Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems

Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan

{uratec,zen,nankaku,ri,tokuda}@sp.nitech.ac.jp

Abstract

This paper proposes a technique of reducing footprint of HMM-based speech synthesis systems by tying all covariance matrices. HMM-based speech synthesis systems usually consume smaller footprint than unit-selection synthesis systems because statistics rather than speech waveforms are stored. However, further reduction is essential to put them on embedded devices which have very small memory. According to the empirical knowledge that covariance matrices have smaller impact for the quality of synthesized speech than mean vectors, here we propose a clustering technique of mean vectors while tying all covariance matrices. Subjective listening test results show that the proposed technique can shrink the footprint of an HMM-based speech synthesis system while retaining the quality of synthesized speech.

Index Terms: HMM, speech synthesis, decision tree, context-clustering, MDL criterion, embedded device

1. Introduction

Currently the most popular speech synthesis technique is unit selection synthesis [1–3], where appropriate sub-word units are selected from large speech databases. Although this technique can synthesize high quality synthesized speech, we need to record large speech databases. Furthermore, this system usually requires too large footprint to put it on embedded devices such as mobile phones, PDAs, car navigation systems, and game machines.

A statistical parametric speech synthesis system based on HMMs [4] has grown in popularity in recent years. Figure 1 illustrates the overview of a typical HMM-based speech synthesis system. In this system, the spectrum, excitation, and duration of speech are modeled simultaneously by context-dependent HMMs, and speech parameter trajectories are generated from the HMMs themselves under constraints between static and dynamic features. One of the attractive points of HMM-based speech synthesis is its footprint. The HMM-based system usually has smaller footprint than the unit selection system, because statistics rather than speech waveforms are required to be stored. However, further reduction is essential to put it on embedded devices which have very small memory.

Speech parameters such as spectrum, excitation, and durations depend on a variety of contextual factors such as phoneme identities, accent types, and part-of-speech. In the HMM-based speech synthesis system, context-dependent models are used to capture these contextual factors. If more combinations of the above contextual factors are taken into account, we should be able to obtain more accurate models. However, as the number of contextual factors increases, the number of possible combinations also increases exponentially. As a result, it is difficult to robustly estimate model parameters because of lack of training data. Furthermore, it is impossible to cover every possible combination of contextual factors with a finite set of training data. Although a variety of parameter tying techniques have been developed [5–9] to avoid this problem, a decision tree-based context-clustering technique [10] has been widely used. In the HMM-based speech synthesis system, distributions of spectrum, excitation, and duration are clustered individually because they have their own contextual dependency.

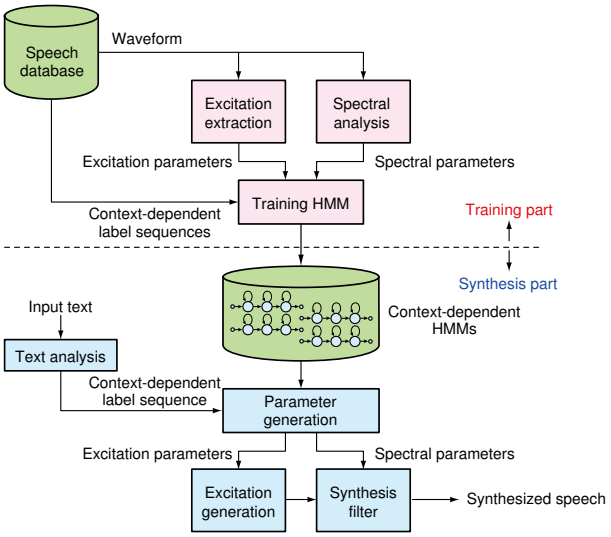

Figure 1: Overview of HMM-based speech synthesis system.

In this technique, a top-down clustering is performed so as to maximize the likelihood of model to the training data by using questions about contexts. Then, HMM states (or streams) which are clustered into the same leaf node are tied. Unseen models can be generated by traversing the decision trees. Various criteria to select questions to be used and nodes to be split [11–13] and techniques to extend single Gaussian distribution to mixture of Gaussian distributions [14, 15] have been proposed.

Conventionally we construct an HMM stream-level tying structure in HMM-based speech synthesis, i.e., mean vectors and covariance matrices have exactly the same parameter tying structure (Fig. 2 (a)). However, we empirically know that covariance matrices have smaller impact for the quality of synthesized speech than mean vectors. Based on this knowledge, this paper proposes a context-clustering technique of mean vectors while tying all covariance matrices (Fig. 2 (b)). If each parameter is stored in single-precision floating-point number (4 Bytes) and the dimensionality of Gaussian distributions is 120, approx-

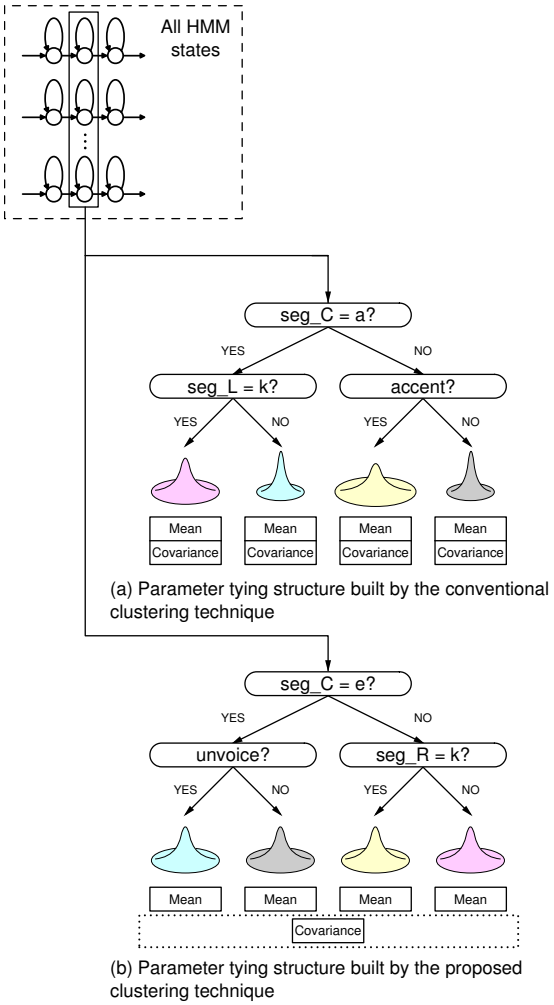

Figure 2: Context-dependent parameter tying structure built by conventional and proposed clustering techniques.

imately 938 KBytes are required to store 1,000 Gaussian with diagonal covariance matrices distributions (statistics associated to the leaf nodes). However, by tying all covariance matrices, it reduced to almost half (469 KBytes).

The rest for this paper is organized as follows. Section 2 describes the proposed decision tree-based context-clustering technique of mean vectors while tying all covariance matrices. Subjective listening test results are shown in Section 3. Concluding remarks and future plans are presented in Section 4.

2. Tying covariance matrices

2.1. Decision tree-based context clustering

In the decision tree-based context-clustering technique, a top-down clustering is performed so as to locally maximize the likelihood of model to the training data using pre-defined questions about contexts. Then, mean vectors and covariance matrices of HMM states (or streams) clustered to the same leaf (terminal) node are tied. As a result, HMM state-level (or stream-level) tying structure can be constructed. The mean vector and the covariance matrix associated to the leaf node S, $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$, can be estimated based on the ML criterion as

$$\hat{\boldsymbol{\mu}}_S = \frac{\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\,\boldsymbol{o}_t}{\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)}, \quad (1)$$

$$\hat{\boldsymbol{\Sigma}}_S = \frac{\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\,(\boldsymbol{o}_t-\hat{\boldsymbol{\mu}}_S)(\boldsymbol{o}_t-\hat{\boldsymbol{\mu}}_S)^\top}{\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)}, \quad (2)$$

where T is the total number of frames in the training data, $\boldsymbol{M}_S$ is a set of HMM states (or streams) clustered to the leaf node S, and $\gamma_m(t)$ is the posterior probability of an HMM state (or stream) m for an observation vector at frame t, $\boldsymbol{o}_t$. The total log likelihood of the Gaussian distribution of node S to the associated training data is calculated as

$$\begin{aligned}\mathcal{L}(S) &= \sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\log\mathcal{N}\left(\boldsymbol{o}_t;\hat{\boldsymbol{\mu}}_S,\hat{\boldsymbol{\Sigma}}_S\right)\\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\left\{n+\log\left(2\pi\left|\hat{\boldsymbol{\Sigma}}_S\right|\right)\right\},\end{aligned} \quad (3)$$

where n is the dimensionality of $\boldsymbol{\mu}_S$.

The minimum description length (MDL) criterion [11] has been used in the HMM-based speech synthesis system to automatically control the size of decision trees. When cluster S is divided to S_{q+} and S_{q-} by a question q, the change of total description length by this split is calculated as follows:

$$\Delta_q = \mathcal{L}(S) - \left\{\mathcal{L}(S_{q+})+\mathcal{L}(S_{q-})\right\} + \alpha\frac{N}{2}\log\Gamma(S_0), \quad (4)$$

where S_0 denotes a root node, α is a heuristic weight for the penalty term of the MDL criterion, N is the number of parameters increased by this split, and

$$\Gamma(S) = \sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t). \quad (5)$$

If all covariance matrices are diagonal covariance matrices, $N = n+n$. Note that the context-clustering based on the MDL criterion can be viewed as that based on the ML criterion whose threshold is given as $\alpha\frac{N}{2}\log\Gamma(S_0)$.

2.2. Context clustering while tying all covariance matrices

The decision tree-based context-clustering techniques used in HMM-based speech synthesis system construct HMM state- or stream-level tying structure, i.e., the same tying structure is used for both mean vectors and covariance matrices. However, we empirically know that mean vectors have more impact for the quality of synthesized speech than covariance matrices. For example, even we manually modify values of covariance matrices, speech parameter trajectories generated from the original and modified models are almost identical. In this paper, we construct tying structure of mean vectors by using decision trees while tying all covariance matrices.

If all covariance matrices are tied, the total log likelihood of the leaf node S to the associated training data is calculated as follows:

$$\begin{aligned}\mathcal{L}'(S) &= \sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\log\mathcal{N}(\boldsymbol{o}_t;\hat{\boldsymbol{\mu}}_S,\boldsymbol{\Sigma}_g)\\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\Big\{\mathrm{Tr}\left(\hat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_g^{-1}\right)\\ &\quad+\log\left(2\pi\left|\boldsymbol{\Sigma}_g\right|\right)\Big\},\end{aligned}\tag{6}$$

where $\boldsymbol{\Sigma}_g$ is a globally tied covariance matrix. Note that $\boldsymbol{\Sigma}_g$ is fixed in the context-clustering process because of large computational cost.

When cluster S is divided to S_{q+} and S_{q-} by a question q, the change of total description length by this split is calculated as follows:

$$\Delta'_q = \mathcal{L}'(S) - \left\{\mathcal{L}'(S_{q+}) + \mathcal{L}'(S_{q-})\right\} + \alpha\frac{N}{2}\log\Gamma(S_0)\,.\tag{7}$$

Unlike Eq. (4), N becomes n in this case because only mean vectors are split. We can expect that the proposed technique can efficiently reduce the footprint of HMM-based speech synthesis systems while retaining the quality of synthesized speech.

3. Experiments

3.1. Experimental condition

To evaluate the effectiveness of the proposed technique, subjective listening tests were conducted. The first 450 sentences of the phonetically balanced 503 sentences from the ATR Japanese speech database B-set [16], uttered by male speaker MHT, were used for training. The remaining 53 sentences were used for evaluation. Speech signals were sampled at 16kHz and windowed with a 5-ms shift, and mel-cepstral coefficients [17] were obtained from STRAIGHT spectrum [18]. Feature vectors consisted of spectrum and excitation parameters. The spectrum parameter vectors consisted of 39 STRAIGHT mel-cepstral coefficients including the zero coefficient, their delta and delta-delta coefficients. The excitation parameter vectors consisted of log F_0, its delta and delta-delta. A seven-sate (including the beginning and ending null states), left-to-right, no skip structure was used for hidden semi-Markov model [19]. The spectrum stream was modeled by single multi-variate Gaussian distributions. The excitation stream was modeled by a multi-space probability distribution consisting of a Gaussian distribution for voiced frames and discrete distribution for unvoiced frames. Each state-duration distribution was modeled by a five-dimensional (equal to the number of emitting states in each phoneme model) multi-variate Gaussian distribution. The decision tree-based context-clustering technique was separately applied to distributions for spectrum, excitation, and state duration. A speech parameter generation algorithm considering global variance (GV) [20] was used for parameter generation.

The MDL criterion [11] was used to control the size of decision trees. We changed the heuristic weight for the penalty term of α to construct acoustic models with various number of parameters. The weights used here were 8.0, 4.0, 2.0, 1.0, 0.5, and 0.25. Although the decision tree-based context-clustering technique was separately applied to distributions for spectrum and excitation, the same α was used.

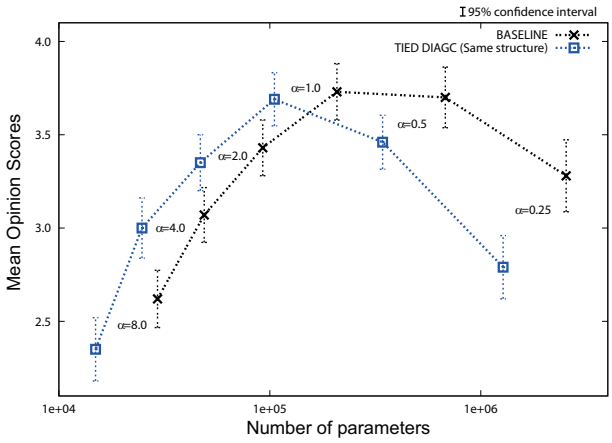

Figure 3: Experimental results: Type of covariance tying structure. The same mean tying structures are constructed.

Ten subjects participated in these listening tests. Ten sentences were randomly selected from 53 sentences for each subject. The subjects were asked to rate the naturalness of synthesized speech with a scale from 1 (completely unnatural) to 5 (natural). All experiments were carried out in a sound-proof room using head-phones.

3.2. Experimental results

The first listening test was designed to confirm the empirical knowledge that covariance matrices have small impact for the quality of synthesized speech. The following two methods were evaluated;

BASELINE : The same tying structure was used for both mean vectors and covariance matrices.

TIED DIAGC (Same structure) : Although the tying structure of mean vectors was exactly the same as **BASELINE**, all covariance matrices were tied.

Figure 3 shows the subjective listening test results. It can be seen from the figure that **TIED DIAGC (Same structure)** achieved almost the same subjective scores with almost the half number of parameters (footprint) when $\alpha = 1.0$. It also shows that tying covariance matrices looks more efficient than reducing the size of decision trees to achieve the same footprint.

The second listening test evaluated the performance of the proposed clustering technique while tying all covariance matrices. The following two methods were compared;

BASELINE : The same tying structure was used for both mean vectors and covariance matrices.

TIED DIAGC (Proposed) : Mean vectors were clustered by decision trees while tying all covariance matrices.

Figure 4 shows the experimental results. It can be seen from the figure that **TIEDGC (Proposed)** significantly reduced the number of parameters. Furthermore, it achieved slightly better subjective scores. On the other hand, only **TIED DIAGC** system with $\alpha = 0.25$ reduced scores. Although the number of parameters decreased by using the proposed technique, the number of mean parameters increased. Therefore, it seems that

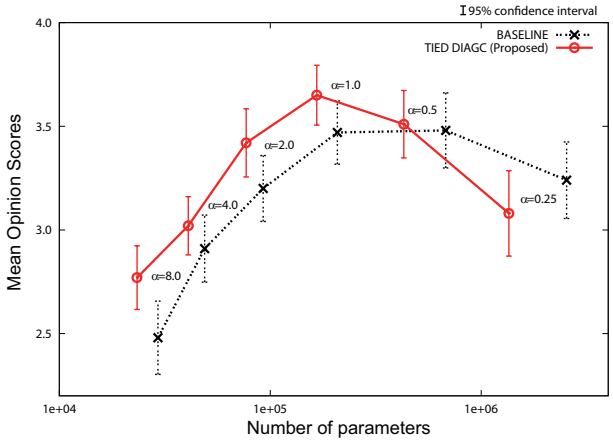

Figure 4: Experimental results: Type of covariance tying structure. Different mean tying structures are constructed.

training data was lacking. When each parameter is stored in single-precision floating-point number (4 Bytes), the footprint of **BASELINE** with $\alpha = 1.0$ is about 813 KBytes. However, **TIED DIAGC (Proposed)** with $\alpha = 1.0$ requires 649KBytes. Furthermore, **TIED DIAGC (Proposed)** with $\alpha = 2.0$ consumes only 300KBytes while retaining the quality of synthesized speech close to **BASELINE** with $\alpha = 1.0$.

Table 1 shows the number of leaf nodes of each system with $\alpha = 1.0$. By using the proposed technique, the number of mean parameters increased. It seems that degradation of quality of synthesized speech by tying covariance matrices was reduced by incrementation of the number of mean parameters.

4. Conclusions

This paper proposed a technique of reducing the footprint of HMM-based speech synthesis systems by tying all covariance matrices. The experimental results showed that the proposed technique efficiently shrinked the footprint of an HMM-based speech synthesis system to less than half of its original size while retaining the quality of synthesized speech. Future work includes applying this technique to full covariance matrices and comparing it with semi-tied covariance matrices [21].

5. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org).

Table 1: Comparison of number of leaf nodes.

	Number of leaf nodes			
	Spectrum		F0	
	Mean	Covariance	Mean	Covariance
BASELINE	808	808	2015	2015
TIED DIAGC (Proposed)	1311	1	2210	1

6. References

[1] A. W. Black and P. Taylor, "CHATR: a generic speech synthesis system," in Proc. COLING94, 1994.

[2] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in Proc. ICASSP, 1996, pp. 373–376.

[3] R. E. Donovan and P. C. Woodland, "Automatic speech synthesizer parameter estimation using HMMs," in Proc. ICASSP, 1995, pp. 640–643.

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

[5] K. F. Lee, "Context-Dependent Phonetic Hidden Markov models for Speaker-Independent Continuous Speech Recognition," IEEE Trans. Acoustic Speech and Signal Processing, vol. 38, no. 4, pp. 599–609, 1990.

[6] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. ICASSP'92, pp. 573–576, 1992.

[7] P. Woodland and S. Young, "Benchmark DARPA RM results with the HTK portable HMM toolkit," Proc. DARPA Continuous Speech Recognition Workshop, pp. 71–76, 1992.

[8] M. Y. Hwang, X. Huang, and F. Alleva, "Predicting Unseen Triphones with Senones," Proc. ICASSP'93, pp .311–314, 1993.

[9] M. Ostendorf and H .Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech Language, vol. 1, no. 1, pp. 17–41, 1997.

[10] J. J. Odell, "The Use of Context in Large Vocabulary Speech Recognition, PhD dissertation," Cambridge University, 1995.

[11] K. Shinoda and T. Watanabe, "MDL-based Context-Dependent Subword Modeling for Speech Recognition," J. Acoust. Soc. Jpn.(E), vol .21, no. 2, pp. 79-8-6, 2000.

[12] W. Chou and W. Reichl, "Decision Tree State Tying based on Penalized Bayesain Information Criterion," Proc. ICASSP'99, pp. 345–348, 1999.

[13] S. Watanabe, Y. Minami, A. Nakamura, and N.Ueda, "Training of Shared States in Hidden Markov Model Based on Bayesian Approach" IEICE technical report. Speech, vol. 102, no. 35, pp. 43–48, 2002.

[14] T. Kato, S. Kuroiwa, T. Shimizu, and N. Higuchi, "Tree-based Clustering for Gaussian Mixture HMMs," IEICE Trans. vol. J83-D-II, no. 11, pp. 2128–2136, 2000.

[15] H. J. Nock, "Context Clustering for Triphone-based Speech Recognition," Master Thesis, Cambridge University, 1996.

[16] A. Kuramatsu, K. Takeda, Y. Sagisaka, S. Katagiri, H .Kawabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, pp. 357–363, 1990.

[17] K. Tokuda, T. Kobayashi, T. Chiba, and S. Imai, "Spectral Estimation of Speech by Mel-Generalized Cepstral Analysis," IEICE Trans. vol. 75-A, no. 7, pp. 1124–1134, 1992.

[18] H. Kawahara, M. K. Ikuyo, A. Cheneigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27, pp. 187–207, 1999.

[19] H. Zen, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura, "A Hidden Semi-Markov Model-Based Speech Synthesis System," IEICE Trans. Inf. & Sys., vol. 90D, no. 5, pp. 825–834, 2007.

[20] T. Toda, K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," Interspeech 2005, pp. 2801–2804, 2005.

[21] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 272–281, 1999.