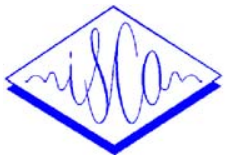
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

DECISION-TREE BACKING-OFF IN HMM-BASED SPEECH SYNTHESIS

Shunsuke Kataoka, Nobuaki Mizutani, Keiichi Tokuda, Tadashi Kitamura

Department of Computer Science and Engineering, Nagoya Institute of Technology
{shunsuke,miztani,tokuda,kitamura}@ics.nitech.ac.jp

10.21437/Interspeech.2004-447

Abstract

This paper proposes a decision-tree backing-off technique for an HMM-based speech synthesis system. In the system, a decision-tree based context clustering technique is used for constructing parameter tying structures. In the context clustering, the MDL criterion has been used as a stopping criterion. In this paper, however, huge decision-trees are constructed without any stopping criterion. In the synthesis phase, decision-trees obtained in this way are used in the proposed backing-off scheme. This enables us to adjust the cluster size dynamically at run-time according to the text to be synthesized. Results of subjective listening tests show that the proposed technique improves the synthesized speech quality.

1. Introduction

In the last decade, extensive studies have been carried out on text-to-speech synthesis. Although significant improvements in the quality of synthesized speech have been shown by various kinds of speech synthesis systems, it is still difficult to synthesize speech with various voice characteristics such as speaker individualities, speaking styles, and emotions. To synthesize various speech, an HMM-based speech synthesis system [1] was proposed. The feature of the system is that it generates synthetic speech from HMMs themselves [2]. Accordingly, it can generate various voice characteristics by transforming HMM parameters with a speaker adaptation technique [3], a speaker interpolation technique [4], or an eigen voice technique [5], even though its synthetic speech is "vocoded."

In the HMM-based speech synthesis system, decision-tree based context clustering technique is used for constructing parameter tying structure. It can assign a sufficient amount of training data to each cluster. In the context clustering, factors which affect acoustic property such as preceding or succeeding phoneme, part of speech, and accentual phrase are regarded as contexts. Generally, decision-tree construction procedure is stopped with a threshold such as a gain in log likelihood. In the HMM-based speech synthesis system, a Minimum Length Description (MDL) criterion is used as the stopping criterion [6]. However, we have observed that the speech quality can be improved by enlarging decision-trees. Meanwhile, overly large decision-trees make the speech quality discontinuous and unstable. It is also observed that the optimal tree size varies according to the text to be synthesized.

From this point of view, this paper proposes a decision-tree backing-off technique for HMM-based speech synthesis. It can vary each cluster size dynamically at run-time according to the text to be synthesized. The unit selection speech synthesis approaches using backing-off techniques based on clustering have also been proposed (e.g., [7]). Our technique, however, differs from them in that backing-off is performed based on a statistical criterion: the algorithm searches for an optimal size of the decition-tree in such a way that the output probability of the speech parameter sequence generated from HMMs whose output vector consist of static and dynamic features is maximized. The results of subjective listening tests show that the backing-off technique improves the quality of synthesized speech.

The rest of this paper is organized as follows. Section 2 and Section 3 describe the HMM-based speech synthesis system and the backing-off technique, respectively. Experimental results are presented in Section 4, and concluding remarks and our plans for future work are presented in the final section.

2. HMM-Based Speech Synthesis System

2.1. Spectram, $F0$, State Duration Modeling

In HMM-based speech synthesis system [1], spectrum, $F0$ and state duration are modeled simultaneously in a statistical framework. We use mel-cepstral coefficients as spectral parameter and model sequences of mel-cepstral coefficient vectors with continuous density HMMs. On the other hand, we cannot apply the conventional discrete or continuous HMMs to $F0$ pattern modeling since the observation sequence of $F0$ pattern is composed of one-dimensional continuous values and a discrete symbol which represents "unvoiced." Then, we apply a hidden Markov model based on multi-space probability distribution (MSD-HMM) [8] for $F0$ pattern modeling. The MSD-HMM includes discrete HMM and continuous mixture HMM as special cases and furthermore can model the sequence of observation vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols.

We construct spectral and $F0$ models by using embedded training. The embedded training dose not need label boundaries when appropriate initial models are available. However, if spectral models and $F0$ models are trained separately, speech segmentations may be discrepant between them. To avoid this problem, context dependent HMMs are trained with feature vectors each of which consists of spectrum, $F0$ and their dynamic features.

State duration densities are modeled by single Gaussian distributions. Since we assume left-to-right models, the dimensionality of state duration densities is equal to the number of states in an HMM, i.e., the n-th dimension of state duration densities is corresponding to the n-th state of HMMs. State duration densities are estimated by using statistics which are obtained in the last iteration of embedded training.

2.2. Decision-Tree Based Context Clustering

As contextual factors increase, the number of their combinations, i.e., the number of the context dependent models also increases exponentially. Thus model parameters with sufficient accuracy cannot be estimated with limited training data. Fur-

thermore, it is impossible to prepare speech database which includes all combinations of contextual factors.

To overcome this problem, we apply a decision-tree based context clustering technique for spectral, $F0$ and state duration models. The "questions" about contextual factors are used for cluster (node) separation in the decision-tree construction. This seems to be good idea for constructing parameter tying structure of context-dependent models that effectively share model parameters. Once we have a decision-tree, any contextual combinations can descend to one of the leaf nodes while answering questions assigned to every node. Furthermore, even unseen contextual combinations can find proper model uniquely by using decision-trees.

Generally, the decision-tree construction procedure is stopped with a threshold such as a gain in log likelihood. In the HMM-based speech synthesis system, an MDL criterion is used as the stopping criterion [6]. Since each of spectrum, $F0$ and duration have its own contextual influences, the distributions for spectrum, $F0$ and state duration are clustered independently. Since it is also considered that contextual influences are different according to the state position in an HMM, the clustering algorithm is applied not for a model unit but for every state-position separately.

2.3. The Synthetic Part

In the synthetic part, a given text to be synthesized is converted to a context-based-label sequence. Then, according to the label sequence, a sentence HMM λ is constructed by concatenating context-dependent HMMs. A sequence of mel-cepstral coefficients and $F0$ parameters including voiced/unvoiced decisions is generated from the sentence HMM by using a speech parameter generation algorithm based on a criterion which maximizes the output probability [2]. First, state durations are determined so as to maximize the probability of the state sequence $\boldsymbol{q}$:

$$\log P(\boldsymbol{q}|\lambda) = \sum_{k=1}^{K} \log p(d_k|n_k) \quad (1)$$

where

$$\begin{aligned} \boldsymbol{q} &= \{q_1, q_2, \ldots, q_T\} \\ &= \{\overbrace{\underbrace{n_1, \ldots, n_1}_{d_1}, \underbrace{n_2, \ldots, n_2}_{d_2}, \ldots, \underbrace{n_K, \ldots, n_K}_{d_K}}^{T}\} \end{aligned} \quad (2)$$

and $p(d|n)$ is the probability of d consecutive observations in state n, K is the total number of states which have been visited during T observations. It should be noted that we assume left-to-right models with no skip. Then, for the state sequence $\boldsymbol{q}$, a speech parameter sequence $\boldsymbol{o} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']'$ is determined in such a way that the output probability of the speech parameter sequence:

$$\begin{aligned} &\log P(\boldsymbol{o} \,|\, \boldsymbol{q}, \lambda) = \log P(\boldsymbol{o} \,|\, \boldsymbol{\mu_q}, \boldsymbol{U_q}) \\ &= -\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu_q})' \boldsymbol{U_q}^{-1} (\boldsymbol{o} - \boldsymbol{\mu_q}) - \frac{1}{2} \log |\boldsymbol{U_q}| + \text{Const} \end{aligned} \quad (3)$$

is maximized, where

$$\boldsymbol{\mu_q} = [\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \ldots, \boldsymbol{\mu}_{q_T}']' \quad (4)$$

$$\boldsymbol{U_q} = \text{diag}\,[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \ldots, \boldsymbol{U}_{q_T}] \quad (5)$$

and $\boldsymbol{\mu}_{q_t}, \boldsymbol{U}_{q_t}$ are mean vector and the covariance matrix of state q_t in t-th frame.

It should be noted that (3) is maximized when $\boldsymbol{o} = \boldsymbol{\mu_q}$, which is a step-wise function. To avoid this problem, we assume that each state output vector $\boldsymbol{o}_t$ consist of static and dynamic vectors:

$$\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta \boldsymbol{c}_t', \Delta^2 \boldsymbol{c}_t']' \quad (6)$$

where

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L^{(1)}}^{L^{(1)}} w^{(1)}(\tau) \boldsymbol{c}_{t+\tau}, \quad (7)$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L^{(2)}}^{L^{(2)}} w^{(2)}(\tau) \boldsymbol{c}_{t+\tau}. \quad (8)$$

These equations can be arranged in a matrix form:

$$\boldsymbol{o} = \boldsymbol{W} \boldsymbol{c} \quad (9)$$

where

$$\boldsymbol{c} = [\boldsymbol{c}_1, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_T]' \quad (10)$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_T]', \quad (11)$$

$$\boldsymbol{w}_t = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}], \quad (12)$$

$$\begin{aligned} \boldsymbol{w}_t^{(n)} = [&\underbrace{0, \ldots, 0}_{1\text{st}}, \underbrace{w^{(n)}(-L^{(n)})}_{(t-L^{(n)})\text{-th}}, \ldots, \underbrace{w^{(n)}(0)}_{t\text{-th}}, \\ &\ldots, \underbrace{w^{(n)}(L^{(n)})}_{(t+L^{(n)})\text{-th}}, 0, \ldots, \underbrace{0}_{T\text{-th}}]' \\ &\otimes \boldsymbol{I}_{M \times M}, n = 0, 1, 2 \end{aligned} \quad (13)$$

and $L^{(0)} = 0$, $w^{(0)}(0) = 1$, M is the size of vector $\boldsymbol{c}_t$. Thus, we maximized (3) under the constraint of (9). It has been shown that the generated parameter sequence is realistic as defined by statistics of both static and dynamic features.

Finally, speech is synthesized directly from the generated spectral parameters and $F0$ parameters by the MLSA filter [9].

3. Speech Parameter Generation Based on Decision-Tree Backing-off

In the context clustering, the MDL criterion has been used to stop splitting decision-trees [6]. However, we have observed that the speech quality can be improved by enlarging decision-trees. Meanwhile, overly large decision-trees make the speech quality discontinuous and unstable. Moreover, the optimal tree size varies according to the text to be synthesized. In fact, larger decision-trees achieve a better speech quality for closed sentences.

To determine the cluster size adaptively according to the text to be synthesized, this paper proposes a decision-tree backing-off technique. In the training phase, the decision-tree is grown without any stopping criterion until no questions can split any leaf nodes. Every node of decision-trees holds model parameters corresponding to each cluster in the proposed technique, whereas only leaf nodes hold model parameters in the conventional decision-tree construction. When synthesizing speech, as shown in Fig. 1, one of the nodes on the path from the root node to the leaf nodes corresponding to the requested context-based-label is selected in such a way that the output probability of the generated speech parameter sequence is maximized. The output probability can be written by

$$\max_{\boldsymbol{O}} \log P(\boldsymbol{o} \,|\, \boldsymbol{q_h}, \lambda) = \max_{\boldsymbol{O}} \log P(\boldsymbol{o} \,|\, \boldsymbol{\mu_{q_h}}, \boldsymbol{U_{q_h}}) \quad (14)$$

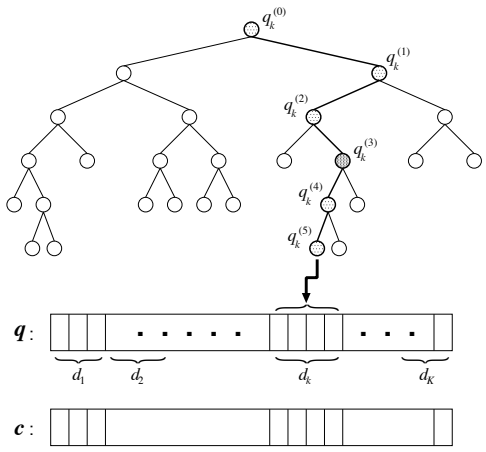


Figure 1: Decision-tree Backing-off.

where

$$\boldsymbol{q}_{\boldsymbol{h}} = \{\overbrace{\underbrace{n_1^{(h_1)}, \ldots, n_1^{(h_1)}}_{d_1}, \ldots, \underbrace{n_K^{(h_K)}, \ldots, n_K^{(h_K)}}_{d_K}}^{T}\}, \quad (15)$$

$$\boldsymbol{h} = \{h_1, h_2, \ldots, h_K\}. \quad (16)$$

It is noted that without the constraint of (9), leaf nodes are always selected. However, with the constraint of (9), appropriate nodes on the paths from leaf nodes to the root node are selected by compromising between the output probabilities of static and dynamic features: at the leaf nodes, the output probabilities of static feature vectors are maximized though they could cause a decrease in the output probabilities of dynamic feature at each state transition.

Unfortunately, there are too many combinations of $h_1, h_2, \ldots, h_K$. Therefore, we constrain $\boldsymbol{h}$ as

$$\boldsymbol{h} = \{\llcorner \alpha H_1 \lrcorner, \llcorner \alpha H_2 \lrcorner, \ldots, \llcorner \alpha H_K \lrcorner\}, \quad 0 \leq \alpha \leq 1 \quad (17)$$

where $\llcorner \cdot \lrcorner$ is the Gauss' symbol and $H_k + 1$ is the number of nodes on each path from the root node to the leaf node. Then, for a given text to be synthesized, we optimize α by a simple one-dimensional optimization algorithm.

4. Experiment

4.1. Experimental Environment

We used phonetically balanced 503 sentences uttered by a male from the ATR Japanese speech database B-set. The 450 sentences were used for training HMMs. Speech signals were sampled at 16kHz and windowed by a 25-ms Blackman window with a 5-ms shift, and then mel-cepstral coefficients were obtained by a mel-cepstral analysis technique. An output vector consists of a spectrum part and $F0$ part. Spectrum part consists of 25 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. Similarly, $F0$ part consists of log $F0$, its delta and delta-delta. We used 5-state left-to-right HMMs with no skip for spectral and $F0$ modeling and 5-dimentional Gaussian' for duration modeling.

This paper applies the backing-off technique only to the spectral part of HMM parameters. To evaluate the speech samples synthesized with and without the backing-off technique, state durations were determined from the phonetic alignments on the natural speech calculated with the Viterbi algorithm.

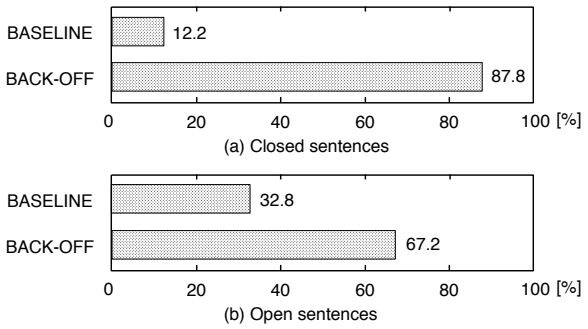


Figure 2: Preference score.

To evaluate the technique using the MDL criterion to stop decision-tree splitting and not applying the backing-off technique (BASELINE) and the proposed technique not using any criterion to stop decision-tree splitting and applying the backing-off technique (BACK-OFF), preference tests were conducted with nine subjects. In the subjective test, the following two sets of sentences were tested separately:

(a) 20 sentences were chosen randomly from 50 sentences included in the training data (closed sentences).

(b) 20 sentences were chosen randomly from 53 sentences not included in the training data (open sentences).

In decision-tree backing-off, huge size decision-trees are conducted by not using any criterion to stop decision-tree splitting. Generally, likelihood of the training data become larger so that decision-trees are enlarged. This is because the output probability is maxmized by using the leaf nodes or the nodes close to the leaf nodes for synthesizing the speech for texts included in the training data. In order to confirm this point, the backing-off technique is applied also to synthesizing speech for closed sentences.

4.2. Experimental Result

Fig. 2 shows the experimental results. From this figure, the decision-tree backing-off technique improves the speech quality for both closed and open sentences. This is also comfirmed from generated running spectra shown in Fig. 3, 4: the structure of the spectral formant gets clear especially in the closed sentence even though the sequence of the generated spectra is varying continuously.

Fig. 5 shows the location of the selected node in decision-tree backing-off. In addition, "BASELINE" means the nodes which satisfy the MDL criterion in trees constructed by the proposed technique. From Fig. 5, it can be seen that the leaf nodes themselves or the nodes very close to leaf nodes are selected for the closed sentence while nodes closer "BASELINE" nodes were selected for the open sentence. From these observations, we see that appropriate nodes were selected automatically for each text to be synthesized by the decision-tree backing-off technique.

5. Conclusion

In this paper, we propose a desicion-tree backing-off technique for an HMM-based speech synthesis system. The proposed technique enables us to adjust the cluster size dynamically at

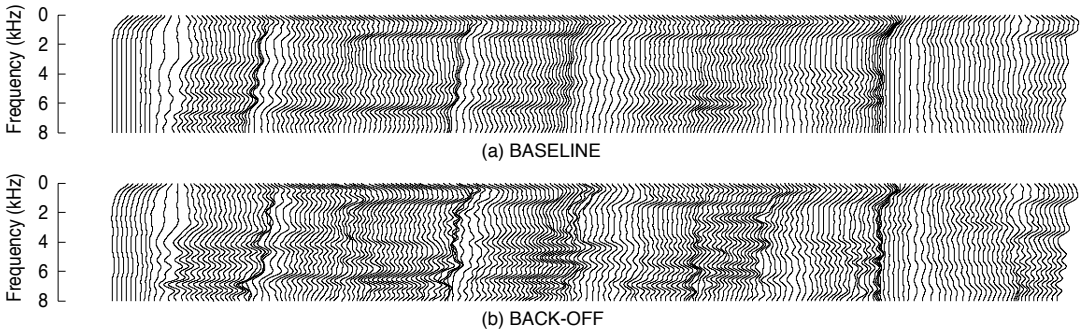


Figure 3: Generated running spectra for closed sentence: "/a-r-a-y-u-r-u-g-e-n-j-i-ts-u-o-· · · /"

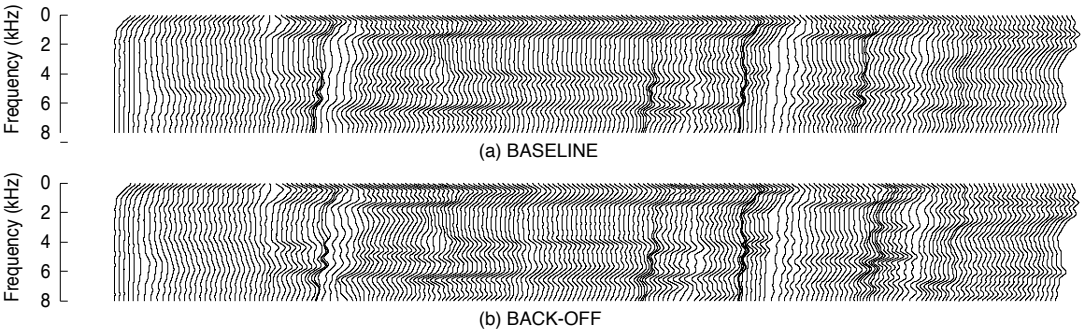
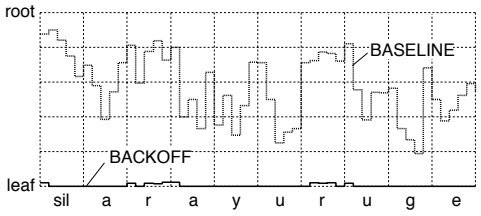
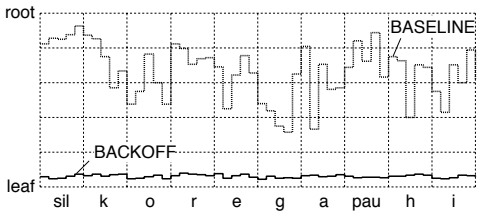


Figure 4: Generated running spectra for open sentence: "/h-i-r-o-i-i-m-i-d-e-n-o-· · · /"



(a) Closed sentence "/a-r-a-y-u-r-u-g-e-N-j-i-· · · /"



(b) Open sentence "/k-o-r-e-g-a-h-i-r-o-i-· · · /"

Figure 5: Node location in decision-tree backing-off.

run-time according to the text to be synthesized. The result of subjective tests show that the proposed technique improves the quality of synthesized speech.

In this paper, we applyed the backing-off technique only for the spectral models but not for the $F0$ and duration models. Further improvements of the quality of synthesized speech is expected by applying the proposed technique to these models. In addition, the experiment and evaluation for synthesizing speech of the limited-domain are also mentioned as the future works since it is expected to be effective in limited domain speech synthesis.

6. References

[1] T. Yoshimura, K. Tokuda, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis", *Proc. of EUROSPEECH*, vol. 5, pp.2347–2350, Sep. 1999.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithm for HMM-based Speech Synthesis, " *Proc. of ICASSP*, vol. 3, pp.1315–1518, 2000.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of Pitch and Spectrum for HMM-based Speech Synthesis Using MLLR, " *Proc. of ICASSP* pp.208–808, May. 2001.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker Interpolation in HMM-based Speech Synthesis System, " *Proc. of EUROSPEECH*, vol.5, pp.2523–2526, Sep. 1997.

[5] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoice for HMM-based Speech Synthesis," in *Proc of ICSLP*, vol. 1, pp. 1269–1272, Sep. 2002.

[6] K. Shinoda and T. Watanabe, "Acoustic Modeling Based on the MDL Principle for speech recognition," *Proc of Eurospech*, Rhodes, vol. 1, pp. 99–102, 1997.

[7] R. E. Donovan and E. M. Eide, "The IBM Trainable Speech Synthesis System," *Proc. of ICSLP'98, Sydony* pp. 1703–1706.

[8] T. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," *Proc of ICASSP*, pp.229–232, Mar. 1999.

[9] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. of ICASSP*, vol. 1, pp. 137–140, 1992.