Iterative Unit Selection with Unnatural Prosody Detection

Dacheng LIN¹*, Yong ZHAO², Frank K. SOONG², Min CHU² and Jieyu ZHAO³

¹Institute of Computing technology, Chinese Academy of Science, China
²Speech Group, Microsoft Research Asia, China
³dchlin04@mails.gucas.ac.cn (zhaoy, frankkps, mchu)@microsoft.com

Abstract

Corpus-driven speech synthesis is hampered by the occurrence of occasional glitches which ruin the impression of the whole utterance. We propose an iterative unit selection integrated with an unnatural prosody detection model to identify any unnatural prosody. The system searches an optimal path in the lattice, verifies its naturalness by the unnatural prosody model and replaces the bad section with a better candidate, until it passes the verification test. In light of hypothesis testing, we show this trial-and-error approach takes effective advantage of abundant candidate samples in the database. Also, in contrast to conventional prosody prediction, an unnatural prosody detection model still leaves enough room for the prosody variations. Unnaturalness confidence measures are studied. The combined model can reduce the objective distortion by 16.3%. Perceptual experiments also confirm the proposed approach improves the synthetic speech quality appreciably.

Index Terms: speech synthesis, unit selection, confidence measure, unnatural prosody detection, iterative synthesis

1. Introduction

State-of-the-art text-to-speech (TTS) systems [1-3] adopt corpus-driven approaches due to their capability to generate highly natural speech. These systems rely on a very large database of segmental samples, from which the best unit sequence with a minimum distortion cost is retrieved for generating speech output. However, although such a sample-based approach generally synthesizes speech with high-level intelligibility and naturalness, it is bothered by the stability problem that critical errors will occasionally occur and ruin the perception of the whole utterance. This issue hinders TTS from widely spreading in applications of commercial service.

In the literature, various approaches have been proposed to solve the issue with respect to optimizing the cost function [4][5], so as to reflect unnatural just as a human listener might perceive it. The cost function has not yet been explored in full-scale. Here, we cast the focus on the cost function itself. Minimization with a cost function is a process of global optimization, which does not guarantee the goodness of local units. Unfortunately, human perceptions are especially affected by any local inconsistency, and they score intrinsically different than the cost function does.

On the other hand, human intervention is a practical and efficient approach to fix unsmoothed synthesis. For example, we have listeners label unnatural sections in the resulted optimal path, prune out bad units from the search lattice, and re-synthesize to get an alternative path. With several trials, we always succeed in correcting the synthesis problem. In [6], the paper exploited this kind of intuition and presented an interactive speech unit selection scheme, where user's feedback can intervene Viterbi search by removing indicated units from the search lattice.

Inspired by these studies, we propose an iterative speech synthesis framework, named N-pass tuning, to automate this process of human intervention. It involves post-processing the optimized unit path with a confidence measure module, pruning out those incommensurate units and redoing search, until the whole unit path passes.

On the other hand, this framework recently takes into account the implementation of the prosody model. Conventional prosody prediction models aim to predict deterministic prosodic values with input of text transcriptions [7][8]. Repetitious and monotonous prosody patterns are perceived since natural variations in prosody of human speech are replaced with the most frequently used pattern. An important characteristic of a high-quality TTS engine is to inhibit any violation of naturalness and leave room for acceptable prosody variations [9]. Thus, we propose unnaturalness confidence measures to detect any unnatural prosody in synthetic speech, different from conventional prosody models.

Moreover, we demonstrate that in case of a sample-based system, an unnatural prosody detection model even with high false alarm can somewhat meet the requirement. Various unnaturalness confident measures are examined. The objective and subjective evaluation showed our new system significantly improves the voice quality of the synthesized utterances.

This paper is organized as follows. Section 2 introduces the N-pass tuning framework for unit selection. The concept and implementation of unnatural prosody models are described in Section 3. The objective experiments and the perceptual evaluation results are presented in Section 4, and the conclusions in Section 5.

2. N-pass tuning synthesis

N-pass tuning synthesis is literally an iterative procedure with rounds of two-pass scoring. In the first stage, a Viterbi search is performed to find a best unit path conforming to the guidance of the transcription; in the second stage, the sequence of units is scored by verification models to compute likelihood ratios. If there are some unnatural units that do not pass the test, they are pruned out from the lattice, and the next iteration continues. The iterations would continue until a unit sequence entirely passes the verification, or it reaches a preset value of maximum iterations. Figure 1 shows a schematic diagram of this framework.

* This work was carried out while the author was a visiting student at Microsoft Research Asia.
Such a framework takes effective use of plentiful candidate units within the unit database. We will explain this advantage in light of hypothesis testing theory.

![Diagram](image)

Figure 1. Schematic diagram of N-pass tuning synthesis.

### 2.1. Unnatural prosody detection

An unnatural prosody detection model is aimed to detect any occurrence in the synthesized speech that sounds unnatural in prosody. Unnatural prosody includes badly-uttered segments, unsmoothed concatenation or wrong accents and intonations. Given the feature $X$ observed from the synthetic speech, we make a choice between two hypotheses:

- $H_0$: $X$ is natural in prosody
- $H_1$: $X$ is unnatural in prosody

The decision is based on a likelihood ratio test:

$$LR(X) = \frac{P(X | H_0)}{P(X | H_1)} \geq \theta \quad \text{choose } H_0$$

$$\frac{P(X | H_0)}{P(X | H_1)} < \theta \quad \text{choose } H_1$$

(1)

Where $P(X | H_i)$ is the likelihood of the hypothesis $H_i$ with respect to the observed feature $X$.

If $\lambda$ is the loss of deciding $D_i$, when the true class is $H_j$, then the expected risks for two types of errors, false alarm and miss, are:

$$R_f = \lambda_f P(D_f | H_0) P(H_0)$$

(2)

$$R_m = \lambda_m P(D_m | H_1) P(H_1)$$

(3)

As we mentioned, unnatural segments would destroy the perception of the whole utterance, which means that the cost of miss $\lambda_m$ is really high. On the other hand, N-pass tuning will remove detected unnatural tokens, and re-synthesize the utterance. Provided the candidate units are available in a sufficient amount, the cost of mistakenly removing a natural token $\lambda_{01}$, is as small as a run of lattice search.

In summary, this is a two-class classification problem with unequal misclassification costs, and the loss of false alarm is remarkably less than the loss of miss. To minimize the total risk, a sum of $R_f$ and $R_m$, the optimal decision boundary should be intentionally biased against $H_1$, as illustrated in Figure 2. Thus, it means that the unnatural prosody model works at a high false alarm rate, an undersampling requirement for the implementation of confidence measure.

![Figure 2](image)

Another reason why we prefer an unnatural prosody detection model to a prosody prediction model is that deterministic prosody prediction can result in mechanic and monotonous synthetic speech. Actually, people speak with noticeably prosodic variations, even for the same sentence. Multiple realizations in prosody are equally acceptable. Thus, unnatural prosody detection model is expected to outperform a prosody prediction model, because it prevents unnatural prosody from being generated and leaves room for acceptable variations. Besides, this framework has other advantages:

- Loosely coupling with the existing cost function, and making minimum side-effects on cost function.
- Capability to take into account long-term prosodic features, like at syllable and word level.

One disadvantage is that multiple runs of Viterbi search will increase the computation overhead. However, in many scenarios, it is worth pursuing better synthesis quality at the expense of increase of computations, e.g. the system permits users to improve the synthetic quality offline. Another scenario is to provide an option for users to balance between naturalness and computational complexity, so that it particularly operates on words receiving more priorities, such as proper nouns.

### 2.2. Alternative framework

An alternative framework with an unnatural prosody module is to embed it into the Viterbi search and on-line turn off those unnatural paths, without the assistance of synthesis iterations. In this way, an unnatural prosody module actually defines a non-linear cost function, where the cost is close to 0 when the feature distance is below a threshold, and becomes infinity when above that threshold.

This framework may drop several advantages presented in the former approach, such as a high false alarm and loosely coupling with the cost function. In the rest of this paper, we focus on studying the N-pass tuning approach.

### 3. Training unnatural prosody model

As described above, an unnatural prosody model is designed to detect any unnatural prosody in synthetic speech. The ideal approach is to collect an amount of synthetic utterances generated by a TTS system of interest, mark up perceptual scores by experiments or language experts, and learn patterns of unnaturalness from the labeled dataset. However, this approach for unnaturalness learning consumes huge time and effort. Besides, since concatenated speech is a kind of artificial data, distributions of utterances become more arbitrary than those of natural speech, which often results in the poor generalization of the trained models. We decline this method in the paper.

Here, we propose an approach to learn patterns of naturalness from real speech. It is assumed that a synthetic
utterance sounding natural in perception exhibits prosodic characteristics similar to those of real speech:

\[ P(X | H_0) = P(X | N) \]  

(4)

Where \( P(X | N) \) is the probability density of feature \( X \) given in situation of real speech \( N \). Thus, the issue is turned to learn natural prosody from source speech corpus and measure the likelihood of naturalness given synthetic tokens. A decision threshold is chosen in terms of \( P(X | N) \), independent of the distribution of alternative hypothesis \( H_1 \). In this way, it works at Constant False Alarm Rate (CFAR).

We employ decision trees to characterize prosody pattern of real speech. The splitting criterion is to maximize the reduction of Mean Square Error (MSE). Phonetic and prosodic contextual factors, such as phonemes, break indices, stress and emphasis, are taken into account to split trees.

In time of unnaturalness detection, given the observation \( X \) of a token, a leaf node is found by traversing the tree with context features of that token, and the distance between \( X \) and the kernel of the leaf node is used to reflect the likelihood of unnaturalness:

\[ z(X) = \sum_{j=1}^{N} \sigma_j \left( \frac{(x_j - \mu_j)^2}{\sigma_j^2} \right) \]  

(5)

Where \( \mu_j \) and \( \sigma_j \) denotes the mean and standard deviation of the \( j \)-th dimension of the leaf node. When \( z(X) \) is larger than a preset value, it is decided as unnaturalness.

Table 1. Contextual factors involved in decision trees to learn unnatural prosody patterns. \( X \) means the item being checked. As for boundary models, \( D \) denotes the difference between left/right tokens, and \( L/R \) denotes including both left/right tokens.

<table>
<thead>
<tr>
<th>Contextual factors</th>
<th>Phon</th>
<th>PhonBnd</th>
<th>Syll</th>
<th>SyllBnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position of word in phrase</td>
<td>X</td>
<td>L/R</td>
<td>X</td>
<td>L/R</td>
</tr>
<tr>
<td>Position of syllable in word</td>
<td>X</td>
<td>L/R</td>
<td>X</td>
<td>L/R</td>
</tr>
<tr>
<td>Stress</td>
<td>X</td>
<td>L/R</td>
<td></td>
<td>L/R</td>
</tr>
<tr>
<td>Current phoneme</td>
<td>X</td>
<td>L/R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left/right phoneme</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Break index of boundary</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 2. Acoustic features used in unnatural prosody model.

<table>
<thead>
<tr>
<th>Acoustic features</th>
<th>Phon</th>
<th>PhonBnd</th>
<th>Syll</th>
<th>SyllBnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>X</td>
<td>D</td>
<td>X</td>
<td>D</td>
</tr>
<tr>
<td>( F_0 ) mean, std. dev.</td>
<td>D</td>
<td>X</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>( F_0 ) at head, middle,</td>
<td>D</td>
<td>X</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>tail</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
</tr>
</tbody>
</table>

In this paper, we explored confidence measures in 4 token types, including phoneme (Phn), phoneme boundary (PhonBnd), syllable (Syll) and syllable boundary (SyllBnd). Models Phn and Syll are aimed to measure the fitness of prosody, and Models PhonBnd and SyllBnd are to reflect the transition smoothness of spliced units. The contextual factors and observation features for each decision tree are described in Table 1 and Table 2.

Phonemes are basic units in our TTS system, and likewise basic units for lattice pruning. The system removes from the lattice units with a score above a threshold. As for Models Phn and Syll, confidence scores estimated by models are duplicated to all phonemes enclosed by the focused tokens. As for Models PhonBnd and SyllBnd, confidence scores are divided into halves and assigned to the left/right tokens. We also build a combined model, which consists of the above four models, and sums up their scores with equal weights.

4. Experiments

4.1. Experimental setup

In our previous work on English TTS [3], a soft prosody-constrained unit selection algorithm was developed, which bypasses the prosody model that predicts numeric prosodic parameters for synthetic speech. Units in the database are labeled and clustered by their phonetic and prosodic contexts. For each target unit, a leaf node is found by traveling through the tree with context features, and all its instances are pooled in the search lattice. Viterbi search is adopted to find the best path by minimizing the overall cost function. Speech is generated by directly concatenating segments without any pitch and duration modification.

Based on this framework, we implemented N-pass tuning synthesis by integrating unnatural prosody models and the recursive dataflow on search stage. Other parts, including the cost function and speech concatenation, are unchanged.

The Microsoft Mulan English speech corpus is used to evaluate the performance of the proposed method. It was recorded by a professional speaker and annotated with symbolic prosodic labels, like break level, stress, emphasis, etc. The speech corpus consists of about 6,000 sentences, among which 5,000 sentences are used to build the voice font and train the unnatural prosody model, and the rest 1,000 sentences are held for objective evaluation.

4.2. Distortion measures

An objective function is designed to evaluate the speech synthesis system by calculating a distortion measure between the synthetic speech and its natural counterpart. Let \( O = \{a_1, a_2, \ldots, a_N\} \) and \( \tilde{O} = \{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_N\} \) be a string of feature vectors observed on \( N \) units in natural speech and synthesized speech, and let \( B = \{b_1, b_2, \ldots, b_N\} \) and \( \tilde{B} = \{\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_N\} \) be a string of feature vectors observed on boundaries between those units in natural speech and synthesized speech, then the distortion measure can be written as:

\[ E(O, \tilde{O}) = \sum_{j=1}^{N} E_j(O, \tilde{O}) + \sum_{j=1}^{Q} E'_j(B, \tilde{B}) \]  

(6)

Where \( E_j(O, \tilde{O}) \) denotes the similarity distortion on unit feature \( j \). \( E'_j(B, \tilde{B}) \) denotes the smoothness distortion on boundary feature \( j \). \( P \) and \( Q \) are the dimension of unit and boundary features respectively, and weight \( \kappa \) is to balance the two factors. These two sub-terms are estimated by:

\[ E_j(O, \tilde{O}) = \frac{1}{N\sigma_j} \sum_{i=1}^{N} (a_i - \tilde{a}_i)^2 \]  

(7)

\[ E'_j(B, \tilde{B}) = \frac{1}{(N-1)\zeta_j} \sum_{i=1}^{N} (\tilde{b}_i)^2 \]  

(8)

Where \( \sigma_j \) and \( \zeta_j \) are the standard deviation of the \( j \)-th feature of unit string \( O \) and boundary string \( B \) respectively.

Features for the similarity distortion include duration, \( F_0 \) mean of units, and features for the smoothness distortion include \( F_0 \) difference at boundaries.
4.3. Objective evaluation

The first experiment was to objectively evaluate the N-pass tuning synthesis framework with various unnaturalness confidence measures. We compared the distance between the synthesized speech and the natural speech, by synthesizing sentences from the held-out dataset. The average search iterations to produce a valid path are referred for comparison among different models. Figure 3 shows the characteristics of objective distortions with respect to average iterations by modifying the decision thresholds.

All models show significant improvements over the baseline (Average iterations = 1), and the best performance is obtained after 4-7 iterations. The combined method exhibits the best results among 5 models. 4 individual models can be arranged in descending order of the performance, Phn, PhnBnd, Syl and SyIBnd. The phone-based models (Phn and PhnBnd) distinctively outperform the syllable-based models (Syl and SyIBnd). One possible reason is the weakness of models in capturing long term features. Also, it is observed that the performance may begin to slowly degrade after 6 or 7 runs. We offer an explanation that the lattice may be overpruned, and the unit database we used does not involve so abundant candidate units to support such a frequency of lattice pruning.

Table 3. Distortions in each acoustic feature for unnatural prosody models with their best configurations.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dur</th>
<th>F0</th>
<th>F0 Diff</th>
<th>Total</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.74</td>
<td>1.04</td>
<td>1.72</td>
<td>3.93</td>
<td>-</td>
</tr>
<tr>
<td>Phn</td>
<td>0.66</td>
<td>0.89</td>
<td>1.33</td>
<td>3.36</td>
<td>14.5</td>
</tr>
<tr>
<td>PhnBnd</td>
<td>0.68</td>
<td>0.97</td>
<td>1.17</td>
<td>3.44</td>
<td>12.5</td>
</tr>
<tr>
<td>Syl</td>
<td>0.70</td>
<td>0.95</td>
<td>1.51</td>
<td>3.62</td>
<td>7.9</td>
</tr>
<tr>
<td>SyIBnd</td>
<td>0.73</td>
<td>0.98</td>
<td>1.42</td>
<td>3.64</td>
<td>7.4</td>
</tr>
<tr>
<td>Combined</td>
<td>0.67</td>
<td>0.90</td>
<td>1.19</td>
<td>3.29</td>
<td>16.3</td>
</tr>
</tbody>
</table>

Table 3 lists distortions in each acoustic feature for those models with their best configurations. Dur, F0 mean and F0 Diff stand for duration and F0 mean of units, and F0 difference at boundaries respectively. The shadowed cells indicate which individual model obtains the best performance for the feature in focus. It shows that Model Phn obtains the best result in features related to the fitness of unit, while Model PhnBnd obtains the best result in features related to the prosody smoothness, an observation consistent with the objective of those models. The combined model gets a distortion reduction of 16.3%.

4.4. Perceptual evaluation

A preference test was conducted to perceptually evaluate the performance of the N-pass tuning synthesis. 30 pairs of sentences were used for listening test. 8 subjects participated in the test and they were forced to choose the one from each pair which sounds more natural.

The result for the preference test is given in Table 4. It shows that the synthetic speech obtained by the proposed method sounds much better than that from baseline system.

Table 4. Preference test results.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>N-pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>36.7%</td>
<td>63.3%</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we proposed a novel unit selection algorithm, which is an iterative synthesis integrated with an unnatural prosody model to detect any prosodic unnaturalness. The framework takes advantage of the abundance of candidate units in the database, to improve the synthesis quality. We discussed the advantages of an unnatural prosody detection model over a conventional prosody prediction model and pointed out that the new model accommodates different prosodic variations in synthesized speech. Both objective and subjective experiments show that the proposed approach improves the synthetic speech quality significantly.

Future work involves new unnaturalness confidence measure and efficient combination of individual models through model selection and weight tuning.

6. References