A Minimum V/U Error Approach to F0 Generation in HMM-based TTS

Yao Qian, Frank Soong, Miaomiao Wang, Zhizheng Wu

1Microsoft Research Asia, Beijing, China
2Department of Information and Communication Engineering, University of Tokyo
3College of Software, Nankai University, China

Abstract

The HMM-based TTS can produce a highly intelligible and decent quality voice. However, HMM model degrades when feature vectors used in training are noisy. Among all noisy features, pitch tracking errors and corresponding flawed voiced/unvoiced (v/u) decisions are identified as two key factors in voice quality problems. In this paper, we propose a minimum v/u error approach to F0 generation. A prior knowledge of v/u is imposed in each Mandarin phone and accumulated v/u posterior probabilities are used to search for the optimal v/u switching point in each VU or UV segment in generation. Objectively the new approach is shown to improve v/u prediction performance, specifically on voiced to unvoiced swapping errors. They are reduced from 3.7% (baseline) down to 2.0% (new approach). The improvement is also subjectively confirmed by an AB preference test score, 72% (new approach) versus 22% (baseline).

Index Terms: speech synthesis, HMM-based TTS, v/u decision

1. Introduction

HMM-based speech synthesis has been successfully applied to TTS synthesis in many different languages, e.g., Japanese, English and Mandarin [1-3]. In this framework, spectral envelope, fundamental frequency, and duration are modeled simultaneously by the corresponding HMMs. For a given text, speech parameter trajectories are generated by the trained HMMs in the Maximum Likelihood (ML) sense. Compared with the large corpus, example unit selection based unit 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 [4,5].

However, HMM-based TTS voice quality degrades when acoustic features used in training are noisy or flawed. Among them, F0 tracking errors and companion flawed v/u decisions are key causes of voice quality degradation. Different approaches have been proposed to improve the pitch tracking performance. Many HMM-based TTS systems use STRAIGHT [6], a high quality speech analysis-synthesis system, to extract acoustic parameters for HMM training. In [7], a voting method, which combines the IFAS [8] algorithm, a fixed-point analysis called TEMPO [9] and ESPS robust pitch tracking (RAPT) algorithm [10], to alleviate F0 extraction errors such as F0 halving and doubling, and voiced/unvoiced swapping.

In this paper, we propose a new minimum v/u error approach to F0 trajectory synthesis for HMM-based TTS. The new approach is for producing more consistent and better v/u prediction in synthesis than the conventional baseline system. A prior knowledge of v/u label for each Mandarin phone is incorporated into v/u prediction and accumulated v/u probabilities are used to search for the optimal v/u switching point. Comparing with the baseline system, the new approach can significantly reduce v/u prediction errors in F0 generation and produce more pleasant synthesized voice.

The rest of the paper is organized as follows. In section 2, the conventional F0 modeling and generation in HMM-based TTS system is reviewed. In section 3, we present our new minimum v/u error approach. In section 4, experiments and evaluation results are described. In section 5, we give our conclusions.

2. F0 Modeling and Generation in HMM-based TTS System

2.1. MSD-HMM for F0 modeling

In HMM-based speech synthesis, Multi-Space Distribution (MSD) HMM was proposed [11] to model stochastically the piece-wise continuous F0 trajectory. It assumes that the observation space \( \Omega \) of an event is made up of \( G \) sub-spaces. Each sub-space \( \Omega_g \) has its own probability

\[
    w_g = p(\Omega_g) \quad \text{and they are summed up to } 1, \quad \sum_{g=1}^{G} w_g = 1.
\]

An observation vector, \( o \), in each sub-space is distributed according to an underlying pdf, \( p_g(o) \). The dimension of the observation vector can be different in each sub-space. The observation probability of \( o \) is

\[
    b(o) = \sum_{g \in s(o)} w_g p_g(o) \tag{1}
\]

where \( s(o) \) is the index set of the sub-spaces that observation \( o \) belongs to and it is determined by the extracted features at each observation time. For voiced frames with F0 values and unvoiced frames, MSD-HMM models two, discrete and continuous, probability spaces: discrete for the unvoiced regions and continuous for the voiced F0 contours. The output pdf of the zero-dimensional, unvoiced sub-space is a Kronecker delta function and continuous F0 is a one-dimensional, voiced sub-space with a Gaussian mixture distribution. The probability of subspace \( g \) and state \( j \) is estimated by
where \( \gamma_j(j,g) \) is the posterior probability of an observation in state \( j \) and subspace \( g \) at time \( t \), which is efficiently estimated by the Forward-Backward algorithm.

In MSD-HMM training, all contexts including rich and long ones are used to capture F0 co-articulation effects. However, in practice, limited by insufficient training data, we usually have to tie models of rich contexts into generalized ones for predicting unseen contexts in testing. State tying via a clustered decision tree is commonly used. F0 is modeled in a stream separated from the spectral feature stream and a stream-dependent F0 decision tree is built.

2.2. V/U prediction and F0 generation

In parameter trajectory generation, contextual MSD-HMM parameters are retrieved by traversing the trained decision trees and the duration of each state is obtained from corresponding duration model. Voiced or unvoiced decision of state \( j \) is determined by the voiced subspace probability \( W_{j,g} \). A maximum likelihood F0 trajectory is generated with dynamic feature constraints [1]. For a given HMM model \( \lambda \), it determines a sequence of F0 values, \( F_0 = f_0(1),...,f_0(T) \), which maximizes \( \log P(O|\lambda) \) with respect to \( O = WF \). \( W \) is the transformation matrix in computing static, delta, and delta-delta features. The state labeled, voiced frame sequence \( Q \) is determined by the corresponding duration statistics and we set

$$\frac{\partial \log P(WF|Q,\lambda)}{\partial F} = 0$$  \hspace{1cm} (3)$$

for the F0 extraction in training and erroneous v/u predictions due to the state-dependent memoryless F0 generation.

$$W^T U^{-1}WF = W^T U^{-1}M$$  \hspace{1cm} (4)$$

where \( U \) and \( M \) are the covariance matrix and mean vector of F0.

2.3. Limitations of the conventional approach

The HMM-based TTS can produce a highly intelligible and decent voice quality. However, sometimes the synthesized speech exhibits perceptually annoying glitches due to voiced/unvoiced swapping errors in F0 trajectory generation. After analyzing 50 synthesized Mandarin testing sentences, we find that nearly 8% of voiced phones have these erroneous v/u switching errors. Generally, v/u switching should only appear at the boundary states of a voiced phone. Five examples of erroneous voiced phones in their tri-phone forms are shown in Figure 1, where the probabilities of voiced subspaces in successive state sequences are plotted. These erroneous v/u predictions mainly occur in nasals, phones preceding or succeeding to silence, and phones in 3rd tone Mandarin syllables. We think these v/u decision errors are caused by errors in both training and synthesis, i.e., flawed pitch estimates in the F0 extraction in training and erroneous v/u predictions due to the state-dependent memoryless F0 generation.

Figure 1: Five examples of voiced phones with erroneous v/u predictions.

3. Minimum V/U Error to F0 Generation

Although the MSD-HMM provides an elegant stochastic framework for F0 modeling and generation, it is still not perfect in practice. One of the shortcomings of MSD is the assumption that all observations of a state depend only upon that state but not upon its neighboring states. This memoryless state dependency creates fluctuating v/u spurious decisions from one state to the next, hence perceptually annoying v/u swapping noise in synthesized voice. We propose a minimum v/u error approach to F0 generation for producing a more continuous, less “jumpy” v/u decisions. In F0 generation, voiced/unvoiced decision of state \( j \) is made by accumulated v/u errors \( e_j \), instead of voiced subspace probability \( W_{j,g} \) in Eq (2).

We labeled each Mandarin phone as either voiced or unvoiced, based on their manner of vocal fold vibration. Accordingly, each state has a voiced (v) or unvoiced (u) label to inherit from its parent phone. Two kinds of state sequence are defined for any two successive segments: an UV sequence, which has only one unvoiced to voiced state and succeeding v states, and a VU sequence, similar to UV sequence but V sequence precedes U sequence. An example is illustrated in Figure 2.

We define accumulated v/u errors \( e_j^n, j = 1,...,N \) and \( e_j^u, j = 1,...,M \) for UV and VU state sequences, respectively.

$$e_j^n = V_j^n + U_j^n$$  \hspace{1cm} (5)$$

$$V_j^n = V_{j+1}^n + \gamma(j,g = v), \quad V_0^n = 0, \quad j = 1,...,N$$

$$U_j^u = U_{j+1}^u + \gamma(j,g = u), \quad U_N^u = 0, \quad j = N,...1$$

$$e_j^u = V_j^u + U_j^u$$  \hspace{1cm} (6)$$

$$V_j^u = V_{j+1}^u + \gamma(j,g = v), \quad V_M^u = 0, \quad j = M,...1$$

$$U_j^u = U_{j+1}^u + \gamma(j,g = u), \quad U_0^u = 0, \quad j = 1,...,M$$
where \( \gamma(j, g=v) \) and \( \gamma(j, g=u) \) are the accumulated posterior probabilities summing over all frames in state \( j \) and in a voiced subspace \( (g=v) \) or an unvoiced subspace \( (g=u) \), i.e., \( \gamma(j, g) = \sum_{t} \gamma_{t}(j, g) \). Only one v/u switching point is allowed and it is set at the minimum \( e^{vu}_{i} \) and \( e^{uv}_{i} \) for each UV or VU state sequence. Accordingly, for an UV state sequence, \( i = \text{min}(e^{uv}_{i}) \), i.e., all states precede \( i \) are all unvoiced, those succeed \( i \) are voiced, and the v/u ratio for state \( i \) given state duration is according to the probability in state \( i \) and subspace \( g \), \( w_{i,g} \), which is calculated in Eq (2). The v/u decision for UV state sequence is similar to UV state sequence but searching for the optimal voiced to unvoiced switching point instead. The successive minimum accumulated v/u error operations are illustrated well in two adjacent UV and VU state sequences in Figure 2.

4. Experiments and Results

4.1. Experimental Setup

A phonetically and prosodically rich, broadcast news style corpus in Mandarin, recorded by a female professional radio announcer, is used in our experiments. The corpus consists of 1,000 training and 50 testing sentences. Speech signals are sampled at 16 kHz, windowed by a 25-ms window with a 5-ms shift, and the 40th order LPC coefficients are transformed into static LSPs and their dynamic counterparts. Five-state, left-to-right HMM phone models, where each state is modeled with a single Gaussian, diagonal covariance output distribution, are adopted.

The phonetic and prosodic contexts are used as question set in growing decision tree. They include tones and breaks, quin-phone context, positions of phone, syllable and word in phrase and sentence, and the length of word and phrase. The questions for splitting the nodes of tree are automatically selected in the ML sense. Minimum description length (MDL) criterion [12] for balancing model complexity and training data size, is used as a stopping criterion for state clustering in decision tree growing.

4.2. Evaluation Results and Analysis

Objective and subjective measures are used to evaluate the performance of the proposed approach on the testing set of 50 sentences.

4.2.1. Objective measure

Synthesis quality is measured objectively in terms of distortions between natural and synthesized speech frame synchronously where oracles state durations (obtained by forced alignment) of natural speech are used. The objective measure is voiced/unvoiced (v/u) swapping errors between the natural and synthesized F0 trajectories. To avoid pitch tracking errors in the automatic pitch extraction algorithm, the natural F0 trajectories are manually checked by visualizing spectrogram and glottal wave signals. Table 1 lists the v/u swapping error percentage by comparing the natural and synthesized F0 trajectories. It shows our minimum v/u errors approach can significantly reduce v/u errors, especially the voiced to unvoiced swappings.

Table 1: The v/u swapping errors between the natural and synthesized F0 trajectories.

<table>
<thead>
<tr>
<th></th>
<th>Swapping Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>v -&gt; u</td>
<td>3.7%</td>
</tr>
<tr>
<td>u -&gt; v</td>
<td>4.1%</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
</tr>
<tr>
<td>Minimum v/u errors</td>
<td>2.0%</td>
</tr>
<tr>
<td></td>
<td>3.9%</td>
</tr>
</tbody>
</table>

4.2.2. Subjective measure

The subjective evaluation of the approach is an AB preference test on speech sentence pairs synthesized with the same segmental but two F0 trajectories synthesized by the baseline and our new minimum v/u error approach. Fifty Mandarin sentences are evaluated. Two professional
subjects participated in the preference listening test where pair preference choices are: 1) the former is better; 2) the latter is better; 3) no preference (The difference between the paired sentences cannot be perceived or the difference can be perceived but it is difficult to choose which one is better). The preference scores of the baseline and our approach are summarized in Figure 3. It shows that our minimum v/u error approach outperforms the baseline system significantly, 72% vs 22% while 6% of sentences are selected as no preference.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>No preference</th>
<th>Minimum v/u errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22%</td>
<td>6%</td>
<td>72%</td>
</tr>
</tbody>
</table>

Figure 3: The preference scores of the baseline system and our minimum v/u error approach.

4.2.3. Results analysis

Figure 4 shows two examples of F0 trajectories predicted by the baseline and our minimum v/u error approach, along with that of corresponding natural speech. In the two Mandarin syllables “dian3” and “shuo1”, our approach predicts the v/u switching point more accurately than the baseline system.

Figure 4: Two examples of F0 trajectories predicted by baseline and our minimum v/u error approach, along with corresponding natural speech trajectories.

We also investigate whether our approach can remove sporadic v/u prediction errors within an all voiced phone, like the examples shown in Figure 1. The statistical results show that the percentage of intra-phone v/u switching is reduced from 8% to 4%. Most of the remaining errors are in phones with left or right silence context. Due to the statistical nature of HMM, these voiced phone models may absorb silence frames into their bordering and internal states.

5. Conclusions

We propose a new minimum error approach to F0 generation in HMM-based speech synthesis. Mandarin phones are labeled as either voiced or unvoiced and it is used as given prior in synthesis. The posterior probabilities of voiced and unvoiced states are accumulated and then used to search for the optimal v/u switching points within each UV or VU speech segment in generation. The experimental results show our approach significantly outperforms the conventional baseline system in both objective evaluation and subjective AB preference test.

6. References