMENTAL RESULTS

The validity of an approach can be tested by its ability to analyze observed samples. We adopted 5,779 speech samples extracted from existing speech corpora for this purpose. Mandarin samples were partly extracted from CoSS-1, and the others from a multilingual speech corpus at ATR. The evaluation was done in three experiments. Experiment 1 investigated parameters δ and k. Through Experiment 2, we evaluated the reliability of this Poisson-process-based approach by using automatic approximations of Mandarin F_0 contours. Experiment 3 was conducted on 50 utterances in English and Japanese to further examine the assumption of using the Poisson process for generating F_0 contours. The measured F_0 contours were interpolated in frame intervals of 5 ms. No corrections for F_0 extraction errors, if any, were made.

4.1. Experiment 1: Investigation of parameters δ and k

The speech samples were 279 Mandarin isolated tri-syllabic words from CoSS-1. All the words consisted of voiced segments except for the first initial consonants; thus, the observed F_0 contours were continuous over time. Combinations of lexical tones were balanced. These samples were produced by a native male whose vocal range was fixed at [50 Hz and 250 Hz] in the experiments. The target points were manually determined at the peaks and valleys of F_0 contours. In this experiment, δ was 0.01, 0.05, and 0.1, in turn. For the value set to δ, parameter k was re-estimated for each target point using *Algorithm A* where k_{max} was set at 100 for test purposes.

Figure 2 shows the experimental results. Two observations can be made. (i) δ is preferred to 0.1 than either 0.05 or 0.01. For $\delta = 0.1$, the average absolute error is 2.97 Hz, and 92.9% of the voiced frames drop into an error interval [–5 Hz and 5 Hz]. (ii) Varied k values are necessary for the interpolations between targets. This is evident in that the mean of k values is 6.78 with a standard deviation of 7.66. After this, δ is fixed at 0.1.

4.2. Experiment 2: Automatic F_0 contour approximations

Mandarin speech samples were used in this experiment, including (i) 2,000 three-syllabic and 2,049 four-syllabic words produced by a native male, as used in Experiment 1, and (ii) 1,680 dialog sentences uttered by a native female whose vocal range was fixed at [100 Hz and 450 Hz] (a slight extension of the frequency ranges measured from these samples). Tone combinations were balanced at the phase of prompt design.

This experiment can roughly be described as follows. (i) We predicted target number n for each utterance by doubling the number of syllabic tones in the utterance (light tones were not counted). Two targets were usually required for a lexical tone [6]. (ii) We estimated the target position for each utterance using *Algorithm B* with input n.

Figure 3 plots the percentage of voiced frames as a function of frame errors: $\hat{F}_0(t_j) - F_0(t_j)$ at the jth voiced frame. More than 76% of the voiced frames for the 5,729 utterances

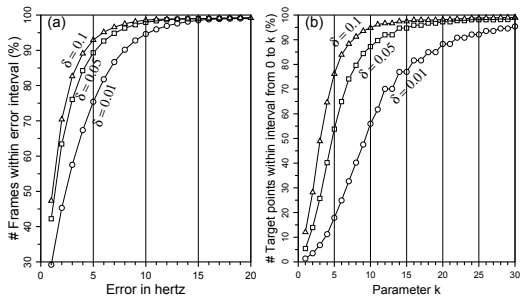

Fig. 2. Counts of voiced frames in percent according to F_0 error intervals in (a) and counts of target points according to intervals of k values in (b). δ is set at 0.01, 0.05, and 0.1.

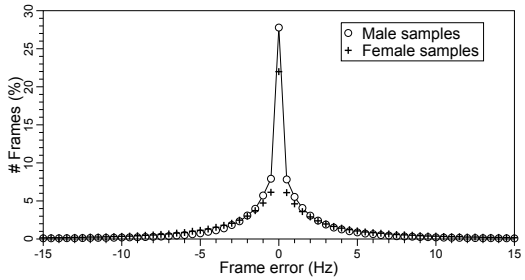

Fig. 3. Percentage of voiced frames as function of errors for experiment on automatic F_0 contour approximations.

drop into an error interval of [–5 Hz and 5 Hz], i.e., 86.3% for the male's samples and 76.9% for the female's. For [–10 Hz and 10 Hz], these are 93.8% for the male's and 89.0% for the female's. The excellent symmetry in the *curves* in Fig. 3 also indicates the model-based reproductions accurately trace the observed F_0 contours. However, the algorithm may need to be optimized to improve its strategy of selecting appropriate target points from potential candidates.

4.3. Experiment 3: Testing non-Mandarin F_0 contours

Another experiment similar to Experiment 2 was conducted on 30 Japanese and 20 English utterances. In short, given an appropriate input of target number n for each utterance, *Algorithm B* could find optimal approximations of the F_0 contours by minimizing weighting errors. There is an example in Fig. 4 and the corresponding parameters are listed in Table 1. As we can see from the figure, the micro-prosodic effects, such as the valley between the 3rd and 4th targets, on target detection were suppressed by the measure of weighting errors.

Table 1 Parameters for approximations in Fig. 4.

i	1	2	3	4	5	6	7	...
t_i (sec)	0.34	0.43	0.65	0.88	1.03	1.15	1.33	...
f_{0_i} (Hz)	84	71	144	101	114	84	67	...
λ_i	1.52	1.63	1.28	1.42	1.37	1.52	1.67	...
k_i	9	2	3	13	3	10	2	...

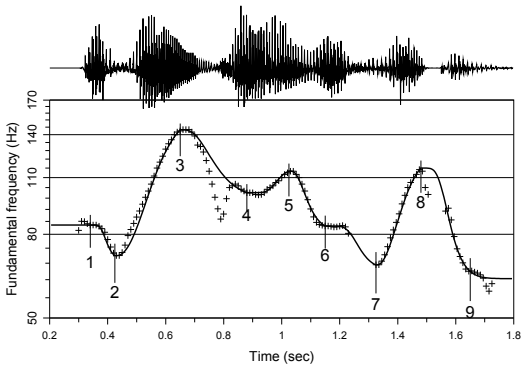

Fig. 4. Approximations (solid lines) of observed F_0 contour ("+" sequences) for English sentence "A whole joy was reaping." uttered by an American. The short vertical lines indicate the position of target points (t_i, f_{0_i}), $i = 1, ..., 9$.

5. CONCLUSIONS

This paper presented an enhanced functional model that can optimally represent F_0 contours, undertaken by AbS. Qualitative and quantitative analyses revealed that the process of generating F_0 contours may have the simple mathematical structure of the Poisson process; this consistently yields a family of constrained decaying exponential functions that are widely used for generating F_0 contours [1]. In practice, this enhancement makes it possible to find underlying target points through tracing the dynamics of F_0 movements as a stochastic process. Thus, it is more promising for automatic parameterization of the observed F_0 contours than the previous model [5]. However, the problem of prosody modeling is difficult because it is multidisciplinary, heavily involving linguistics and acoustics. It remains to be seen how the detected target points are related to communication factors. Constraints as discussed in [1] and [6] and phonetic labeling of the signal are expected to be useful for improving the confidence of target point detection in future work.

6. REFERENCES

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

[2] S. Prom-on, Y. Xu, and B. Thipakorn, 2006. "Quantitative target approximation model: simulating underlying mechanisms of tones and intonations," *ICASSP2006*, I-749 – I-752.

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

[4] D. Hirst, A. D. Cristo, and R. Espesser, 2000. "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), 51-87.

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

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