Generation of Fundamental Frequency Contours of Mandarin in HMM-based Speech Synthesis using Generation Process Model

Miaomiao Wang 1, Keikichi Hirose 2, Nobuaki Minematsu 2

1 Department of Electrical Engineering and Information Systems, the University of Tokyo, Tokyo
2 Department of Information and Communication Engineering, the University of Tokyo, Tokyo

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

Abstract

The HMM-based speech synthesis system can produce high quality synthetic speech with flexible modeling of spectral and prosodic parameters. In this approach, short term spectra, fundamental frequency ($F_0$) and duration are generated by multi-stream HMMs separately. 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, because the pitch of these vowels can be very low and sometimes treated as aperiodic signal. On the other hand, $F_0$ values in unvoiced regions, such as consonants, are normally defined as unavailable; 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 estimate $F_0$ values in the region of pitch tracking errors, as well as in unvoiced regions. A prior knowledge of $VU$ is imposed in each Mandarin phoneme and then used for $VU$ decision. Thus the $F_0$ can be modeled within the standard HMM framework.

Index Terms: Mandarin speech synthesis, Generation process model, $F_0$ contour, HMM-based 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]. Compared with the large corpus, example the unit selection based speech synthesis, 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 the generated speech, we can modify HMM parameters flexibly [2, 3].

However, in HMM-based synthesis, 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], 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 those 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, 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]. By handing $F_0$ contours in the $F_0$ model framework, a clear relationship is obtainable between generated $F_0$ contours and their background linguistic (and para-/non-linguistic) information, enabling “flexible” control of prosodic features. And in Mandarin, there is a clear set of constraints on the phonetic structure of each syllable. Initials may be consonants or vowels, medials are vowels, and finals are vowels or nasals. 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.

The rest of the paper is organized as follows. In section 2, the generation process model of $F_0$ contours for Mandarin utterances is introduced. In section 3, after a brief discussion of $F_0$ extraction errors in Mandarin syllable of Tone 3 and Tone 4, the conventional $F_0$ modelling and generation in HMM-based synthesis system is reviewed. In section 4, we present our method of $F_0$ modelling in HMM-based synthesis using generation process model. In section 5, experiment result is described and in section 6, we give our conclusion.
2. A Model for 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) and a baseline level $F_0$. 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 (pedestals) 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

$$\log_e F_0(t) = \log_e F_0 + \sum_{i=1}^{1} Ap_i Gp(t - T_{0i})$$

$$+ \sum_{j=1}^{1} Aa_j (Ga(t - T_{1j}) - Ga(t - T_{2j}))$$ (1)

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

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

where $Gp(t)$ represents the impulse response function of the phrase control mechanism and $Ga(t)$ represents the stop response function of the tone control mechanism.

The model consists of the following parameters: $Ap_i$ and $T_{0i}$ denote the magnitude and time of the $i$th phrase command respectively, while $Aa_j$, $T_{1j}$ 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: T1 (high tone), T2 (rising tone), T3 (low tone), T4 (falling tone) and T0 (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). For T2 and T4, the offset of the 1st tone command is assumed to coincide with the onset of the 2nd tone command. The command pattern for T0 is assumed to depend on the context and usually have reduced amplitudes. Figure 2 shows an example of $F_0$ contours of a Mandarin utterance that are generated by extracted tone and phrase parameters. By handing tone and phrase commands in the generation process model frame-work, a clear relationship is obtainable between generated $F_0$ contours and their background linguistic (and para-non-linguistic) information, enabling flexible control of prosodic features.

A prior knowledge of VU or UV switch in Mandarin is that, each syllable or we could call it each Chinese character 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. Conventional Pitch Tracking Method and $F_0$ Modelling in HMM-based TTS System

In recent HMM-based synthesis, which need 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. Because the pitch of those syllables can be very low and somewhat are not strong in periodicity.

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

Figure 2: An example of $F_0$ contour of Chinese utterance "ta1 yi1 jiu3 san1 er4 nian2 si4 Yue4 chan1 jia1 zhong1 guo2 gong1 nong2 hong2 jun1 (He joined the Chinese Workers’ and Peasants’ Red Army in April 1932)." From top to bottom: observed $F_0$ contour with its generation process model approximation, tone components/commands, and phrase components commands.

Figure 3: An example of $F_0$ contours of Mandarin syllable "su4": From top to bottom: original wave, $F_0$ by manually check, $F_0$ calculated by RAPT algorithm.
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, handling $F_0$ contours with their background linguistic knowledge and the problems encountered in HMM-based TTS when $F_0$ values were mis-calculated in voiced regions, discrete probability mass for unvoiced regions. 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.

In order to investigate the validity of our proposed method of continuous $F_0$ contours generation when it is applied in a HMM-based TTS system, a full speech synthesis algorithm was constructed as show in Fig. 5.
Then a continuous $F_0$ contour can be generated using Generation Process Model. We could select the continuous $F_0$ contour as the $F_0$ observation for the unvoiced frames.

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

<table>
<thead>
<tr>
<th>Voiced/Unvoiced</th>
<th>Unvoiced Initials</th>
<th>Voiced Initials</th>
<th>Voiced Tonal Finals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b, c, ch, d, f, g, h, j, k, p, q, s, sh, t, x, zh</td>
<td>ga, ge, go, l, m, n, r, w, y</td>
<td>a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ib, ian, iang,iao, ie, if, in, ing, iong, iu, o, ong, ou, u, ua, uai, uan, uang, ui, un, uo, v, van, ve, vn</td>
</tr>
</tbody>
</table>

Together with extracted spectral parameters, the continuous $F_0$ contours will be applied to a HMM-based TTS. In the synthesis stage, 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 mis-calculated $F_0$ can be fixed before training, and also there is only one single $F_0$ stream, there are no redundant component weights parameters.

5. Experiment

To evaluate the performance of our proposed method compare to the MSDHMM, 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. 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 an NLP 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) is used [13]. The MSDHMM 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 all most 22.37% syllables of the total have the error $VU$ decisions. And among all those errors, 33% failures are occurred in T4 and 39% are in T3. After training process of MSDHMM, this error will be increased.

Figure 6: Error $VU$ decisions for Mandarin syllables in different tones

We use the FujiiParaEditor to find continuous $F_0$ contours for the corpus. Figure 7 shows an example of our method compared to the conventional pitch tracking method. We can see that during a voiced vowel “i” in T4, the conventional method failed to find $F_0$ values in the voiced regions.

Figure 7: An example of the continuous $F_0$ contours for the Mandarin word “l+i4 sh+i2”. The spot line is by ESPS algorithm and the continuous line is by our method.

6. Conclusion

In this paper, we proposed a method to generate continuous $F_0$ contours for HMM-based speech synthesis by applying the generation process model. It can fix the $VU$ errors of $F_0$ 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. The $VU$ errors are fixed before HMM training. And compared to MSDHMM, there will be no more unvoiced frames during the voiced regions.

7. References

http://hts.sp.nitech.ac.jp.