A Perceptual Study of Acceleration Parameters in HMM-based TTS

Yi-Ning Chen*  Zhi-Jie Yan  Frank K. Soong

Microsoft Research Asia, Beijing, China
{ynchen,zhijiey,frankkps}@microsoft.com

Abstract

In HMM-based TTS, statistical models of static, velocity (delta), and acceleration (delta-delta) parameters are jointly trained in a unified, ML-based framework. Previous study has shown that the acceleration parameters are able to generate smoother trajectory with less distortions, but the effect has never been investigated for formal objective and subjective tests. In this paper, the effect of the acceleration parameters, in addition to their static and velocity counterparts, in trajectory generation is studied in depth. We show that discarding acceleration parameters only introduces small additional distortion compared to the reference generated with full model parameters. But human subjects can easily perceive the voice quality degradation, because saw-tooth-like trajectories are commonly generated. Several methods to alleviate the discontinuity are discussed, and we choose the upper- and lower-bounded envelopes of the saw-tooth trajectories for further analysis. Experimental results show that both envelope trajectories have larger objective distortions than the saw-tooth ones. However, the speech synthesized using the envelope trajectory becomes perceptually transparent to the reference. This study, in addition to its subjective and objective significance in measuring the distortion of the synthesized speech, facilitates efficient implementation of low-cost TTS systems, as well as low bit rate speech coding and reconstruction.

Index Terms: speech synthesis, HMM, acceleration parameters

1. Introduction

Hidden Markov Model (HMM) based approach has become one of the most important methods for speech synthesis in the last decades. In HMM-based TTS, a universal Maximum Likelihood (ML) criterion is used for both training and synthesis. The ML criterion is capable of obtaining a reliable estimate of both static and dynamic parameters during the training phase. And most importantly, it imposes a static-dynamic constraint in the synthesis phase, which helps to generate smooth and highly intelligible parametric speech trajectories [1–3].

Dynamic properties of speech trajectory are conventionally characterized by velocity (delta) and acceleration (delta-delta) parameters of the HMMs. During synthesis, those parameters are considered along with the static parameters, and a weighted normal equation is solved to generate trajectories maximizing the overall likelihood. Without dynamic parameters, the generated speech would only reflect the static mean parameters. The ML solution of the trajectory then becomes a piecewise constant function, which leads to perceptually noticeable discontinuity of the synthesized speech [2]. However, after dynamic parameters are imposed, not only the delta and delta-delta mean parameters, but also all the covariance parameters are taken into account. The output speech trajectory becomes significantly smoother, and is also favored in subjective listening tests.

It was reported in [2] that the difference between the synthesized speech with/without delta-delta (acceleration) parameters is relatively small. However, no quantitative objective test nor formal subjective test were performed for further analysis. This initiates this study to investigate the perceptual effect of the acceleration parameters in HMM-based TTS by theoretical analysis, as well as both objective and subjective tests.

In this paper, we use the trajectory generated using full model parameters as the reference. We show that discarding acceleration parameters only introduces small objective distortion to the reference. However, human subjects can easily perceive the quality degradation, because saw-tooth-like trajectories are commonly generated without acceleration parameters. Similar observation was also reported recently in a different context using trajectory HMMs and various delta windows [4]. We presume that the unsmooth trajectory is the main reason for the voice quality degradation. Therefore, the upper- and lower-bounded envelopes of the saw-tooth trajectory are manually generated for further analysis. Experimental results suggest that although the objective distortions of the envelope trajectories are both larger than that of the saw-tooth trajectories, the human subjects can no longer perceive any quality degradation with the speech synthesized using the envelope trajectories. This study inspires us on how to generate perceptually transparent speech trajectory in the absence of some model parameters (specifically acceleration parameters in this case): 1) the distortion introduced compared with the reference should be relatively small, and 2) the generated speech should not contain other artifacts like the unsmooth, saw-tooth trajectory. This observation is not only helpful in building low-cost TTS systems, but also enlightening for other research topics like speech coding and reconstruction.

The rest of this paper is organized as follows: In Section 2, we briefly review the conventional trajectory generation algorithm, and show the voice quality degradation if acceleration parameters are discarded; In Section 3, we investigate the reason for the quality degradation. The saw-tooth trajectory, as well as its upper- and lower-bounded envelope trajectories are compared with the reference in both objective and subjective experiments; Finally in Section 4, we draw our conclusions.

2. Trajectory Generation With / Without Acceleration Parameters

2.1. Conventional trajectory generation algorithm

This study was initiated in an effort to develop low-cost HMM-based TTS algorithms when computation and storage resources are limited (e.g., TTS running on Windows Mobile platform).

*This author is now with Alibaba Cloud Computing, Beijing.
One of the most straightforward methods to achieve such a goal is to discard the acceleration parameters of the HMMs while maintaining the same level of voice quality. In conventional HMM-based TTS, given the state sequence \( q = [q_1, q_2, \ldots, q_T] \), parametric speech trajectory is generated so as to maximize the overall likelihood \( \log P(W|q) \), which leads to the well-known weighted normal equation:
\[
W^T U^{-1} W C = W^T U^{-1} M^T,
\]
where
\[
U^{-1} = \text{diag}(U_{q_1}^{-1}, U_{q_2}^{-1}, \ldots, U_{q_T}^{-1})
\]
are the covariance and mean matrices, \( C = [c_1, c_2, \ldots, c_T] \) the trajectory to be solved, and \( W = [w_1, w_2, \ldots, w_T] \) the weighing matrix for calculating dynamic features, respectively.

The none-zero elements of \( w_0 \) are typically composed of three parts, i.e., static, velocity (delta) and acceleration (delta-delta). Usually the window functions are chosen as [5]:
\[
\begin{align*}
w_0^{(0)} &= [0, 1, 0], \\
w_0^{(1)} &= [-0.5, 0, 0.5], \\
w_0^{(2)} &= [-1, 2, 1],
\end{align*}
\]
in which three adjacent frames are considered for dynamic feature calculation.

### 2.2. Trajectory generation discarding acceleration parameters

An objective experiment is performed first to measure the distance between trajectories generated with and without acceleration parameters. A phonetically rich, broadcasting news style, American English speech corpus recorded by a female speaker is used in this study. This corpus contains 12,000 utterances (about 15 hours), and is sampled at 16 kHz. A frame rate of 5 ms is used to perform a 40th-order LPC analysis. The LPC parameters are then converted to LSP coefficients plus gain, as well as their delta and delta-delta counterparts. These features along with the \( f_0 \) features are to train ML-based, context-dependent phone HMMs. Each HMM has a 5-state, left-to-right, and no-skip topology.

100 sentences excluded from the training set are prepared for testing. The reference testing sentences are synthesized using conventional trajectory generation algorithm with full model parameters. LSP and \( f_0 \) trajectories are generated separately. Meanwhile, the sentences without acceleration parameters are also generated for comparison. This can be done by simply setting \( w_0^{(2)} = [0, 0, 0] \) in Eq. (4) during trajectory generation (note that the acceleration parameters are discarded only in trajectory generation, but not in training).

#### 2.2.1. Objective tests

The following objective measures are used to measure the distance between the reference sentences and the ones generated without acceleration parameters:

- Log-spectral distance (dB) for LPC spectra:
\[
D_{\text{spec}} = 20 \sum_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} \left[ \log |S_{\text{ref}}(t, n)| - \log |S(t, n)| \right]^2,
\]
where \( S_{\text{ref}} \) and \( S \) are the magnitude spectra with / without acceleration parameters, \( T \) the number of frames, and \( N \) the number of frequency samples, respectively.

- Root mean square error (RMSE) for \( f_0 \) trajectory:
\[
D_{f0} = \sqrt{\frac{1}{T_{\text{voiced}}} \sum_{t=1}^{T_{\text{voiced}}} \left[ f_{0\text{ref}}(t) - f_0(t) \right]^2},
\]
in which \( T_{\text{voiced}} \) is the total number of voiced frames.

The objective distances of the testing sentences with / without acceleration parameters are summarized in Table 1. The results suggest that the distortion introduced by discarding acceleration parameters is relatively small, i.e., with only 0.36 dB log-spectral distance and 1.29 Hz \( f_0 \) RMSE distance. In speech coding, a spectral distance less than 1 dB between two LPC spectra has long been used as the Difference Limen (DL) for transparent coding [6, 7]. So the speech synthesized with / without acceleration parameters should be perceptually transparent in general. However, our subjective test results indicates the opposite in the following subsection.

#### 2.2.2. Subjective tests

Three native American English speakers are invited to conduct preference testing of the 100 sentence pairs synthesized with / without acceleration parameters. The results of their preference are summarized in Fig. 1.

It is interesting to see that the sentences generated with acceleration parameters are significantly better than those without acceleration. Human subjects can easily perceive the voice quality degradation without acceleration parameters, despite the fact that the objective distortion introduced is rather small. The feedback from the listeners shows that the sentences synthesized without acceleration parameters sound “hoarse,” “sandy,” or not as smooth as the ones generated with full parameters. This result challenges our old knowledge got from speech coding, and inspires us to further investigate the effect of the acceleration parameters in trajectory generation.

### 3. The Effect of Acceleration Parameters in Trajectory Generation

#### 3.1. The saw-tooth trajectory without acceleration constraints

We first compare the trajectories generated with / without acceleration parameters by plotting them together. A typical trajectory segment of about 20 frames is shown in Fig. 2. This

<table>
<thead>
<tr>
<th>Log-spectral Distance</th>
<th>RMSE of f0 Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o acc.</td>
<td>0.36 dB</td>
</tr>
<tr>
<td>w/ acc.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: The subjective preference test results of the testing sentences with / without acceleration parameters.
segment is chopped from one of the synthesized testing sentences, and represents the gain dimension. It is not difficult to see that the trajectory generated without acceleration constraints has saw-tooth-like fluctuation along time axis, although it still closely follows the trend of the reference. Similar observation was also reported in [4], where different orders of the delta windows were compared. We believe the unsymmetric trajectories generated because of the absence of acceleration constraints are the main reason for the strong preference reported in our previous subjective tests.

3.2. Analysis of the Acceleration Constraints in Trajectory Generation

Let’s revisit the trajectory generation algorithm in Eq. (1). When diagonal covariance matrix \( \mathbf{U} \) is used, the left-hand side of the equation can be decomposed into three parts, representing the static, velocity, and acceleration constraints:

\[
\mathbf{W}^\top \mathbf{U}^{-1} \mathbf{W} = \mathbf{W}^{(0)} \mathbf{U}^{(0)-1} \mathbf{W}^{(0)} + \mathbf{W}^{(1)} \mathbf{U}^{(1)-1} \mathbf{W}^{(1)} + \mathbf{W}^{(2)} \mathbf{U}^{(2)-1} \mathbf{W}^{(2)},
\]

where \( \mathbf{U}^{(n)} = \text{diag}(\rho^{(n)}) \) are diagonal variance matrices, and \( \mathbf{W}^{(n)} \) are band diagonal weighting matrices.

Let \( \mathbf{A}^{(n)} \) denote each part of the three summation term, i.e.,

\[
\mathbf{A}^{(n)} = \mathbf{W}^{(n)} \mathbf{U}^{(n)-1} \mathbf{W}^{(n)},
\]

because only a few elements in \( \mathbf{W}^{(n)} \) are non-zero, \( \mathbf{A}^{(n)} \) will be symmetric, band diagonal matrices, too. More specifically, after some mathematical derivation, each element in \( \mathbf{A}^{(n)} \) can be calculated by:

\[
a_{i,i+\delta}^{(n)} = \sum_{l=-L}^{L} w_{-l}^{(n)} w_{-l+i}^{(n)} / w_{i+\delta}^{(n)},
\]

in which \( \delta \) is the offset from the main diagonal, \( L \) is half of the delta window length; and \( w_{i}^{(n)} \) represents the window coefficient for the previous frame, \( w_{0}^{(n)} \) for the current frame, and \( w_{i}^{(n)} \) for the next frame, etc.

From Eq. (9), and when the conventional window coefficients in Eq. (4) are used, the following properties of \( a_{i,i+\delta} \) can be derived:

\[
a_{i,i+\delta}^{(0)} = 0, \quad \text{when } \delta \neq 0,
\]

\[
a_{i,i+\delta}^{(1)} = 0, \quad \text{when } |\delta| = 1 \text{ or } |\delta| > 2,
\]

\[
a_{i,i+\delta}^{(2)} = 0, \quad \text{when } |\delta| > 2.
\]

That is, \( \mathbf{A}^{(0)} \) is a simple diagonal matrix, \( \mathbf{A}^{(1)} \) only has non-zero elements along its main diagonal plus the second diagonals above and below the main diagonal (refer to the matrix structure in Eq. (11) as an example), while \( \mathbf{A}^{(2)} \) is a band diagonal matrix with 5 non-zero diagonals.

Since \( \mathbf{W}^\top \mathbf{U}^{-1} \mathbf{W} = \sum_{n=0}^{2} A^{(n)} \), when acceleration constraints \( \mathbf{A}^{(2)} \) are discarded, the elements along the first diagonals above and below the main diagonal become zero. The left-hand side of Eq. (1) turns out to have the following band diagonal structure:

\[
\begin{pmatrix}
0 & a_{1,3} & 0 & 0 & \cdots \\
0 & 0 & a_{2,4} & 0 & \cdots \\
a_{3,1} & 0 & a_{3,3} & 0 & \cdots \\
0 & a_{4,2} & 0 & a_{4,4} & 0 & \cdots \\
0 & 0 & a_{5,3} & 0 & a_{5,5} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots \\
\end{pmatrix}
\]

Therefore, the odd and even frames of \( \mathbf{C} \) can be solved independently without any constraint. Furthermore, when velocity constraints \( \mathbf{A}^{(1)} \) are removed, the left-hand side of Eq. (1) will become:

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & \cdots \\
0 & 0 & a_{2,4} & 0 & \cdots \\
0 & 0 & 0 & a_{4,4} & 0 & \cdots \\
0 & 0 & 0 & 0 & a_{5,5} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots \\
\end{pmatrix}
\]

which means all frames can be solved independently. To conclude, the special band diagonal structures in Eqs. (11) and (12) explain why a saw-tooth trajectory is generated without acceleration parameters, and why a piecewise constant trajectory is generated without both velocity and acceleration parameters.
3.3. Generating perceptually transparent trajectory without acceleration parameters

There could be several methods to alleviate the perceptual distortion introduced by the saw-tooth trajectory. The most straightforward way is to perform smoothing over the generated trajectory. Other methods may include the use of alternative delta window coefficients, e.g., by setting $w^{(1)} = [-0.5, 0.5, 0]$ or $w^{(2)} = [0, -0.5, 0.5]$. Since the later method may introduce other issues because the reduction of the saw-tooth length also loosens the temporal constraint, we focus on analyzing the first method in this study.

A simple, envelope-based smoothing is chosen for analysis in this paper. The saw-tooth trajectory is smoothed by generating its upper- and lower-bounded envelope as in Fig. 3. This envelope-based smoothing is performed on LSP, gain, and also f0 parameters. It can be shown from the figure that the two envelopes are absolutely smoother than the saw-tooth trajectory. However, it is also clear that the distance between either of the envelope trajectories and the reference does not become smaller than the saw-tooth one. Again, both objective and subjective tests are performed using the envelope trajectories. We want to confirm our previous assumption that it is the unsmooth saw-tooth trajectory that introduces perceptual distortions.

3.3.1. Objective tests

The experimental setups are the same as in Section 2, and the distances of the upper- and lower- enveolpes from the reference are given in Table 2. Compared with the results given in Table 1, it is not surprising to see that both the envelope trajectories have slightly larger spectral distortions than the saw-tooth trajectory. However, the f0 distance actually becomes slightly smaller in this case. Since the differences on the values of the objective measures are rather small, subjective tests are still needed to compare their voice quality.

3.3.2. Subjective tests

We choose the sentences generated by the upper-bounded trajectory to compare with the reference ones generated with full model parameters. The results of the preference test are given in Fig. 4. It is interesting to note that, although the upper-bounded trajectory has a larger spectral distortion than that of the saw-tooth trajectory, the subjects can no longer perceive the voice quality degradation as in our previous subjective test. The results suggest that we are able to generate perceptually transparent speech trajectory without the use of acceleration parameters under certain conditions. And the smoothness of the generated trajectories is as important as our effort of generating speech with less distortions. The envelope-based smoothing method used in our experiments is only a simple trial. Other smoothing methods can also be applied to achieve similar effect.

Our knowledge gained from speech coding is actually renewed by these new results got from HMM-based TTS, i.e., perceptually transparent speech could be generated when 1) the distortion between two samples is relatively small, and 2) the speech does not contain artifacts such as the saw-tooth trajectory presented in this work.

### 4. Conclusions

In this paper, we investigate the effect of the acceleration parameters during trajectory generation in both objective and subjective perspective. We find that discarding the acceleration parameters results in an objectively small (0.36 dB), but subjectively noticeable distortion. The saw-tooth trajectory which causes the voice quality degradation has been investigated. The specific band diagonal matrix structure in the underlying ML-based trajectory generation equation is analyzed. Moreover, we show that the upper- and lower-bounded envelopes of the saw-tooth trajectory which have larger spectral distortions (0.43 and 0.39 dB) are perceptually transparent to the reference synthesized with full parameters. These results shed significant light upon our understanding of human speech perception, and should be useful in many research areas like low-cost TTS system, speech coding, and speech reconstruction.

### 5. Acknowledgements

The authors would like to thank Lyenna Wood, Ian Alcock, and Sebastian Cohen for their participation in the subjective listening tests and their valuable feedbacks.

### 6. References


