

# Improved Generation of Fundamental Frequency in HMM-Based Speech Synthesis Using Generation Process Model

Miaomiao Wang<sup>1</sup>, Miaomiao Wen<sup>1</sup>, Keikichi Hirose<sup>2</sup>, Nobuaki Minematsu<sup>2</sup>

<sup>1</sup> Department of Electrical Engineering and Information Systems, the University of Tokyo, Tokyo

<sup>2</sup> Department of Information and Communication Engineering, the University of Tokyo, Tokyo

{wangm, wenm, hirose, mine}@gavo.t.u-tokyo.ac.jp

## Abstract

The HMM-based Text-to-Speech System can produce high quality synthetic speech with flexible modeling of spectral and prosodic parameters. However the quality of synthetic speech degrades when feature vectors used in training are noisy. Among all noisy features, pitch tracking errors and corresponding flawed voiced/unvoiced (*VU*) decisions are the two key factors in voice quality problems. Pitch tracking errors occur more often in Mandarin vowels of Tone 3 and Tone 4. On the other hand, due to the dis-continuous  $F_0$  values in voiced and unvoiced regions, it is then impossible to use standard HMMs for  $F_0$  modeling. Currently a preferred method to solve this is to use a multi-space distribution HMM (MSD-HMM). In this approach, discrete distributions are used for modeling the *VU* decision and continuous Gaussian distributions are used for  $F_0$  modeling within the voiced regions. Due to this assumption of undefined  $F_0$  values in unvoiced regions and the special structure of MSDHMM, the generated  $F_0$  values are limited in accuracy. In this paper, an  $F_0$  generation process model is used to re-estimate  $F_0$  values in the regions of pitch tracking errors, as well as in unvoiced regions. A prior knowledge of *VU* is imposed in each Mandarin phoneme and they are used for *VU* decision. Then the  $F_0$  can be modeled within the standard HMM framework.

**Index Terms:** Mandarin speech synthesis,  $F_0$  generation, generation process model, HMM-based TTS

## 1. Introduction

Recently the HMM-based speech synthesis has been demonstrated to be very effective in synthesizing acceptable speech, in which short term spectra, fundamental frequency ( $F_0$ ) and duration are simultaneously modeled by the corresponding HMMs. It has compact and flexible representation of voice characteristics and has been successfully applied to Text-To-Speech system in many different languages, e.g., Japanese, English and Mandarin [1]. Compared with the unit selection based speech synthesis which based on large corpus, HMM-based synthesis is statistically oriented and model based. The speech generated by the HMMs is fairly smooth and exhibits no concatenation glitches occur in unit-selection synthesis. To change the segmental or supra-segmental quality of generated speech, we can modify HMM parameters flexibly [2, 3].

However, in HMM-based synthesis, the voice quality degrades when acoustic features used in training are noisy or flawed. Among them, pitch tracking errors and companion flawed voiced or unvoiced decisions are key causes of voice quality degradation. Different approaches have been proposed to improve the pitch tracking performance. Many HMM-based systems use STRAIGHT [4], a high quality speech analysis-synthesis system, to extract acoustic parameters for HMM training. In [5], a voting method, which combines the IFAS [6]

algorithm, a fixed-point analysis called TEMPO [7] and ESPS robust pitch tracking (RAPT) algorithm [8], is used to alleviate  $F_0$  extraction errors such as  $F_0$  halving and doubling, and voiced/unvoiced swapping. But still as we look into pitch tracking of Mandarin syllables, the tracking errors occur more often in vowels of Tone 3 and Tone 4. Because the pitch of these syllables can be very low and somewhat are not strong in periodicity. Thus the synthesized vowels sound very dry and hoarse, which greatly hurt the overall quality of synthesized speech.

Furthermore, even though the pitch tracking errors are manually checked before training which requires lots of work and time, there still can be *VU* decision errors in synthesized speech. In HMM-based synthesis, the modeling of  $F_0$  is difficult due to the discontinuity of  $F_0$  across voiced and unvoiced region. The multi-space distribution HMM (MSD-HMM) provides a solution to this problem by using a combination of discrete and continuous distributions [9] and it is now the default modeling approach in state-of-the-art HMM synthesis systems. However, although good performance can be achieved using MSDHMMs, this type of mixed distribution  $F_0$  modeling has some issues arising from the discontinuities at the boundaries of unvoiced regions and the need to keep the discrete and continuous density regions distinct. Therefore, the use of MSDHMMs makes it more difficult to exploit standard techniques for HMM modeling, such as adaptation, which cannot be readily applied to the mixed discrete or continuous  $F_0$  distributions.

From this consideration, we have developed a corpus-based method of synthesizing  $F_0$  contours in the framework of the generation process model, which represents continues sentence  $F_0$  contours as a superposition of tone components on phrase components [10]. The generation model is based on the physiological and physical properties of the vocal fold and the laryngeal structure involving laryngeal muscles. By applying this model,  $F_0$  contours can be smoothed and re-estimated from the extracted parameters, and also give us a possible way for interpolation of  $F_0$  in the unvoiced region [11]. Usually initials can be divided as voiced or unvoiced consonant, and all medials and finals are voiced in Mandarin. We can use the phoneme information for *VU* decision in synthesizing  $F_0$  contours.

## 2. A Model for the Generation Process of $F_0$ contours of Mandarin utterances

The generation process model is a command-response model that describes  $F_0$  contours in the logarithmic scale as the super-position of phrase components, accent components (or tone components for tonal languages, like Mandarin) and a baseline level  $F_b$ . The exact relationships between these components of an  $F_0$  contour and the underlying linguistic information have been formulated by Fujisaki and his coworkers [10]. The model diagram for Mandarin is shown in

Figure 1, where the phrase commands (impulses) produce phrase components through the phrase control mechanism, giving the global shape of the  $F_0$  contour at sentence level, while the tone commands generate tone components through the tone control mechanism, characterizing the local  $F_0$  changes. Both mechanisms are assumed to be critically-damped second-order linear systems.

In this model, the  $F_0$  contour is expressed by

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^I A p_i G_p(t - T_{0i}) + \sum_{j=1}^J A a_j \{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (1)$$

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0, \\ 0, & \text{for } t < 0 \end{cases} \quad (2)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t) \exp(-\beta t), \gamma], & \text{for } t \geq 0 \\ 0, & \text{for } t < 0 \end{cases} \quad (3)$$

where  $G_p(t)$  represents the impulse response function of the phrase control mechanism and  $G_a(t)$  represents the step response function of the tone control mechanism.

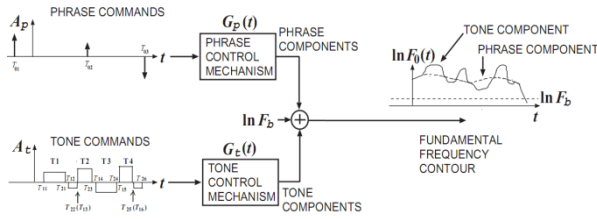


Figure 1: A Functional model for the process of generating  $F_0$  contours.

The model consists of the following parameters:  $A p_i$  and  $T_{0i}$  denote the magnitude and time of the  $i$ th phrase command respectively, while  $A a_j$ ,  $T_{1j}$  and  $T_{2j}$  denote the amplitude, onset time and offset time of the  $j$ th tone command respectively. The constants  $\alpha$ ,  $\beta$  and  $\gamma$  are set at their respective default values 3.0 (1/s), 20.0 (1/s) and 0.9 respectively in the current study.

Unlike most non-tone languages, e.g. English and Japanese, Mandarin requires both positive and negative tone commands. In Mandarin there are four lexical tones and a neutral tone. These tones are attached to each syllable. As shown in Figure 1, T1 to T4 are assumed to correspond to their respective tone command patterns (intrinsic patterns): T1 (positive), T2 (negative followed by positive), T3 (negative) and T4 (positive followed by negative). Figure 2 shows an example of  $F_0$  contours of a Mandarin utterance that are generated by extracted tone and phrase parameters.

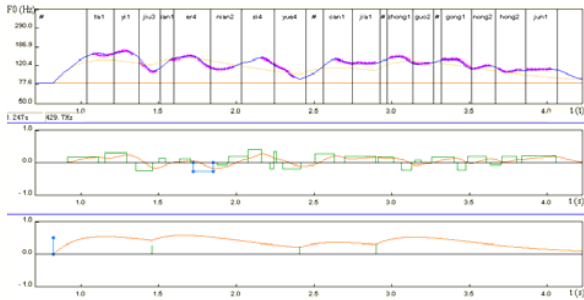


Figure 2: An example of  $F_0$  contour of Chinese utterance "tal yi1 jiu3 san1 er4 nian2 si4 yue4 chan1 jial zhong1 guo2 gong1 nong2 hong2 jun1 (He joined the Chinese Workers' and Peasants' Red Army in April 1932)."

A prior knowledge of  $VU$  or  $UV$  switch in Mandarin is that, each syllable has the phonemic structure of a single vowel or a consonant followed by a vowel. So there will be no more than one  $VU$  or  $UV$  switch during one syllable period.

### 3. Pitch Tracking Method and MSD-HMM for $F_0$ Modelling and Generation

In recent HMM-based synthesis, which needs a large corpus for training, an automatic pitch tracking method is needed. And a common assumption is that  $F_0$  has a continuous value in voiced regions and no value in unvoiced regions.

Firstly, ESPS RAPT algorithm is successful in automatic pitch tracking, and can alleviate  $F_0$  extraction errors such as  $F_0$  halving and doubling, and voiced/unvoiced swapping. But still as we look into pitch tracking of Mandarin syllables, the tracking errors occur more often in vowels of T3 and T4, for their pitch contours change greatly. Sometimes the T3 creates creaky sounds at the turning point of the pitch contour. The speaker may reach lower end of her/his modal register during such a tone, and to even more lower  $F_0$ , has to go into laryngealization. Thus it makes conventional pitch tracking methods fail to find  $F_0$  values in these regions and may lead to the  $VU$  decision errors at this point.

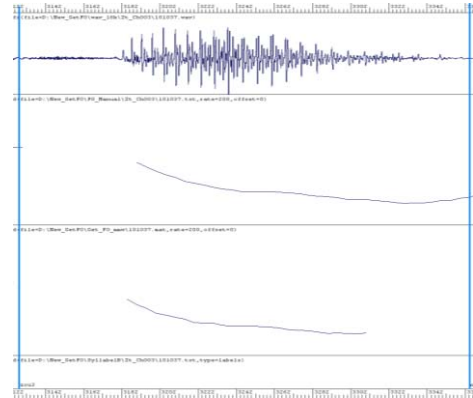


Figure 3: An example of  $F_0$  contours of Mandarin syllable "zou3". From top to bottom: original wave,  $F_0$  by manually check,  $F_0$  calculated by RAPT algorithm.

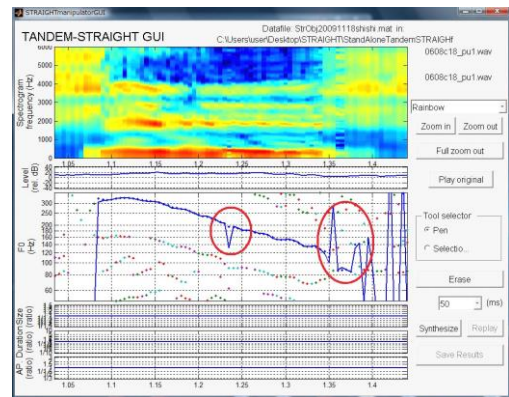


Figure 4: A typical  $F_0$  halving and doubling error (in circle) of Mandarin syllable "shi4" calculated by STRAIGHT algorithm.

Figure 3 shows comparison of target  $F_0$  and  $F_0$  extracted by ESPS RAPT algorithm. At the end of diphthong "ou" in T3, pitch detection algorithm fails to find  $F_0$  in voiced region. And Figure 4 shows a typical  $F_0$  halving and doubling errors in the Mandarin vowel "i" in T4 calculated by STRAIGHT algorithm. Thus these fails in  $F_0$  tracking of phonemes will

lead to a shorter duration of the vowel and sometimes noisy sound inside a vowel when re-synthesis. And more unvoiced utterances will occur in the synthesized speech from a HMM-based TTS which leads to unnatural sound.

Furthermore, in HMM-based speech synthesis system, the Voiced/Unvoiced ( $VU$ ) decision of each state is independently made based on the multi-space distribution of  $F_0$  parameters of that state. The MSD of  $F_0$  parameters of one state is estimated by traversing the decision tree by the contextual features till a leaf node. Due to some pitch tracking errors or some bad pronounced vowels, one leaf of the state belong to a vowel may contain more unvoiced occurrences than voiced occurrences. Thus, if choosing that leaf, the state will be decided as an unvoiced. Then the voice quality degrades not only because of the error pitch tracking, but also of the error  $VU$  decisions in HMM training.

In order to simultaneously model the discrete  $VU$  decision and the continuous  $F_0$  trajectory variables, multi-space distribution HMMs (MSDHMM) are commonly used [9]. The state output distribution in an MSDHMM is

$$b_{\theta}(o) = \begin{cases} c_v \mathcal{N}(o; \mu_{\theta}, \sigma_{\theta}) & o \in \text{voiced region,} \\ c_{uv}, & o \in \text{unvoiced region} \end{cases} \quad (4)$$

$$c_v + c_{uv} = 1 \quad (5)$$

where  $o$  is the observation at state  $\theta$ ,  $c_v$  and  $c_{uv}$  are the probabilities of voiced and unvoiced regions,  $\mu_{\theta}$  and  $\sigma_{\theta}$  are the means and variances of Gaussian distribution of  $F_0$  in the voiced regions. This MSDHMM framework results in some inherent limitations. Since  $b_{\theta}(o)$  represents a continuous density in voiced regions and a discrete probability mass in unvoiced regions, each observation can only be either voiced or unvoiced, but not both at the same time. Consequently, during the forward-backward calculation for any  $F_0$  stream in training, the state posterior occupancy will always be wholly assigned to one of the two components depending on the voicing condition of the observation. This hard assignment limits the ability of the unvoiced component to learn from voiced data and vice versa, and it prevents any possibility of using a soft assignment to reduce the effect of  $F_0$  estimation errors.

It's also hard for this state-of-art HMM-based TTS to handle prosodic features especially at the phrase or sentence level. In this method, both segmental and prosodic features of speech are processed together in a frame-by-frame manner. Prosodic features cover a wider time span than segmental features, and should be treated differently.

#### 4. $F_0$ Modelling in HMM-based TTS using Generation Process Model

The previous sections highlighted the Generation Process Model which can generate continuous  $F_0$  contours; the problems encountered in HMM-based TTS were successfully solved. In the model that we proposed in this section, we used Generation Process Model to generate continuous  $F_0$  contours and assumed to exist in unvoiced regions, together with the  $VU$  decision of phoneme information.

Here we defined Mandarin phonemes with either voiced or unvoiced as show in Table 1. In some respects, the phonemic structure of Mandarin is quite simple. It's either a consonant-vowel ( $CV$ ) structure or single vowel ( $V$ ) structure. Mandarin contains 21 consonants, 5 semi-vowels, 4 diphthong vowels, and 14 mono-phthong vowels. We can define them either voiced or unvoiced depending on the pervious knowledge of their waveforms.

Table 1. Mandarin Initial and Tonal Final units with Voiced/Unvoiced decision

Unvoiced Initials	b, c, ch, d, f, g, h, j, k, p, q, s, sh, t, x, z, zh
Voiced Initials	l, m, n, r, u, y
Voiced Tonal Finals	a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, ii, iii, in, ing, iong, o, ong, ou, u, ua, uai, uan, uang, uei, uen, uo, v, van, ve, vn

After labeling each phoneme with  $VU$  information, together with the  $F_0$  values estimated from an ESPS waves-based  $F_0$  contours, Fujisaki parameters are extracted by a FujiPara Editor [12]. Then a continuous  $F_0$  contour can be re-estimated using Fujisaki parameters [11]. Together with extracted spectral parameters, continuous  $F_0$  contours will be applied for the HMMs training. Here  $F_0$  is modeled in one single stream, and each state is modeled with a single Gaussian diagonal covariance output. In the synthesis stage,  $F_0$  trajectory is generated for a given state sequence in the maximum likelihood sense. And the  $VU$  decision will be made based on the phonemic information and white noise will be used as unvoiced excitation source to synthesize the unvoiced frames.

By making the continuous  $F_0$  using the generation process model, the problems in section 3 are effectively addressed. Since the miscalculated  $F_0$ , either error  $VU$  decisions, or doubling and halving, can be fixed before training. Also there is only one single  $F_0$  stream, so there are no redundant component weights parameters.

### 5. Experiment Results and Discussion

To evaluate the performance of our proposed method compared to the MSD-HMM, a manually checked female speaker's corpus is used for both methods. Prof. Renhua Wang, from the University of Science and Technology of China provided us the Mandarin speech corpus which consists of 270 training and 30 testing sentences. The labels of unvoiced initials are used as the boundaries of  $VU$  switch. The input text to the system includes symbols on pronunciation and prosodic boundaries, which can be obtained from orthogonal text using a natural language processing system, developed at University of Science and Technology of China [12].

As for the HMM-based method, the HMM-based Speech Synthesis toolkit (HTS Ver.2.1) [13] is used. Five-state, left-to-right HMM phone models are adopted. The MSD-HMM generates  $F_0$  together with 24-order mel-cepstrum coefficients.

The ESPS RAPT algorithm is used for automatic  $F_0$  extraction. Before training, we found that almost 22.37% syllables of the total have the  $VU$  decision errors. And among these errors, 33% failures are occurred in T4, 39% in T3, 11% in T0, 12% in T2 and 5% in T1. After training process of MSDHMM, the errors will increase.

Figure 6 shows an example of  $F_0$  extracted by ESPS RAPT algorithm and the continuous  $F_0$  contour re-estimated by the tone and phrase components using generation process model. Here we can find that the ESPS RAPT algorithm is failed to find  $F_0$  values in the vowel "u" in T3. In contrast our proposed method is successful to smooth the  $F_0$  contours, find  $F_0$  values in the  $VU$  decision error regions, and interpolate  $F_0$ s in the unvoiced regions. Then all the syllables with  $VU$  decision errors are fixed before training.

Figure 7 shows examples of  $F_0$  contours generated by MSD-HMM and our approach, compared and overlaid with

corresponding original nature speech. In this example, there are 5 Mandarin syllables: “zhe3+ti2+chu1+le0+yi4”. Here the syllable “zhe3” in Tone 3 is difficult to synthesize for its pitch contours change greatly and sometimes sounds creaky. The syllable “le0” in neutral tone is also hard to synthesize for its command pattern depends on the context and usually have reduced amplitudes. As shown in Fig. 7, the MSD synthesizer has  $VU$  decision errors in “zhe3” and “le0” syllables and low accuracy of  $F_0$  contours while our method outperforms in both two syllables. And there are few  $VU$  decision errors in our approach.

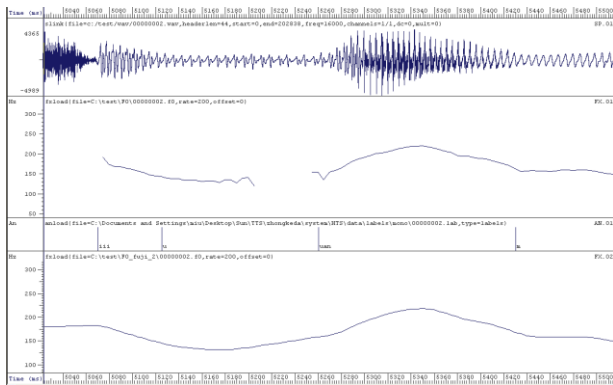


Figure 6: An example of the continuous  $F_0$  contours for the Mandarin syllable “shi2+wu3+wan4+mu3”. From top to bottom: original wave,  $F_0$  calculated by RAPT algorithm, phoneme labels,  $F_0$  re-estimated by Generation Process Model

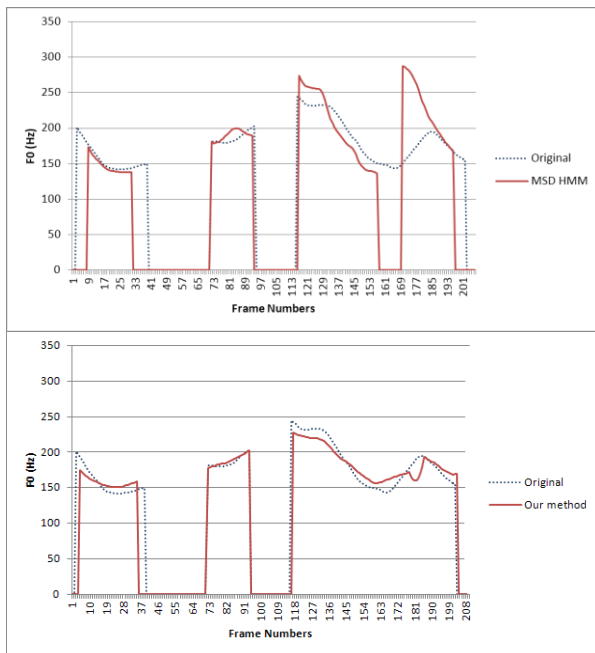


Figure 7: An example of the  $F_0$  contours predicted by MSD and our method, along with corresponding original nature speech contours

Another advantage of continuous  $F_0$  contours generated at synthesis stage is that we can re-estimate the tone and phrase components from them. In the conventional HMM-based speech synthesis system, the synthesized speech sounds are evidently muffled compared with natural speech because the generated speech-parameter trajectories are often over-smoothed, especially the  $F_0$  parameter trajectory. The detailed characteristics of speech parameters are removed in the modeling part and cannot be recovered in the synthesis part. Although using the advanced acoustic models may reduce

over-smoothing, this may still exist because the synthesis algorithm does not explicitly include a recovery mechanism [14]. But by using the generation process model, we can re-estimate the tone components and especially the phrase components, which cover a wider range and should be treated differently in this frame-by-frame works of HMMs. Thus the generation process model gives us a possible way of recovery mechanism, e.g. tone emphasis, refinement of phrase component and duration, at the synthesis part.

## 6. Conclusions

In this paper, we proposed a method to generate continuous  $F_0$  contours for HMM-based speech synthesis system by applying the generation process model. It can fix the  $F_0$  tracking errors and  $VU$  errors before training, and assume that  $F_0$  values are exist in unvoiced regions so there is only one single stream of  $F_0$  in HMM. Then there are no redundant component weights parameters. A prior linguistic knowledge of phonemes of Mandarin is used for the  $VU$  decision at the synthesis stage. And a clear relationship is obtainable between generated  $F_0$  contours and their background linguistic information, enabling flexible control of prosodic features in HMM-based TTS.

## 7. References

- [1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kita-mura, “Speech Parameter Generation Algorithms for HMM-based Speech Synthesis,” in Proc. ICASSP, 2000.
- [2] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training”, IEICE Trans. Inf. & Syst., vol. E90-D, no. 2, pp. 533–543, Feb. 2007.
- [3] T. Nose, J. Yamagishi, and T. Kobayashi, “A style control technique for HMM-based expressive speech synthesis”, IEICE Trans. Inf. & Syst., vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.
- [4] H. Kawahara, I. M. Katsuse, and A. D. Cheveigne, “Restructuring speech representations using a pitch adaptive time frequency smoothing and an instantaneous frequency-based  $F_0$  extraction: possible role of a repetitive structure in sounds”, Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.
- [5] J. Yamagishi, Z. Ling, and S. King, “Robustness of HMM-based Speech Synthesis”, Proc. of Interspeech, 2008.
- [6] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, “Robust  $F_0$  estimation of speech signal using harmonicity measure based on instantaneous frequency”, IEICE Trans. Inf. & Syst., vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.
- [7] H. Kawahara, H. Katayose, A. Cheveign’e, and R. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of  $F_0$  and periodicity”, Proc. of EuroSpeech, 1999.
- [8] D. Talkin, “A robust algorithm for pitch tracking (RAPT)”, in Speech Coding and Synthesis, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, pp. 495–518.
- [9] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, “Multispace probability distribution HMM,” IEICE Trans. Inf. & Syst., vol. E85-D, no. 3, pp. 455 – 464, 2002.
- [10] 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)
- [11] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, “Evaluation of an improved method for automatic extraction of model parameters from fundamental frequency contours of speech,” Proc. Int. Conf. Speech Prosody, pp.443-446 (2004-3)
- [12] Mixdorff, H., Hu, Y. and Chen, G. (2003): Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin. In Proceedings of Eurospeech 2003, Geneva.
- [13] “HMM-based Speech Synthesis System (HTS),” <http://hts.sp.nitech.ac.jp> 2009
- [14] H. Zen, K. Tokuda, and A. W. Black : “Statistical parametric speech synthesis,” Speech Communication, Vol. 51, No. 11, pp. 1039-1154 (2009)