Vol.2011-SLP-87 No.1
2011/7/21

Prosody Improvement for HMM-based Mandarin Speech Synthesis Using the Tone Nucleus Model

MIAOMIAO WANG[†] MIAOMIAO WEN[†]
KEIKICHI HIROSE[††] NOBUAKI MINEMATSU[††]

The HMM-based Text-to-Speech System has attracted great interest due to its compact and flexible modeling of spectral, F_0 and duration parameters. The synthesized speech is highly dependent on the context model. However, the complex F_0 variations make it rather difficult to define the tone type of Mandarin continuous speech. Then the F_0 and duration trajectories, generated by HMM-based speech synthesis are often excessively smoothed and lack of prosodic variance. Tone nucleus of a syllable is assumed to be the target F_0 of the associated lexical tone, and usually conforms more likely to the standard tone pattern. In this paper, by modeling F_0 variations at different levels ranging from segmental factors to tone co-articulations, and apply the tone nucleus model to HMM-based Mandarin speech synthesis.

1. Introduction *

Recently in speech synthesis community, attention has been attracted by HMM-based speech synthesis, 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]. 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 the generated speech, we can modify HMM parameters flexibly [2, 3]. In this system the trained statistical context dependent HMMs are clustered using a tree-based context clustering technique, and then used to predict duration and generate parameters like mel-cepstral coefficients, F_0 values, and bandpass voicing strengths using the maximum likelihood parameter generation algorithm including global variance (GV).

Mandarin, the standard Chinese, is a well-known tonal language in which pitch tones play an important phonemic role in Mandarin: each syllable corresponds to a morpheme and is associated with a lexical tone. Syllables with different lexical tones may have different meanings even if they own the same segmental structure. When researchers tried to apply the present speech synthesis systems, originally designated for read speech, to spontaneous speech, they met many new problems such as ungrammatical utterances, different reading style, ambiguous syntax, etc. All the researches depend on correct understanding and modeling of prosodic features including F_0 contours, duration and intensity. In the case of Mandarin, due to F_0 variations for lexical tones, F_0 contours show larger undulations than those in the non-tonal languages, like English and Japanese. The lexical tones show consistent tonal F_0 patterns when uttered in isolation, but show complex variations in continuous speech [4, 5]. The invariance problem is the difficulty of giving a physical phonetic definition of a given linguistic category that is constant and always free of context [6].

Thus, in Mandarin continuous speech the complex F_0 variations and intonation make it rather difficult to define the tone type and then there will be some mismatch between the context labels and real F_0 patterns. Then the synthesized speech is usually over-smoothed and lack of prosodic variance.

Tone nucleus model suggest that a syllable F_0 contours can be divided into three segments: onset course, tone nucleus and offset course. The tone nucleus of a syllable is assumed to be the target F_0 of the associated lexical tone, and usually conforms more likely to the standard tone pattern than the articulatory transitions. This model has improved the tone recognition rate in [7] to show that tone nucleus keeps the important discriminant information between tonal F_0 patterns and underlying tone type and successfully improved the tone recognition rate. Those findings lead us to the idea that we can apply the tone nucleus model to improve the F_0 modeling and generation in HMM-based speech synthesis system. So in this paper, firstly we extracted the tone nucleus part of each syllable and adjust the context label according to the F_0 patterns. Then we used only the tone nucleus part, instead of the whole syllable F_0 contours which consist a lot of redundancy and cause data sparseness, to train the context dependent HMMs. Finally, the tone nucleus part is generated with corresponding HMMs and the whole F_0 contours are recovered by SPLINE interpolation and voiced or unvoiced decision will be included in the

* [†] Department of Electrical Engineering and Information Systems, the University of Tokyo
[††] Department of Information and Communication Engineering, the University of Tokyo

duration model.

The rest of this paper is organized as follows. The second section will introduce Mandarin tones and tone nucleus model. In the third section, we will review the conventional F_0 modeling and generation method in HMM-based speech synthesis system and how to adapt the tone nucleus model to it. In section 4, experiments and results are described and discussed. Finally we will give the conclusion in the last section.

2. Mandarin Tones and Tone Nucleus Model

2.1 Basic lexical tones

In Mandarin each syllable corresponds to an ideographic character and is associated with a pitch tone referred to as lexical tone). Phonemically, a syllable is divided into two parts: an Initial and a Final. An Initial can be a consonant or none. A Final may be a vowel, a diphthong, or a triphthong and with an optional nasal ending. There are four basic lexical tones (referred to as Tones 1, 2, 3, 4, respectively) and a neutral tone. The four basic lexical tones are characterized by their perceptually distinctive pitch patterns which are conventionally called by linguists as: high-level (Tone 1), high-rising (Tone 2), low-dipping (Tone 3) and high-falling (Tone 4) tones [8]. The neutral tone, according to [8] does not have any specific pitch pattern, and is highly dependent on the preceding tone and usually perceived to be temporally short and zero pitch range. F_0 contours are the main acoustic manifestations of pitch tones, and there seem to be distinctive F_0 patterns associated with the four basic lexical tones as shown in Figure 1.

2.2 Tone nucleus

For a syllable F_0 contour, as pointed out in [7], lexical tone is not evenly distributed in a syllable as F_0 variations in a syllable F_0 contour in various phonetic contexts. Only its later portion, approximately corresponding to the final vocalic part, is regarded to bear tonal information, whereas the early portion is regarded as physiological transition period from the previous tone. It was also found that there are often cases where voicing period in the ending portion of a syllable also forms a transition period of vocal vibration and contributes nothing to the tonality. From these considerations, we can classify a syllable F_0 contour into underlying target and articulatory transitions:

- Underlying target represents the F_0 target and serves as the main acoustic cue for pitch perception.
- Articulatory transitions are the F_0 variations occurring as the transitions to or from the pitch targets.

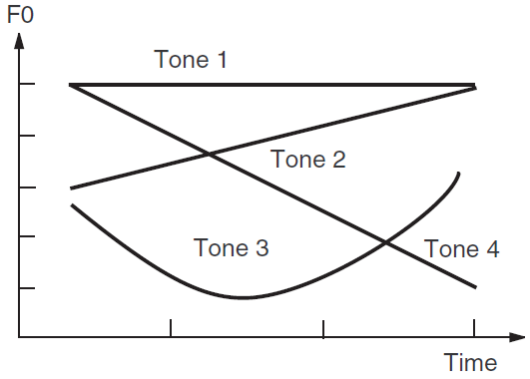


Figure 1. Standard distinctive F_0 patterns of the four basic lexical tones

Figure 2 illustrates some typically observed tonal F_0 variations in continuous speech and their tone nuclei notations. The three segments are called onset course, tone nucleus, and offset course, respectively, which are defined as follows:

- Tone nucleus: a portion of F_0 contour that represents pitch targets of the lexical tone. It is the segment containing the most critical information for tonality perception, thus called as the tone-critical segment..
- Onset course: the asymptotic F_0 transition locus to the tone-onset target from a preceding vocal cords' vibration state.
- Offset course: the F_0 transition locus from the tone-offset target to a succeeding vocal cords' vibration state.

Table 1. Pitch target features of the four lexical tones

Targets	Tone 1	Tone 2	Tone 3	Tone 4
Onset	H	L	L	H
Offset	H	H	L	L

As compared with Figure 1, the tone nucleus part will conform more likely to the standard tone pattern. Tone-onset target and tone-offset target indicate the pitch values, which takes either low (L) or high (H) value, at the tone onset and offset, respectively. These pitch values serve as distinctive features characterizing the four basic lexical tones showed in table 1.

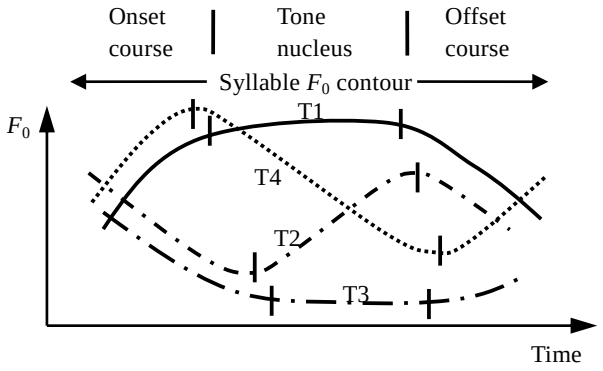


Figure 2: Tonal F_0 contours with possible articulatory transitions and their tone nucleus.

2.3 Tone nucleus extraction

To apply the tone nucleus model for speech synthesis, it is necessary to automatically estimate tone nucleus parameters from F_0 contour. For each syllable F_0, we use a robust tone nucleus segmentation and location method based on statistical means. The method has two steps: the first step is F_0 contour segmentation based on the iterative segmental K means segmentation procedure, with which a T-Test based decision of segment amalgamation is combined in [7]. When segmentation becomes available, which segment is tone nucleus is decided according to the following rules in the second step: (1) For Tone 1, the segment with the biggest average F_0. (2) For Tone 2, the segment with the largest average ΔF_0. (3) For Tone 4, the segment with the lowest average ΔF_0. Among the four basic lexical tones, Tone 3 is different from the other three tones, in that the others are associated with rather stable F_0 patterns, whereas Tone 3 is found with rather wide variety of F_0 patterns. So (4) for Tone 3, only the segment with the lowest average F_0 will be considered. Considering that the syllable's maximum and minimum F_0 points carry important information in expressing emotions, if the chosen segment fails to cover the maximum or the minimum F_0 point, it will be expanded to include these two critical points. Since neutral tones shows no inherent F_0 contour, a stable definition of tone nucleus is difficult, and hence we assume the entire voiced segment of the syllable as tone nucleus for neutral tone.

3. Conventional F_0 modeling and generation in HMM-based speech synthesis and our approach

3.1 Conventional method

In HMM-based speech synthesis, Multi-Space Distribution (MSD) HMM was proposed [9] to model stochastically the piece-wise continuous F_0 trajectory. For the F_0 values in unvoiced and voiced regions, MSD-HMM models two different probability sub spaces: discrete for the unvoiced regions and continuous for the voiced F_0 contours. During synthesis, sequences of F_0 are generated directly from the HMMs under the maximum likelihood criterion considering not only the sequences but also dynamic features and the GV of the sequences. The output probability of the GV is modeled by a single Gaussian distribution. In this method, the observation sequence O includes static and dynamic features of F_0 sequences f, and is defined as

$$O = [o_1^\top, o_2^\top, \dots, o_n^\top, \dots, o_n^\top] \tag{1}$$

where

$$o_n = [f_n, \Delta f_n, \Delta^2 f_n] \tag{2}$$

$$F = [f_1, f_2, \dots, f_n, \dots, f_N]^\top \tag{3}$$

Δf_n and $\Delta^2 f_n$ are dynamic features of F_0. The above equations can also be represents by the determinant of a matrix as

$$O = WF \tag{4}$$

For a given continuous mixture HMM λ, it maximizes $\log P(O|\lambda)$ with respect to (4). W is static, delta and delta-delta coefficient matrix. In the synthesis part, if the state sequence Q is given by state duration, we set

$$\frac{\partial logP(WF|Q,\lambda)}{\partial F} = 0 \tag{5}$$

and obtain

$$W^{\top}U^{-1}WF = W^{\top}U^{-1}M \tag{6}$$

where U and M are covariance matrix and mean vector of F_0. The GV of F_0 sequences is defined as

$$v(\boldsymbol{f}) = \frac{1}{N}\sum_{n=1}^{N}[f_n - \bar{f}(\boldsymbol{f})] \tag{7}$$

where

$$\bar{f}(\boldsymbol{f}) = \frac{1}{N}\sum_{n=1}^{N} f_n \tag{8}$$

An output probability of the GV $P(v(f)|\lambda_v)$ is modeled by a single Gaussian distribution. The Gaussian model λ_v and HMMs λ are independently trained from a speech corpus.

3.2 Our approach using tone nucleus model

Firstly, after F_0 extraction, we will find the tone nucleus part of each syllable using the algorithm discussed in section 2. Then we will compare the template of tone nucleus with corresponding context label. If they do not match, we will modify the context label according to the templates. After that instead of using the whole syllable length F_0 contours for HMM training, we use only tone nucleus part to build the context dependent HMMs. Thus in MSD-HMM models two different probability sub spaces will be continuous for the tone nucleus part and discrete for other regions. Here we assumed that the unvoiced regions are also regarded as physiological transition period as other voiced articulatory transition. In the F_0 generation stage, other than maximum the likelihood for the whole syllable length, only the tone nucleus will be predicted. Then we will smooth and make interpolation for the whole F_0 contours. The voiced or unvoiced decision will be included in the duration model. In some respects, the phonemic structure of Mandarin is quite simple. It's either a consonant-vowel structure or single vowel structure. So there will be no more than on V/U switch within phone duration. We can define them based on the pervious knowledge of their waveforms and airflow source. Then we can use the phoneme segmental boundaries as V/U switch point in the synthesis stage. Figure 3 shows an example sentence with extracted tone nucleus (between dots). Figure 4(a) shows an example of F_0 templates for Tone 2 clustered by using tone nucleus. And for comparison, Figure 4(b) shows the F_0 templates for Tone 2 clustered using whole syllable F_0 contours. It is clear that F_0 templates of the whole syllable segment are scattered and thus hard to predict. The extracted tone nucleus templates could better capture the tone pattern shape (a rising shape for Tone2) and are easier to predict.

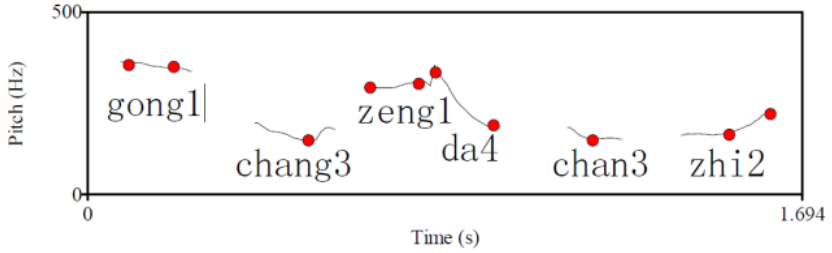
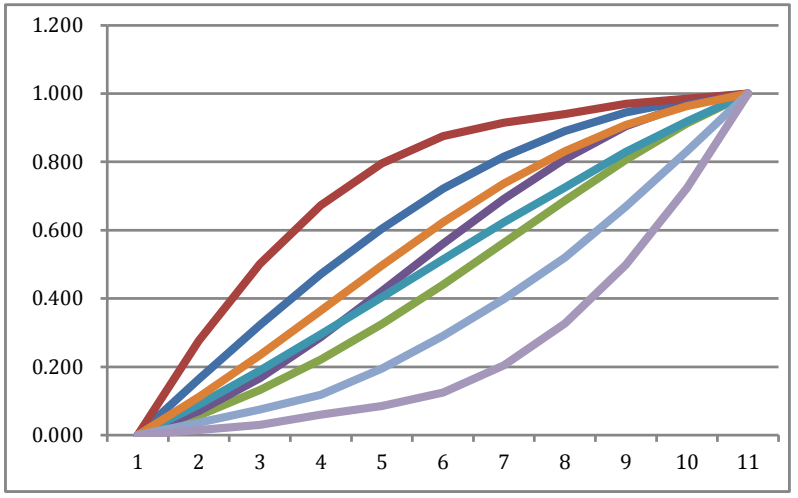
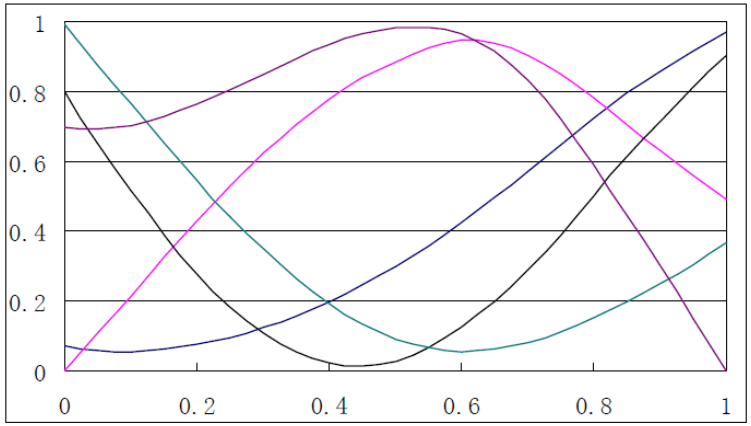


Figure 3 An example sentence: the segment between the dots indicates the tone nucleus part of the syllable



(a)



(b)

Figure 4 (a) F0 templates for Tone2 using tone nucleus. (b) F0 templates for Tone 2 clustered using syllable length F0 contours.

4. Experiment and Results

To evaluate the performance of our proposed method, a female speaker's corpus is used. The Mandarin speech corpus consists of 270 training utterances includes 15392 phones and pauses, and 30 testing utterances. The ESPS RAPT [10] algorithm is used for automatic F_0 extraction. HMM-based Speech Synthesis toolkit (HTS Ver.2.1) [11] 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.

After extracting the tone nucleus of each tonal vowel, we will modify the context label according to the F0 template. That is, for example, if a phone is detected with a Tone 4 shape, but it is labeled as Tone 2. We will modify its label as Tone 6. Here Tone 6 means that there are mismatch between tone nucleus pattern and its label. So we treat it as neutral tones.

Before training, we found that almost 8.22% tonal vowels have different F_0 patterns comparing with their context labels. 2.6% failures are occurred in Tone 0, 65.8% in Tone 1, 26.4% in Tone 2, 5.2% in Tone 4. Tone 3 is a special case because it only has the lowest average F_0. So we didn't consider the case of Tone 3. As shown in Fig. 5, Tone 1 has the most mismatching cases because the intonation of Tone 1 is more complex than other tones.

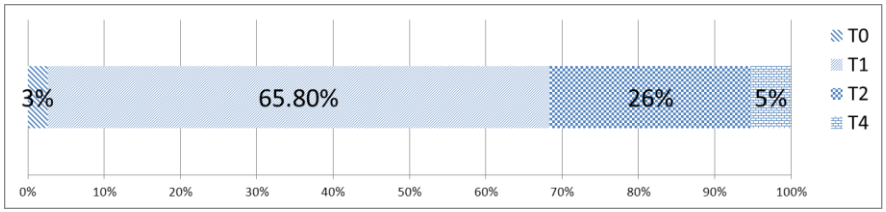


Figure 5 mismatching rate for each tone

But extracting the tone nucleus and modifying the context label, all the mismatching errors are fixed before training. Then we only use the tone nucleus part of the F_0 contours for HMM training. After training, the tone nucleus part is predicted by HMMs and we will use SPLINE interpolation in the voiced regions. Here each phone boundary will be treated as the voiced/unvoiced switching point. Here we divided each phone as voiced or unvoiced in table2.

Table 2. Pitch target features of the four lexical tones

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

The subjective evaluation as shown in Fig.6 is an AB preference choice test on speech sentences synthesized by two different methods. There are two sets of subjective comparisons conducted: one is using duration predicted by HMM, the other is using original duration as the original voiced or unvoiced decision is applied. It shows that our framework outperforms the baseline system.

Figure 7 shows an example of synthesized tone nucleus and the nature F_0 contour. It should be noted our methods provide clearly better results than the baseline system.

Our method works even when a small sized speech corpus is obtainable.

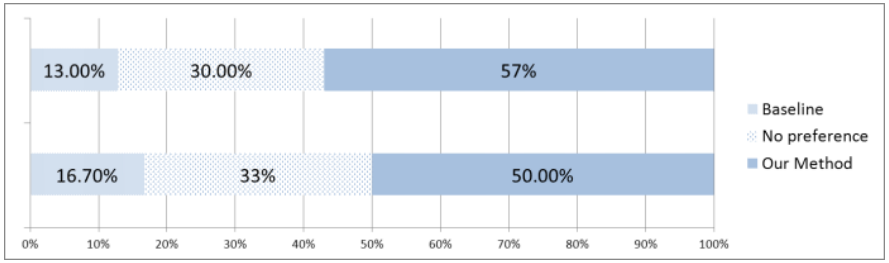


Figure 6: AB preference test. From up to down: duration predicted by HMM and the original duration

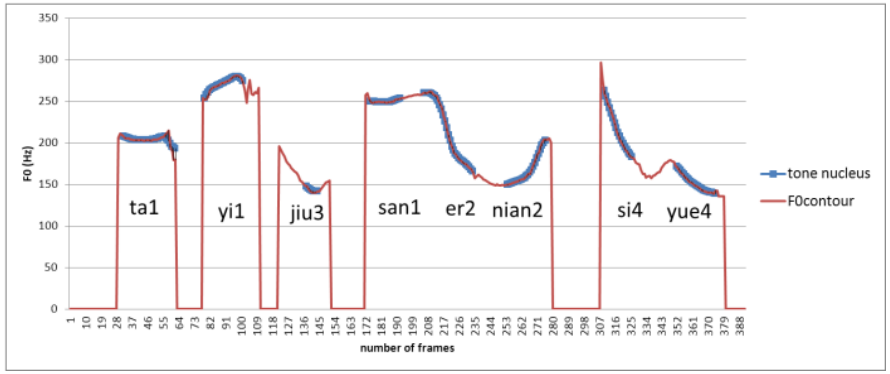


Figure 7: An example of synthesized tone nucleus and nature F0 contour

5. Conclusions

In this paper, we proposed a new method to improve the prosody of the HMM-based speech synthesis using tone nucleus model. As due to the complex variation of F_0 contours in Mandarin, the tone nucleus conforms more likely to the standard tone pattern. It also provides an underlying linguistically and physiological description for surface F_0 contour and often can furnish several compact parameters to represent a long pitch contour. Thus this method enables us to make flexible control of prosodic features in HMM-based TTS. In the future works, we would like to move on the perception study on Mandarin whispered speech as we found that even there is no F_0 values in whispered speech, people still can recognize the tone type especially a meaningful sentence. We would believe that there are other perceptual cues for tone recognition and this would help us to understand better about Mandarin tones, and further improve the prosody of synthesized speech.

6. Acknowledgements

The authors' sincere thanks are due to Prof. Renhua Wang and Prof. Lirong Dai in the University of Science and Technology of China for providing the Mandarin speech database.

Reference

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, pp.1315-1318, 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 tech-nique for HMM-based expressive speech synthesis”, IEICE Trans. Inf. & Syst., vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007
4) S.-H. Chen and Y.-R. Wang, "Tone recognition of continuous Mandarin speech based on Neural Networks", IEEE Trans. on SAP, Vol. 3, No. 2, 1995, pp.146-150.
5) Y. Xu, , Contextual tonal variations in Mandarin. J. Phonetics 25, 61–83, 1997.
6) B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory", W.Hardcastle and A. Marchal (ed.), Speech Production and Speech Modelling. Kluwer Academic Publishers, 1990, pp.403-439.
7) J. Zhang, and K. Hirose, "Tone nucleus modeling for Chinese lexical tone recognition," Speech Communication, Vol. 42, Nos. 3-4, pp. 447-466, 2004.
8) Y.-R. Chao, 1968. A Grammar of Spoken Chinese. University of California Press, Berkeley.
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) 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.
11) "HMM-based Speech Synthesis System (HTS)," http://hts.sp.nitech.ac.jp. 2009