Selection of Training Data for HMM-based Speech Synthesis from Prosodic Features - Use of Generation Process Model of Fundamental Frequency Contours -

Tomoyuki Mizukami[1], Hiroya Hashimoto[2], Keikichi Hirose[1], Daisuke Saito[1], and Nobuaki Minematsu[2]

[1] Graduate School of Information Science and Technology,
[2] Graduate School of Engineering,
University of Tokyo

{mizukami, hiroya, hirose, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

10.21437/SpeechProsody.2014-198

Abstract

Generation process model of fundamental frequency (F_0) contours is ideal to represent global movements of F_0's keeping a clear relation with back-grounding linguistic information of utterances. Using the model, improvements of HMM-based speech synthesis are expected. A new method is developed to cope with erroneous F_0's of utterances included in HMM training corpus. F_0 extraction errors not only cause wrong F_0's, but also degrade segmental features of synthetic speech, since they affect the over-all accuracy of speech analysis. The method is to exclude speech segments from HMM training, where extracted F_0's are largely different from those generated by the generation process model. Experiments on speech synthesis showed a clear improvement in synthetic speech quality when phoneme-based exclusion is conducted with a properly selected threshold.

Index Terms: F_0 contour, generation process model, HMM-based speech synthesis, training segment selection

1. Introduction

HMM-based speech synthesis attains a special attention from researchers in speech communication field, since it can generate a good quality of speech from a rather limited size of speech corpus with flexible controls on voice qualities and utterance styles [1]. It can handle fundamental frequencies (F_0's) of speech in the same frame-by-frame manner with other acoustic features such as mel-cepstral coefficients. It is easy to prepare a training corps, since F_0 values can be directly used for HMM training without assuming any models of F_0 contours. However, this in turn causes demerits; it generally produces over-smoothed F_0 contours with occasional F_0 undulations not observable in human speech. Automatic extraction of F_0's occasionally outputs erroneous results; wrong F_0's and voiced/unvoiced decision errors. These errors not only degrade synthetic speech quality from prosodic feature aspect but also affect extraction of spectral envelope features, resulting in degradation also from segmental feature aspect.

The generation process model of F_0 contours (F_0 model) developed by Fujisaki and his co-workers can solve the problems of HMM-based speech synthesis [2]. The model represents a sentence F_0 contour in logarithmic scale as superposition of accent components on phrase components. These components are known to have clear correspondences with linguistic and para-/non- linguistic information, which is conveyed by prosody. Thus, using this model, a better control is possible in F_0 contour generation than the frame-by-frame control. Because of clear relationship between generated F_0 contours and linguistic and para-/non- linguistic information of utterances, manipulation of generated F_0 contours is possible, leading to a flexible control of prosody. We already have developed several methods, which use the F_0 model-generated F_0's in HMM-based speech synthesis, and showed their advantages in F_0 controls [3-7]. Recently, a method has been proposed, which uses F_0's approximated by the F_0 model instead of observed F_0's of training corpus for HMM training [8]. The method can partly solve the problem, which arises from F_0 extraction errors, but cannot avoid spectral envelope features affecting the synthetic speech quality.

In the current paper, the F_0 model is used in a different way to cope with the issue of erroneous F_0's; to exclude from HMM training samples which have large differences in F_0's from those approximated by the F_0 model. The performance of the method may rely on "how accurately the F_0 model parameters can be extracted from observed F_0 contours." For this purpose, the paper uses the method of automatic extraction of model parameters recently developed by the authors [8]. The method is motivated on how experts do when finding the F_0 model parameters, and utilizes linguistic information of utterances and our knowledge on Japanese prosody.

The rest of the paper is organized as follows: following to the explanation on the F_0 model, a new method of automatic extraction of F_0 model commands is briefly introduced in section 2. Section 3 shows the proposed method of datum selection for HMM training with the results of speech synthesis experiments. Some discussions on the method are given in Section 4. Section 5 concludes the paper.

2. Automatic extraction of F_0 model commands

2.1. F_0 model

Movements of F_0 along time axis are well represented by the F_0 model, which is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components [3]. The *i*-th phrase component $G_{pi}(t)$ of the F_0 model is generated by a second-order, critically-damped linear filter in response to an impulse-like phrase command, while the *j*-th accent component $G_{aj}(t)$ is generated by another second-order, critically-damped linear filter in response to a stepwise accent command:

$$G_{pi}(t) = \begin{cases} \alpha_i^{\,2} t\, e^{-\alpha_i t} & t \geq 0 \\ 0 & t < 0 \end{cases}, \quad (1)$$

$$G_{aj}(t) = \begin{cases} \min\left[1 - (1+\beta_j\, t)\, e^{-\beta_j t}, \gamma\right] & t \geq 0 \\ 0 & t < 0 \end{cases}. \quad (2)$$

Based on the analysis of Japanese utterances, time constants α_i and β_j are known to be fixed to values around 3.0 s^{-1} and 20.0 s^{-1}, respectively. The parameter γ, which thresholds accent components, can also be set to a fixed value around 0.9. An F_0 contour is then given by the following equation (assuming natural logarithm):

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{ G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j}) \} , \quad (3)$$

where, F_b is the bias level, I is the number of phrase components, J is number of accent components, A_{pi} is the magnitude of the i^{th} phrase command, A_{aj} is the amplitude of the j^{th} accent command, T_{0i} is the time of the i^{th} phrase command, T_{1j} is the onset time of the j^{th} accent command, and T_{2j} is the reset time of the j^{th} accent command.

2.2. Method

When searching F_0 model parameters, time constants α_i and β_j, threshold γ, and bias level F_b are usually fixed and are put out of the search process. Therefore, the search is done for F_0 model parameters related to phrase and accent commands, and is called F_0 model command extraction. Several methods have already been developed for automatically extracting F_0 model commands from given F_0 contours. Their basic idea is: smoothing to avoid micro-prosodic and erroneous F_0 movements, interpolating to obtain continuous F_0 contours, and taking derivatives of F_0 contours to extract accent command locations and amplitudes [9, 10]. Phrase commands are extracted from the residual F_0 contours (F_0 contour minus extracted accent components) [9], or from low frequency components of F_0 contours (assuming F_0 contours as waveforms) [10-12]. Extracted phrase and accent commands are optimized by a successive process. These methods, however, are not robust for pitch extraction errors, and produce commands not corresponding to the linguistic information of the utterances to be analyzed. Although attempts have been conducted to reduce extraction errors by introducing constraints (on command locations) induced from the linguistic information, their performances are still not satisfactory [13].

Interpolation of F_0 contours has a drawback since it relies on F_0's around voiced/unvoiced boundaries, where F_0 extraction is not always precise. This situation leads to extraction of false commands. Micro-prosodic F_0 movements at voiced consonants may also degrade the command extraction performance, since they are not counted in the F_0 model. To avoid false extractions, a new method is developed which accounts F_0 contours only of vowel segments [8]. In turn, since no F_0 is available between vowels, it comes difficult to extract accent commands from F_0 contour derivatives. Therefore, the new method takes features of Japanese prosody into account. In Japanese, F_0's of a syllable take either High or Low values corresponding to accent types. The method extracts phrase commands first viewing "Low" parts, and then find accent command amplitudes from "High" parts. The method can extract minor phrase commands, which are difficult to be found from the residual F_0 contours. We can say that the new method is motivated from the human process of command extraction.

The method consists of three steps; pre-processing, parameter extraction, and optimization. As for accent phrase boundaries, and accent types, which are necessary for the following processes, those used in HMM-based speech synthesis are used; the method is in good match with HMM-based speech synthesis. Phoneme boundaries are detected by the forced alignment using Julius as the recognizer [14]. A significant improvement in extraction performance as compared to conventional methods is observed. Since it is developed taking Japanese prosody into account, further investigations are necessary to make it applicable to other languages. The detail of the method is given in [8].

3. Speech synthesis with selected training data

3.1. Selection of training data

Figure 1 shows an example of F_0's extracted by STRAIGHT [15], and their F_0 model approximation. Here, their absolute difference in (natural) logarithmic values in each frame is denoted as F_0 difference:

$$F_{0diff}(t) = \left| \ln F_{0obs}(t) - \ln F_{0\,model}(t) \right| . \quad (4)$$

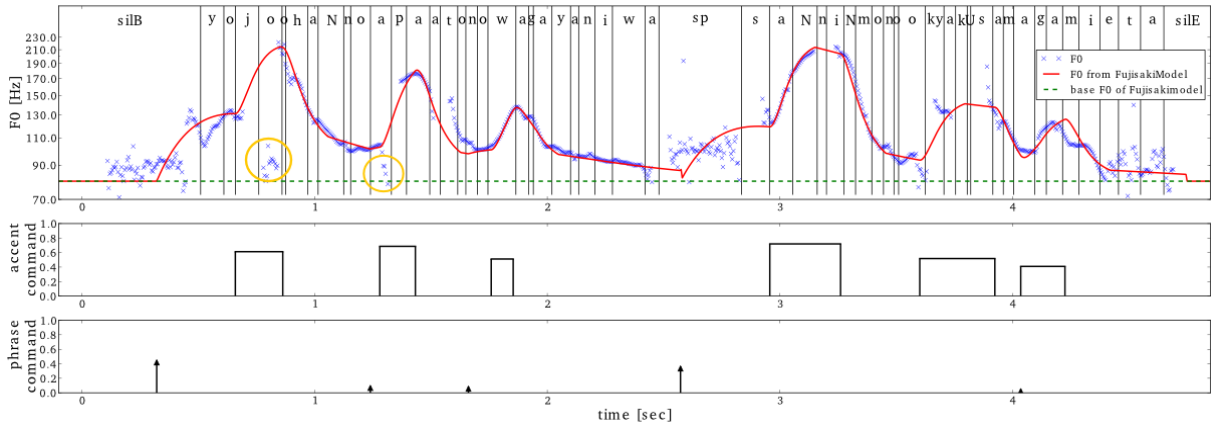

Figure 1: *An example for extracted F_0's (×) and their F_0 model approximation (in red solid line). F_0 model parameters (accent and phrase commands) are also shown. ("yojoohaNno apaatono wagayaniwa, saNniNmono okyakuga mieta": Three guests visited to my house, which was an apartment of only four and half "tatami" space.)*

$F_{0obs}(t)$ and $F_{0\,\mathrm{mod}\,el}(t)$ denote extracted F_0 and F_0 model F_0 at frame t, respectively. Half pitch extraction errors are observable around the second /o/. Also pitch extraction errors occur around the latter half of the second /a/. These segments may affect badly for the HMM training. It is possible to correct F_0's and to use them for HMM training, but the erroneous F_0's also cause errors in mel-cepstral coefficients (through STRAIGHT analysis). Here, speech segments with large F_0 differences are excluded from the data for HMM training. (Although F_0's are "wrongly" extracted at some speechless parts/pauses, they are ignored through F_0 model approximation and selection processes.) Two schemes are checked; one to exclude whole sentences which include many frames with large F_0 differences, and the other to exclude phoneme segments with large F_0 differences.

3.2. Experiments

Speech synthesis experiments are conducted using ATR continuous speech corpus of 503 sentences by male speaker MMI [16]. Out of 503 sentences, 450 sentences are used for HMM training, keeping 53 sentences for evaluation. Utterances for HMM training are analyzed using STRAIGHT with 5 msec frame shift. Fundamental frequencies are searched between 80 Hz to 250 Hz. Mel-cepstral coefficients and aperiodicity indices are calculated from analysis results by STRAIGHT using SPTK [17]. Feature parameters used for HMM training and speech synthesis processes have 138 dimensions: mel-frequency coefficients (0^{th}–39^{th}), aperiodicity indices (0-1 kHz, 1-2 kHz, 2-4 kHz, 4-6 kHz, 6-8 kHz), logarithmic F_0, and their Δ and Δ^2 features. Five-state left-to-right hidden semi-Markov model with single Gaussian distribution for each state, provided in HTS-2.1 [18], are used. A Gaussian distribution is represented by a diagonal conversion matrix. Decision tree-based context clustering is conducted with MDL stop criterion.

F_0 model parameters are extracted automatically from F_0 contours of training sentences using the method explained in section 2.2. Then, F_{0diff} is calculated for each voiced frame. When a frame has F_{0diff} larger than 1.0, and none of F_{0diff}'s of neighboring 10 frames (preceding 5 and following 5 frames) exceeds 1.0, F_0 model parameter extraction is re-conducted by neglecting the frame, and F_{0diff}'s are re-calculated.

As for sentence-based datum selection, utterances are excluded from HMM training, when they include "more than 2 frames with F_{0diff}'s larger than 1.0" or "more than 10 frames with F_{0diff}'s larger than 0.8." Forty three sentences are excluded from the 450 sentences by the process. These parameters for datum selection are determined through a preliminary experiment. Speech synthesis is conducted for two cases; one is when all 450 sentences are used for training (conventional method), and the other is when 43 sentences are excluded from the training (proposed method). Synthetic speeches obtained by these two methods are compared through a listening test with 3 scale scoring (1: better quality by the proposed method, 0: similar quality, -1: better quality by the conventional method). Two native speakers of Japanese conducted the listening test. Each speaker evaluated 10 sentences, which are selected randomly from 53 sentences for testing. The averaged result with 95 % confidence interval is 0.250 ± 1.337, indicating no significant difference between two methods.

Sentence-based datum selection may have a shortcoming that parts without F_0 extraction errors (and thus useful for HMM training) are thrown away together with other parts with erroneous F_0's. In order to cope with this situation, phoneme-based datum selection is conducted. Figure 2 shows the process. When second /a/, which is sand-witched by phonemes /r/ and /y/, includes a frame (frames) with large F_{0diff}, it is deleted from the training data. However, other phonemes in the same sentence without large F_{0diff}'s are left for HMM training. Three thresholds for "large F_{0diff}" are set so that 5 %, 10 % or 30 % of total voiced phoneme segments of the training corpus are excluded from the training. (No selection process is conducted for segments judged as voiceless.) Two versions of synthetic speeches, one by the conventional method and the other by the proposed method are compared through the listening test with 5 scale scores (+2: clearly better quality by the proposed method, +1: slightly better quality by the proposed method, 0: similar quality, -1: slightly better quality by the conventional method, -2: clearly better quality by the conventional method.). Nine native speakers evaluated the synthetic speeches by randomly selecting 20 sentences out of 53 sentences for testing in one turn. Every speaker conducted tree turns. The averaged results with 95 % confidence intervals are summarized in Table 1. Advantage of the proposed method over the conventional method is clear for the 5 % exclusion level.

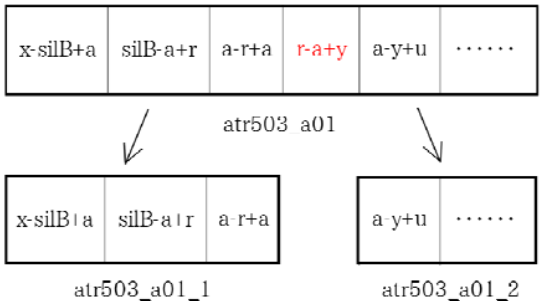

Figure 2: *Process of phoneme-based datum selection. Tri-phone HMM's, which are adopted for Japanese speech synthesis, are assumed. If tri-phone "r-a+y" has erroneous F_0's, it is excluded from the HMM training data.*

Table 1. *Scores of listening test with 95% confidence intervals for three levels of phoneme-based exclusion. Positive values indicate better quality by the proposed method.*

Level of exclusion	Score with 95 % confidence interval
5 %	0.300±0.148
10 %	0.139±0.147
30 %	-0.378±0.162

Figure 3 compares F_0 contours generated by the conventional method and the proposed method. Around sentence initial ("hyogeN"), a stable F_0 contour is obtained by the proposed method. F_0 contours around "nooryoku" and

"tsukeru" show better matches with accent types of the parts by the proposed method.

4. Discussion

Experiments are also conducted similarly to speech samples uttered by another speaker (female speaker FTY). Advantage of the proposed method, however, is not shown. The reason for the result will be due to the fact that F_{0diff}'s for FTY utterances are smaller than those for MMI. Selection of training data decreased the training datum size, and might affect negatively to the synthesized speech quality. Although, in the current experiments, the phoneme selection is conducted so that a fixed percentage of the training corpus remains, a scheme needs to be developed to use an absolute F_{0diff} threshold for selection. Also different thresholds can be used for different phonemes. A further study is necessary on the issue.

Reasons for large F_{0diff} can be divided into two cases; pitch extraction error and micro-prosody. Pitch extraction error can be further categorized several cases; double/half pitch errors which can be corrected through a post processing, pitch errors with large aperiodicity, and voiced/unvoiced decision errors. These errors should be treated differently. It should be noted that the method only counts voiced segments and no selection process works for voiceless frames. However, as is clear from Figure 1, F_0's are extracted for some unvoiced phonemes. When they have large F_{0diff}'s, they can be excluded from the HMM training. It is necessary to check how these exclusions affect the synthetic speech quality.

5. Conclusions

A method is developed to exclude speech segments from HMM training where their extracted F_0's are largely different from F_0's generated by the generation process model. Results on listening test for synthetic speech by the proposed method and the original HMM-based speech synthesis showed a clear improvement in synthetic speech quality when exclusion is done in phoneme-basis. Using F_0 values generated by the F_0 model without excluding segments with large F_{0diff} is another possible way to cope with erroneous F_0's [8]. A combined method is in our future research scope.

6. Acknowledgement

This work is partly supported by Grant-in-Aid for Scientific Research (B) #24300068, JSPS, and the Major Program for the National Social Science Fund of China (13&ZD189).

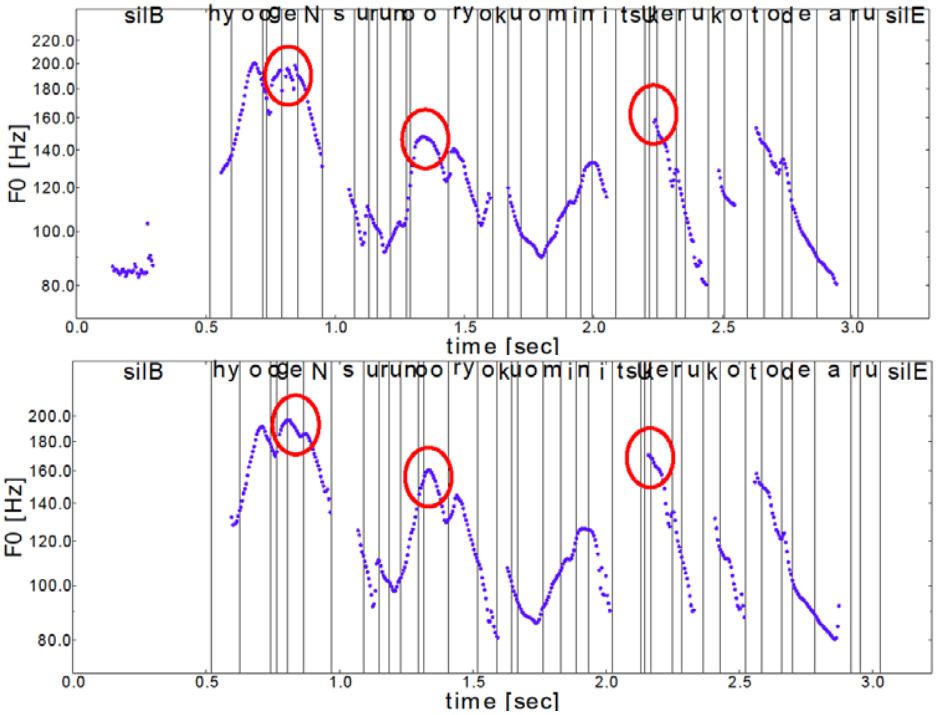

Figure 3: *F_0 contours obtained by the conventional method and the proposed method (phoneme-based). Improvements by the proposed method are observable at parts with red circles. ("hyogeNsuru nooryokuo minitsukeru kotodearu": It is to obtain a skill for presentation.)*

7. References

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. IEEE ICASSP*, pp.1315-1318, 2000.

[2] H. Fujisaki, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol.5, No.4, pp.233-242, 1984.

[3] K. Hirose, K. Sato, Y. Asano, and N. Minematsu, "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, Vol.46, Nos.3-4, pp.385-404, 2005.

[4] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model," *Proc. IEEE ICASSP*, pp.4485-4488, 2009.

[5] K. Hirose, K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu, "Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency," *Proc. INTERSPEECH*, pp.2793-2796, 2011.

[6] T. Matsuda, K. Hirose, and N. Minematsu, "Applying generation process model constraint to fundamental frequency contours generated by hidden-Markov-model-based speech synthesis," *Acoustical Science and Technology, Acoustical Society of Japan*, Vol.33, No.4, pp.221-228, 2012. .

[7] K. Hirose, H. Hashimoto, J. Ikeshima, and N. Minematsu, "Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model," *Proc. International Conf. on Speech Prosody*, pp.171-174, 2012.

[8] 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," *Proc. INTERSPEECH*, 4 pages, 2012.

[9] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujiaski, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. IEEE ICASSP*, pp.509-512, 2002.

[10] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proc. IEEE ICASSP*, vol.3, pp.1281-1284, 2000.

[11] V. Strom, "Detection of accents, phrase boundaries and sentence modality in German with prosodic features," *Proc. EUROSPEECH, V*ol. 3, pp. 2039.2041, 1995.

[12] A. Sakurai and K. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," *Proc. International Conference on Spoken Language Processing*, Vol.2, pp.817-820, 1996.

[13] K. Hirose, Y. Furuyama, and N. Minematsu, "Corpus-based extraction of F_0 contour generation process model parameters," *Proc. INTERSPEECH*, pp. 3257-3260, 2005.

[14] http://julius.sourceforge.jp/

[15] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F_0, and aperiodicity estimation," *Proc. IEEE ICASSP*, pp.3933-3936, 2008.

[16] A. Kurematsu, K.Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, Vol. 9, pp.357–363, 1990.

[17] http://sp-tk.sourceforge.net/

[18] http://hts.sp.nitech.ac.jp/