INTERSPEECH 2009 BRIGHTON

Rich Context Modeling for High Quality HMM-Based TTS

Zhi-Jie Yan *Yao Qian* *Frank K. Soong*

Microsoft Research Asia, Beijing, China

{zhijiey, yaoqian, frankkps}@microsoft.com

Abstract

This paper presents a rich context modeling approach to high quality HMM-based speech synthesis. We first analyze the over-smoothing problem in conventional decision tree tying-based HMM, and then propose to model the training speech tokens with rich context models. Special training procedure is adopted for reliable estimation of the rich context model parameters. In synthesis, a search algorithm following a context-based pre-selection is performed to determine the optimal rich context model sequence which generates natural and crisp output speech. Experimental results show that spectral envelopes synthesized by the rich context models are with crisper formant structures and evolve with richer details than those obtained by the conventional models. The speech quality improvement is also perceived by listeners in a subjective preference test, in which 76% of the sentences synthesized using rich context modeling are preferred.

Index Terms: HMM-based TTS, rich context modeling

1. Introduction

Hidden Markov Model (HMM) based TTS has become one of the most important approaches for speech synthesis in the last decade [1]. Conventional HMM-based speech synthesis uses a universal Maximum Likelihood (ML) criterion in both training and synthesis. The ML criterion is capable of estimating statistical parameters of the HMMs effectively. It also imposes a static-dynamic parameter constraint in synthesis, which helps to generate smooth parametric trajectory that yields highly intelligible speech.

However, speech synthesized using conventional HMM-based approach is generally too smooth, because the ML parameter estimation after decision tree-based tying usually leads to highly averaged HMM parameters. This problem is especially serious in estimating the mean parameters of HMMs, since mean is relatively more important than variance, and carries the most critical information about the spectral "snap-shot" (by static mean) and the spectral "transition" (by dynamic mean). When mean parameters are tied and overly smoothed, the synthesized speech becomes blurred and muffled. As a result, the over-smoothing in conventional HMM-based synthesis is one of the key factors of the degraded speech quality.

Various ways have been proposed to alleviate this over-smoothing in synthesized speech. Some aims at improving modeling resolution, either by increasing the number of leaf-nodes in decision tree-based parameter tying, or by increasing the number of Gaussian components in each decision tree leaf node [2]; some proposes an alternative training criterion, e.g., Minimum Generation Error (MGE) [3], to improve model sharpness; other remedial procedures are proposed in synthesis stage, or as post-processing after the parametric trajectory is generated. They include Global Variance (GV) constraint [4] in trajectory generation, and post-processing like formant sharpening [5].

In this paper, we propose to use rich context models (or "full context models") to model training speech data more faithfully. In synthesis, the same rich context models are used to generate speech trajectory. The main difference of this new approach from the conventional HMM-based one is that the decision tree-based parameter tying is no longer used for synthesis. Because rich context models without tying hold rich segmental and suprasegmental information, they are "crisper" than the conventional, tied models by definition. Therefore, the over-smoothing problem is significantly alleviated.

There are still problems to be resolved in rich context modeling: In training phase, the data alignment of the modeling units (e.g., phonemes) may not be as accurate as the conventional method; variance parameters are usually under-estimated for rich context models due to sparse training data; In synthesis phase, because the conventional decision tree is not built for rich context models, the unseen target labels are not directly mapped to seen models. So a new framework is needed to determine the appropriate rich context model sequence for the target label sequence.

This paper presents our approach to solving the above problems in rich context modeling. In training, a single-pass re-estimation method is adopted to estimate the mean parameters of the rich context models, using data aligned with the conventional, tied models. Variance parameters of rich context models are still tied according to the same tree-structure derived from the conventional method. In synthesis, a context-based, pre-selection is first performed to narrow down the search space to a sausage of rich context model candidates. Then, the optimal rich context model sequence is determined by considering the distance between the guiding, tied model sequence and all candidate rich context model sequences. In this paper, an approximated criterion calculated from the state-aligned, upper-bound of the Kullback-Leibler Divergence (KLD) between two MSD-HMM [6] state sequences is implemented. Finally, the output parametric speech trajectory is generated by the optimal rich context model sequence, using the ML trajectory generation algorithm.

We compared the conventional decision tree-tied modeling and our rich context modeling in an American English database. The experimental results show that synthesized speech spectrum with rich context models has sharper formant structure and richer acoustic details. The speech quality improvement is also proved by a subjective preference test.

The rest of this paper is organized as follows: In Section 2, the over-smoothing problem in conventional HMM-based speech synthesis is analyzed; In Sections 3 and 4, the rich context modeling and synthesis are introduced, respectively; In section 5 we present the experimental results. In Section 6, we draw our conclusions.

Copyright © 2009 ISCA

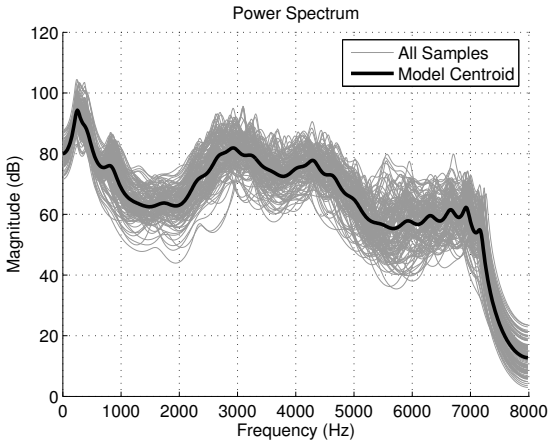

Figure 1: *Example of the over-smoothing problem in a typical decision tree leaf-node.*

2. The Over-Smoothing Problem in HMM-Based Speech Synthesis

The over-smoothing problem in HMM-based speech synthesis is mainly caused by too much averaging in the parameter estimation process. In the training phase, smoothing is performed in the feature, frequency, temporal, and contextual domains. In the feature domain, we assume the speech signal is quasi-stationary within the analysis time window. Some smoothing method may also be performed on the raw features prior to model training; In the frequency domain, parametric speech features (e.g., line spectral pairs (LSPs), Mel-cepstrum coefficients) representing formants are averaged to estimate HMM parameters; In the temporal domain, several successive feature frames are averaged to estimate the corresponding HMM state; Finally in the contextual domain, allophones of different contexts are tied together in a decision tree structure. All speech instances belonging to the same decision tree leaf-node are averaged again to estimate the model parameters.

Smoothing is necessary and important for the robustness of the estimated model parameters. But too much smoothing tends to degrade the synthesized speech quality. Especially for the mean parameters which are computed to represent the spectral "snap-shots" (static mean) and "transitions" (dynamic mean), the over-smoothing problem introduces undesirable distortions that blur the synthesized speech.

In this study, we focus on the over-smoothing problem in the contextual domain, which is introduced mainly by decision tree-based parameter tying. Decision tree was first used in speech recognition, it has several advantages in handling the data sparseness problem and predicting unseen context. However, the situation is different in synthesis compared with recognition. In speech recognition, parameter tying via a tree structure is effective in dealing with speaker and environment variability. But in synthesis, speech models are usually trained with sentences recorded by a single speaker in a well-controlled environment. Decision tree-based parameter tying not only introduces extra smoothing in the contextual domain, but also aggravates smoothing in other domains.

Fig. 1 illustrates an example of the real speech samples belonging to one decision tree leaf-node (gray lines). The estimated spectral envelope for this leaf-node is recovered by its static mean parameters (dark black line). We observe from the figure that the magnitude and center-frequency of the formants of real speech samples vary dramatically even in one leaf-node. But the contextual domain smoothing averages all training instances, and thus can only get a model with blurred, less distinctive spectral peaks (formants). Apparently, the spectral resolution of this leaf-node is far from adequate to represent. The model parameters are overly smoothed by decision tree-based parameter tying.

3. Rich Context Modeling of Training Speech

3.1. Rich context modeling of speech

Many contextual factors can affect the production of human speech. The spectrum, pitch and duration are interacting with one another in synthesizing natural sounding speech. The most important contextual factors include phone identity, stress, accent, position, etc. In HMM-based speech synthesis, the label of the HMMs is composed of a combination of these contextual factors. Before decision tree-based parameters tying, these models are called "rich context models" (or "full context models"). Each rich context model carries rich segmental and suprasegmental information.

In this paper, we propose to directly use rich context models to model training speech segments. Because decision tree is not used in rich context modeling, smoothing in the contextual domain is eliminated.

3.2. Rich context model training for speech synthesis

There are several potential problems in rich context model training: 1) Rich context models are often initialized by cloning mono-phone models. In the conventional training procedure, the data alignment for training rich context models is usually not accurate enough. The poor data alignment may result in ill-estimated model parameters; 2) As the contextual factors increase, the number of rich context models increases exponentially. For each rich context model seen in the training corpus, usually there are only one or at most few instances available for training. As a result, the variance parameters of the rich context models are in general under-estimated. A special training procedure is needed to estimate the HMM parameters of the rich context models reliably.

In our study, the following model training procedure is designed: 1) A decision tree-tied HMM model set is trained using conventional method. This model set is used as the reference; 2) A single-pass re-estimation is performed to estimate the mean parameters of the rich context models. This re-estimation process uses the tied models to get the state-level alignment of the training data. The mean parameters of the rich context models are then estimated according to this alignment; 3) The variance parameters of the rich context models are tied using the conventional tree structure, i.e., the variance parameters of rich context models are set to be equal to that of the tied models.

By using this new training procedure, the accuracy of the data alignment for training is insured by the conventional tied model set. The mean parameters of the rich context models are typically estimated using only a few frames (generally less than 10 frames for each state at a rate of 200 frames per second). For variance parameters, as we are not able to observe enough data, the approximation is necessary to avoid typical numerical singularity problems. Because less smoothing is performed in the new training process, rich context models are "crisper" than

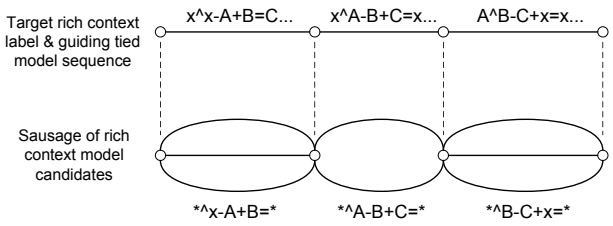

Figure 2: *Sausage of rich context model candidates. Some contextual factors of the target labels are omitted by "..." for convenience, and "*" is the wildcard matching all possible contextual factors.*

the tied models. It then becomes possible to synthesize crisper, high quality speech.

4. Parametric Speech Synthesis Using Rich Context Models

In synthesis, target rich context label sequence is predicted by the front-end text analysis. Due to the large combinational possibilities of all contextual factors, the target rich context labels are usually not seen in the training data. The conventional approach uses the decision tree to map the unseen labels unto seen models. However, when rich context modeling is used, an alternative method is needed. In our approach, a search algorithm following a context-based pre-selection is performed.

4.1. Pre-selection of rich context model candidates

Firstly, pre-selection is performed to compose a rich context model candidate sausage. This process can be illustrated in Fig. 2. For each target rich context label, only its tri-phone-level context is extracted as a pattern. Then, all rich context models seen in the training corpus that match this tri-phone pattern are chosen to form a sausage node (if the tri-phone pattern is unseen in the corpus, a bi-phone-level pre-selection is performed instead). Successive nodes are then connected to compose a sausage.

The purpose of this tri-phone-level, context-based pre-selection is to solve the unseen problem in rich context modeling, while maintaining a reasonable size of the search space for latter process. Considering the scale of the corpus we used in our experiments (refer to Section 5 for more details), the tri-phone-level pre-selection is able to keep a good balance between candidate coverage and search space size. Typically there are less than 150 candidates per node in our experiments after pre-selection.

4.2. Search of the optimal rich context model sequence

Secondly, the "optimal" rich context model sequence is determined in the sausage, which will be used to generate speech trajectory. In our approach, the tied model sequence predicted by conventional HMM-based method is used as the guiding model sequence. The optimal rich context model sequence is defined to be the one which is "closest" to this guidance in global trend.

The reason of choosing the tied model as the guiding sequence is based on its capability in providing a smooth but stable trajectory trend. Following this stable trend, the global smoothness of the trajectory generated using rich context models can be obtained. Meanwhile, the local speech fidelity is improved by using rich context models because they are crisper than conventional, tied models.

The key problem in searching for the "closest" rich context model sequence is how to measure the distance between guiding tied model sequence and rich context model candidate sequences. In this paper, an upper-bound of the state-aligned KLD approximation is adopted as the distance measure, in which spectrum, pitch, and duration information are considered simultaneously.

Given $P = \{p_1, p_2, \ldots, p_N\}$ as the guiding, tied model state sequence. Its state-level duration is determined by using the conventional duration model, denoted as $T = \{t_1, t_2, \ldots, t_N\}$. For each rich context model candidate sequence, we set its state sequence to be aligned to the guiding sequence in a one-to-one mapping. Because of the sausage structure, the number of states is the same for both the guiding and candidate sequences. Therefore, the candidate state sequence is denoted as $Q = \{q_1, q_2, \ldots, q_N\}$, and it shares the same duration with the guiding state sequence.

We define the following approximated criterion to measure the distance between the guiding and the candidate model sequences:

$$D(P, Q) = \sum_n D_{\mathrm{KL}}(p_n, q_n) \cdot t_n, \tag{1}$$

in which $D_{\mathrm{KL}}(p, q) = D^{\mathrm{S}}_{\mathrm{KL}}(p, q) + D^{\mathrm{f0}}_{\mathrm{KL}}(p, q)$ is the sum of the upper-bound KLD for spectrum and pitch parameters between two MSD-HMM states [7]:

$$\begin{aligned} D^{\mathrm{S/f0}}_{\mathrm{KL}}(p, q) \leq\ & (w_0^p - w_0^q) \log \frac{w_0^p}{w_0^q} + (w_1^p - w_1^q) \log \frac{w_1^p}{w_1^q} \\ & + \frac{1}{2} \mathrm{tr} \Big\{ (w_1^p \boldsymbol{\Sigma}_p^{-1} + w_1^q \boldsymbol{\Sigma}_q^{-1})(\mu_p - \mu_q)(\mu_p - \mu_q)^\top \\ & + w_1^p (\boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_q^{-1} - \mathbf{I}) + w_1^q (\boldsymbol{\Sigma}_q \boldsymbol{\Sigma}_p^{-1} - \mathbf{I}) \Big\} \\ & + \frac{1}{2} (w_1^q - w_1^p) \log \left| \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_q^{-1} \right|, \end{aligned} \tag{2}$$

where w_0 and w_1 are prior probabilities of the discrete and continuous sub-space (for $D^{\mathrm{S}}_{\mathrm{KL}}(p, q)$, $w_0 \equiv 0$ and $w_1 \equiv 1$); μ and $\boldsymbol{\Sigma}$ are mean and variance parameters, respectively.

Using Eqs. (1) and (2), spectrum, pitch and duration are embedded in one distance measure. The optimal rich context model sequence is chosen in the sausage so as to minimize the total distance $D(P, Q)$. The search process is quite straightforward in a sausage structure, i.e., the best rich context candidate models for every sausage node are chosen to form the global optimal solution.

Lastly, speech trajectory is generated using the optimal rich context model sequence, by solving the conventional ML generation equation [2]. Because the variance parameters of the rich context models are still tied as the conventional method, there would be no numerical problems (e.g., divide by a small number) in solving the corresponding weighted least squares equations. Since rich context models are used directly, post-processing of the generated trajectory (e.g., formant sharpening) is no longer necessary.

5. Experiments

5.1. Experimental setup

A phonetically rich, broadcast news style, American English speech corpus recorded by a female speaker is used in our experiments. This corpus contains 12,000 sentences (about 15 hours) sampled at 16 kHz. 40^{th}-order LSP coefficients plus

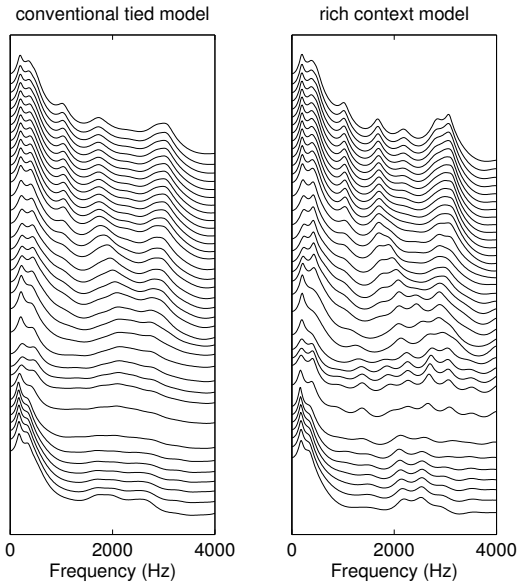

Figure 3: *The running LPC spectral envelope of a syllable composed of two voiced phones (/ax n/), synthesized using conventional decision tree-tied models (left) and rich context models (right).*

gain, as well as their first and second order dynamic features are used to train the ML-based, decision tree-tied baseline model. HMMs of 5-states, left-to-right, no-skip topology with diagonal covariance matrix are used to build all phone models. There are 666,794 different rich context phone models seen in the training corpus. After phone-dependent decision tree-based state tying, there are a total number of 18,828 leaf-nodes for spectral features. So the compression ratio is around 180, i.e., on the average, 180 HMM states are tied into one leaf-node and averaged.

A separated test set of 50 sentences is selected in our experiments. Parametric speech trajectories are synthesized by conventional decision tree-tied models, and our new models using rich context modeling. We compared the synthesized spectrum, and involved six language experts in a subjective preference test. Each subject was asked to listen to 50 pairs of synthesized sentences and provide their feedbacks.

5.2. Implementation issues

In rich context modeling, the number of model parameters is significantly increased. Comparing with conventional method, extra computational cost is needed. In our implementation, all rich context models are re-arranged according to their tri-phone labels. With the tri-phone-based model pre-selection, individual sausage node can be efficiently generated. Meanwhile, the KLD values between the rich context model states and the possible decision tree leaf-nodes are calculated and stored in advance to speed up the search process.

5.3. Experimental results

The parametric speech trajectories generated by the two methods are converted to LPC spectral envelopes and compared. An example of a 35-frame syllable segment is shown in Fig. 3. It can be seen from the figure that, by using rich context models, the synthesized speech has more distinctive, less blurred formant structure than the conventional tied models. The formant

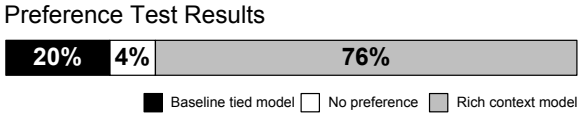

Figure 4: *The subjective test results shown the preference of the synthesized speech using conventional decision tree-tied models and rich context models.*

evolution and the spectral fine details are also richer than that of the tied model. This result demonstrates that rich context models can synthesize more natural, crisper output speech.

The preference test results from all 6 subjects are summarized in Fig. 4. Most sentences (i.e., 76% vs. 20%) synthesized using rich context modeling are preferred by the listeners. The test feedback indicates that the improvement is mainly from the crispness and clarity in speech quality.

6. Conclusions

In this paper, we investigate the over-smoothing problem in the conventional HMM-based speech synthesis, and propose to use rich context models to alleviate or eliminate the sound muffling caused by overly smoothed HMM parameters. In synthesis, rich context model sequence is searched efficiently in a structured sausage. Duration, spectrum and pitch information are universally compared with a guiding model sequence, and the optimal rich context model sequence is obtained by minimizing the corresponding KL divergence. The spectrum synthesized by rich context models is richer and crisper than the conventional baseline. Subject preference test also indicates that the speech synthesized using rich context models is significantly preferred (76%) over the conventional tied models (20%).

7. Acknowledgements

The authors would like to thank all language experts who participated in the subjective tests and provided their valuable feedbacks.

8. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EuroSpeech1999*, vol. 5, 1999, pp. 2347–2350.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP2000*, vol. 3, 2000, pp. 1315–1318.

[3] Y.-J. Wu, W. Guo, and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP2006*, 2006, pp. 89–92.

[4] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. Eurospeech2005*, 2005, pp. 2801–2804.

[5] Y.-J. Wu, "Investigations on HMM based speech synthesis," Ph.D. dissertation, University of Science and Technology of China, 2006, (in Chinese).

[6] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[7] H. Liang, Y. Qian, F. K. Soong, and G. Liu, "A cross-language state mapping approach to bilingual (Mandarin-English) TTS," in *Proc. ICASSP2008*, 2008, pp. 4641–4644.