PAPER

A Covariance-Tying Technique for HMM-Based Speech Synthesis

Keiichiro OURA[†a)], **Heiga ZEN**[†], ***Nonmembers***, **Yoshihiko NANKAKU**[†], **Akinobu LEE**[†], ***and* Keiichi TOKUDA**[†], ***Members***

SUMMARY A technique for reducing the footprints of HMM-based speech synthesis systems by tying all covariance matrices of state distributions is described. HMM-based speech synthesis systems usually leave smaller footprints than unit-selection synthesis systems because they store statistics rather than speech waveforms. However, further reduction is essential to put them on embedded devices, which have limited memory. In accordance with the empirical knowledge that covariance matrices have a smaller impact on the quality of synthesized speech than mean vectors, we propose a technique for clustering mean vectors while tying all covariance matrices. Subjective listening test results showed that the proposed technique can shrink the footprints of an HMM-based speech synthesis system while retaining the quality of the synthesized speech.

key words: *HMM, speech synthesis, decision tree, context-clustering, MDL criterion, embedded device*

1. Introduction

The most widely used speech synthesis technique is unit selection synthesis [1]–[3], in which appropriate sub-word units are selected from large speech databases. Although this technique can synthesize high-quality speech, it requires large databases of recorded speech. Furthermore, it usually requires excessively large footprints to put it on embedded devices such as mobile phones, PDAs, car navigation systems, and game machines.

Statistical parametric speech synthesis based on HMMs [4], [5] has grown in usage. Figure 1 gives an 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 [6]. One of the attractive points of HMM-based speech synthesis is its small footprint. HMM-based systems usually have smaller footprints than unit selection systems, because they store statistics rather than speech waveforms. However, further reduction is essential to put these systems on embedded devices that have little memory.

Speech parameters such as spectrum, excitation, and duration depend on a variety of contextual factors such as phoneme identities, accent types, and parts-of-speech. In the HMM-based speech synthesis system, context-dependent models are used to capture these contextual factors. If more combinations of these 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 due to the lack of training data. Furthermore, it is impossible for a finite set of training data to cover every possible combination of contextual factors. Various parameter-tying techniques have been developed [7]–[11] to avoid this problem. Among them, a decision tree-based context-clustering technique [12] 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 dependencies.

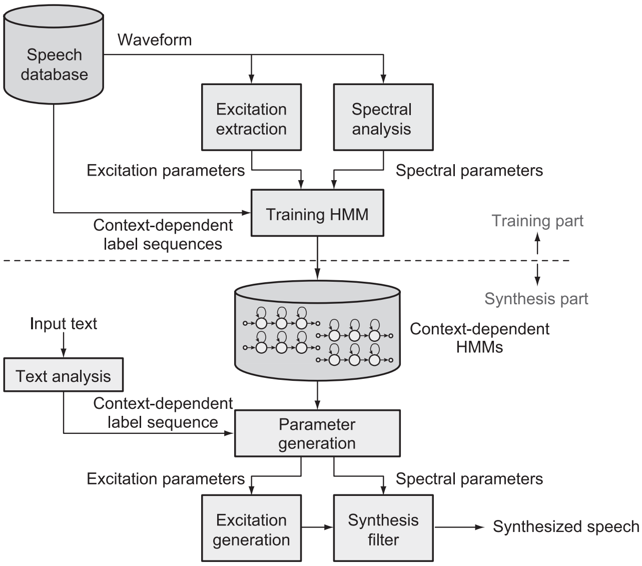


Fig. 1 Overview of HMM-based speech synthesis system.

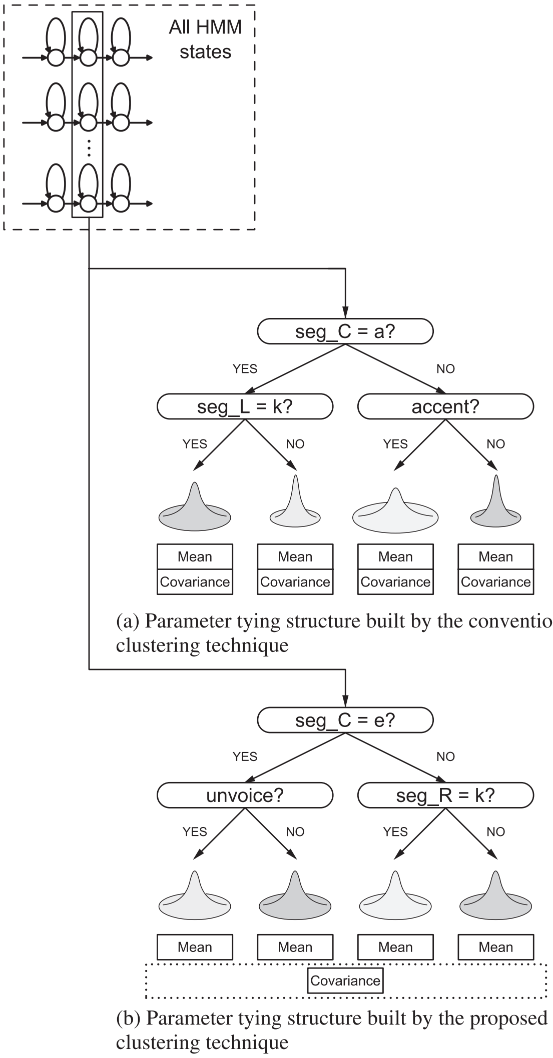
In this technique, top-down clustering is performed to maximize the likelihood of model parameters with respect to the training data by using questions about contexts. Then, HMM identifies which of those clustered into the same leaf node are tied. Unseen models can be generated by traversing the decision trees. Various criteria [13]–[17] have been proposed for selecting the questions to be used.

Conventionally, we construct an HMM stream-level tying structure in HMM-based speech synthesis, i.e., mean

Manuscript received July 15, 2009.
Manuscript revised November 2, 2009.

[†]The authors are with the Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.
a) E-mail: uratec@sp.nitech.ac.jp

DOI: 10.1587/transinf.E93.D.595



(a) Parameter tying structure built by the conventional clustering technique

(b) Parameter tying structure built by the proposed clustering technique

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

vectors and covariance matrices have exactly the same parameter-tying structure (Fig. 2 (a)). However, we empirically know that covariance matrices have a smaller impact on the quality of synthesized speech than mean vectors. On the basis of this knowledge, a technique for context-clustering mean vectors while tying all covariance matrices (Fig. 2 (b)) should be tested in HMM-based speech synthesis. If each parameter is stored in a single-precision floating-point number and the dimensionality of Gaussian distributions is 120, approximately 938 kBytes are required to store 1,000 Gaussian distributions with diagonal covariance matrices (statistics associated to the leaf nodes). However, tying all covariance matrices reduced it almost by half (469 kBytes).

Semi-tied covariance [18] is one of the major covariance-tying techniques. This technique is a simple extension of the standard diagonal or full covariance matrices used with HMMs. Instead of having a distinct covariance matrix for every distribution, each covariance matrix consists of two elements, a component-specific diagonal covariance element and a tied transform. It is important to make the difference between the semi-tied covariance technique and the proposed technique clear.

The rest of this paper is organized as follows. Section 2 describes the decision-tree-based context clustering technique. Context clustering for semi-tied covariance matrices are presented in Sect. 3. Section 4 describes the proposed decision tree-based context-clustering technique for mean vectors while tying all covariance matrices. Subjective listening test results are shown in Sect. 5. Finally, concluding remarks and future plans are presented in Sect. 6.

2. Decision Tree-Based Context Clustering

In the decision-tree-based context-clustering technique, top-down clustering is performed to locally maximize the likelihood of model parameters with respect to the training data using pre-defined questions about contexts. Then, mean vectors and covariance matrices of HMM states clustered to the same leaf (terminal) node are tied. As a result, an HMM state-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 using the ML criterion as

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

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

where T is the total number of frames in the training data, $\boldsymbol{M}_S$ is a set of HMM states clustered to the leaf node S, and $\gamma_m(t)$ is the posterior probability of an HMM state 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

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

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

The minimum description length (MDL) criterion [13] has been used in the HMM-based speech synthesis system to automatically control the size of decision trees. When cluster S is divided into S_{q+} and S_{q-} by a question q, the change of total description length by this split is calculated as follows:

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

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

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

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 with a threshold given by $\alpha\frac{N}{2}\log\Gamma(S_0)$.

3. Context Clustering for Semi-Tied Covariance Matrices

In this section, we describe the semi-tied covariance technique since it will be evaluated as a conventional method in Sect. 5. In the HMM-based speech synthesis system, there is a choice of the form of the covariance matrices. When diagonal covariance matrices are used, elements of the feature vector are assumed to be independent. On the other hand, when full covariance matrices are used, all correlations are explicitly modeled. However, when full covariance matrices are used, the number of parameters per Gaussian component increases exponentially. Compared with a diagonal covariance matrix, the number of parameters per distribution increases to $n + \frac{n(n+1)}{2}$ from $n + n$. Furthermore, the number of training samples per distribution decreases. Due to this massive increase in the number of parameters, diagonal covariance matrices are generally used in HMM-based speech synthesis. Using the semi-tied covariance matrix is a good solution to this problem.

Semi-tied covariance matrices are a simple extension of the standard diagonal or full covariance matrices used with HMMs. Instead of having a distinct covariance matrix for every distribution, each covariance matrix consists of two elements, a component-specific diagonal covariance element, $\boldsymbol{\Sigma}^{(\mathrm{diag})}$, and a tied transform, $\mathbf{H}$. The form of the covariance matrix of state S is defined as

$$\boldsymbol{\Sigma}_S^{(\mathrm{stc})} = \mathbf{H}\boldsymbol{\Sigma}_S^{(\mathrm{diag})}\mathbf{H}^\top. \tag{6}$$

The number of parameters increased by a split, N, becomes $n + n$.

4. Context Clustering while Tying All Covariance Matrices

The decision-tree-based context-clustering techniques used in HMM-based speech synthesis systems construct an HMM state-level tying structure, i.e., the same tying structure is used for both mean vectors and covariance matrices. However, covariance matrices have less impact on the quality of synthesized speech than mean vectors. For example, even if we manually modify values of covariance matrices, the speech parameter trajectories generated from the original and modified models are often close to each other. In this paper, we construct the tying structure of mean vectors 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:

$$\begin{aligned}\mathcal{L}'(S) &= \sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\log\mathcal{N}\left(\boldsymbol{o}_t;\boldsymbol{\mu}_S,\boldsymbol{\Sigma}_g\right)\\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\left(\boldsymbol{o}_t-\boldsymbol{\mu}_S\right)^\top\Sigma_g^{-1}\left(\boldsymbol{o}_t-\boldsymbol{\mu}_S\right)\\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\log\left(2\pi\left|\boldsymbol{\Sigma}_g\right|\right)\\ &= -\frac{1}{2}\mathrm{Tr}\left\{\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\left(\boldsymbol{o}_t-\boldsymbol{\mu}_S\right)\left(\boldsymbol{o}_t-\boldsymbol{\mu}_S\right)^\top\Sigma_g^{-1}\right\}\\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\log\left(2\pi\left|\boldsymbol{\Sigma}_g\right|\right)\\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in\boldsymbol{M}_S}\gamma_m(t)\left\{\mathrm{Tr}\left(\boldsymbol{\Sigma}_S\boldsymbol{\Sigma}_g^{-1}\right)+\log\left(2\pi\left|\boldsymbol{\Sigma}_g\right|\right)\right\},\end{aligned} \tag{7}$$

where $\boldsymbol{\Sigma}_g$ is a globally tied covariance matrix, and $\boldsymbol{\Sigma}_S$ is defined in Eq. (2). Note that $\boldsymbol{\Sigma}_g$ is fixed in the context-clustering process because the computational cost is large.

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

$$\Delta'_q = \mathcal{L}'(S) - \left\{\mathcal{L}'\left(S_{q+}\right) + \mathcal{L}'\left(S_{q-}\right)\right\} + \alpha\frac{N}{2}\log\Gamma(S_0). \tag{8}$$

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

5. Experiments

5.1 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 [19], uttered by male speaker MHT, were used for training. The remaining 53 sentences were used for evaluation. Speech signals were sampled at 16 kHz and windowed with a 5-ms shift, and mel-cepstral coefficients [20] were obtained from STRAIGHT

[†] The standard value of α is unity in the MDL criterion.

spectra [21]. Feature vectors consisted of spectrum and excitation parameters. The spectrum parameter vectors consisted of 39 STRAIGHT mel-cepstral coefficients including the zero coefficient and their delta and delta-delta coefficients. The excitation parameter vectors consisted of log F_0 and its delta and delta-delta. A seven-state (including the beginning and ending null states), left-to-right, no-skip structure was used for the hidden semi-Markov model [22]. The spectrum stream was modeled by single multi-variate Gaussian distributions. The excitation stream was modeled by multi-space probability distributions [23], each of which consists of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. State durations of each model were modeled by a five-dimensional (equal to the number of emitting states in each 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) [24] was used for parameter generation.

The MDL criterion [13] was used to control the size of decision trees. We changed the heuristic weight for the penalty term of α in Eq. (4) and Eq. (8) to construct acoustic models with various numbers 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.

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 the synthesized speech on a scale from 1 (completely unnatural) to 5 (natural). All experiments were carried out using headphones in a soundproof room.

5.2 Semi-Tied Covariance Technique

To reduce the burden on listeners, the listening tests were split into five parts.

To confirm the effect of the covariance matrix type for naturalness and footprint, we evaluated the semi-tied covariance technique in the first experiment. The following two methods were evaluated.

BASELINE: The same structure tied by conventional context-clustering was used for mean vectors and diagonal covariance matrices.

SEMI-TIED (baseline structure): Although the tying structure of mean vectors was exactly the same as the **BASELINE** system, all covariance matrices were semi-tied.

Figure 3 shows the subjective listening test results. The **SEMI-TIED (baseline structure)** system did not improve compared with **BASELINE** system. Although elements of the feature vector are assumed to be dependent in the semi-tied covariance technique, the assumption seems to have little impact on the quality of the synthesized speech.

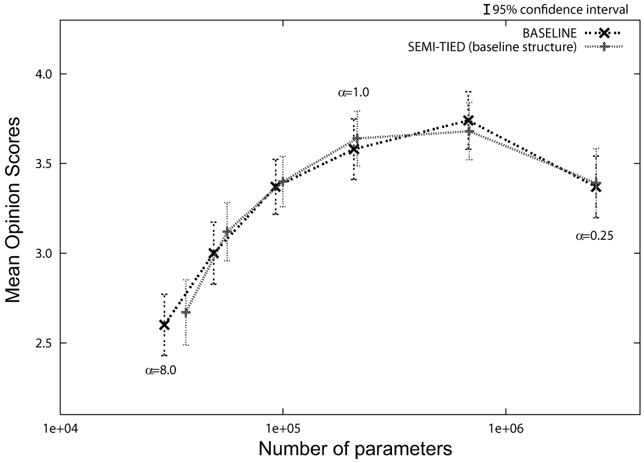


Fig. 3 Subjective experimental results: Conventional method versus semi-tied covariance method. The same mean tying structures are constructed.

5.3 Tied Covariance Technique

Next, a listening test was designed to confirm the empirical knowledge that covariance matrices have little impact on the quality of synthesized speech. The following two methods were evaluated.

BASELINE: The same structure tied by conventional context-clustering was used for mean vectors and diagonal covariance matrices.

PROPOSED0: Although the tying structure of mean vectors was exactly the same as the **BASELINE** system, all full covariance matrices were tied.

Figure 4 shows the subjective listening test results. The **PROPOSED0** system reduced scores slightly compared with the **BASELINE** system. A large amount of memory is required for full covariance matrices, even using the tying technique. Furthermore, covariance matrices without diagonal elements have little impact on the quality of synthesized speech. Therefore, it seems that the use of full covariance matrices is not appropriate for the embedded devices.

Next, we replaced full covariance matrices with diagonal covariance matrices. The following two methods were evaluated.

BASELINE: The same structure tied by conventional context-clustering was used for mean vectors and diagonal covariance matrices.

PROPOSED1: Although the tying structure of mean vectors was exactly the same as the **BASELINE** system, all diagonal covariance matrices were tied.

Figure 5 shows the subjective listening test results. The **PROPOSED1** system achieved almost the same subjective scores with almost half the number of parameters (footprints) when $\alpha = 1.0$. Tying diagonal covariance matrices seems to be more efficient than reducing the size of decision trees to achieve the same footprints.

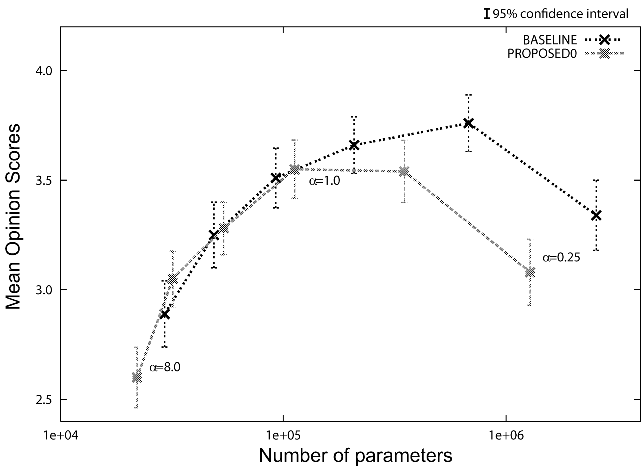


Fig. 4 Subjective experimental results: Conventional method versus tied full covariance method. The same mean tying structures are constructed.

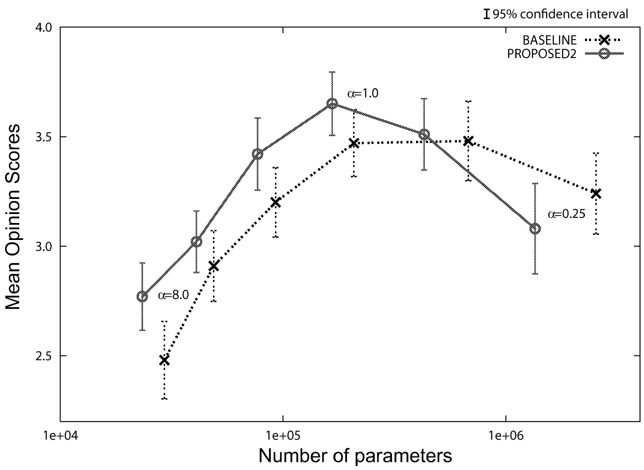


Fig. 6 Subjective experimental results: Conventional method versus tied diagonal covariance method. Different mean tying structures are constructed.

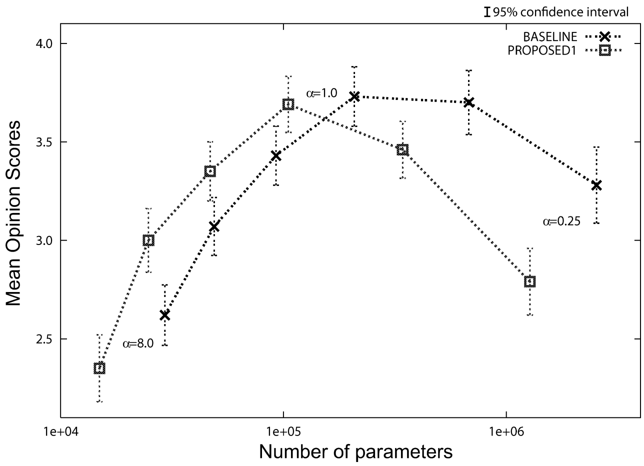


Fig. 5 Subjective experimental results: Conventional method versus tied diagonal covariance method. The same mean tying structures are constructed.

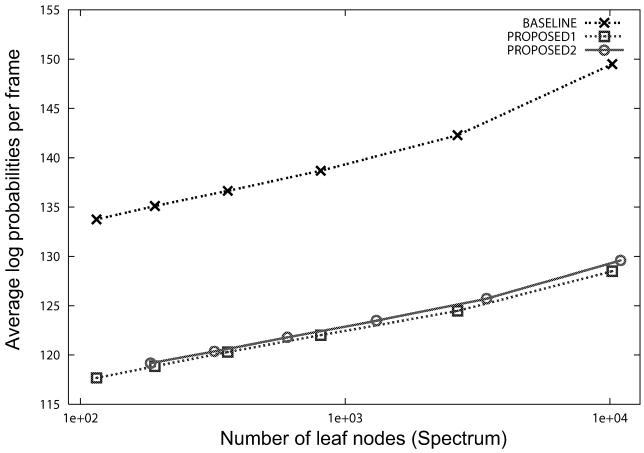


Fig. 7 Objective experimental results: Conventional method versus two proposed methods.

The fourth listening test evaluated the performance of the proposed clustering technique while tying all diagonal covariance matrices. Note that the proposed clustering technique was not applied to full covariance matrices because of its large computational cost. The following two methods were compared.

BASELINE: The same structure tied by the conventional context-clustering was used for mean vectors and diagonal covariance matrices.

PROPOSED2: Mean vectors were clustered by decision trees while tying all diagonal covariance matrices using the technique described in Sect. 4.

Figure 6 shows the experimental results. The **PROPOSED2** system significantly reduced the number of parameters. Furthermore, it achieved slightly better subjective scores than **BASELINE**. When each parameter was stored in a single-precision floating-point number, the footprint of the **BASELINE** system with $\alpha = 1.0$ was about 813 kBytes. On the other hand, that of the **PROPOSED2** system with $\alpha = 1.0$ was 649 kBytes. Furthermore, the **PROPOSED2** system with $\alpha = 2.0$ consumed only 300 kBytes while retaining the quality of synthesized speech close to the **BASELINE** system with $\alpha = 1.0$. Figure 7 shows the average log probabilities per frame of the **BASELINE**, **PROPOSED1**, and **PROPOSED2** systems. In terms of ML estimation of HMM parameters, tying all the covariance matrices decreased the likelihood function. The **PROPOSED2** system had a slightly higher probability than the **PROPOSED1** system because of the proposed context-clustering technique, which constructs appropriate mean vector structures while tying all covariance matrices.

The final listening test evaluated the performance of the two proposed systems compared with the baseline system. To guarantee the generalizability of the proposed method, the first 450 sentences of the phonetically balanced 503 sentences from the ATR Japanese speech database B-set [19], uttered by male speaker MHT and female speaker FKN,

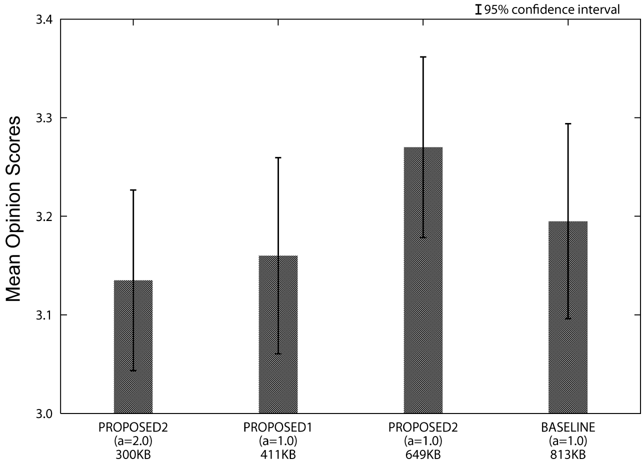


Fig. 8 Subjective experimental results: Conventional method versus two proposed methods. Their footprints were calculated on the assumption that each parameter was stored in a single-precision floating-point number.

Table 1 Comparison of the number of leaf nodes.

	Number of leaf nodes			
	Spectrum		F0	
	Mean	Covariance	Mean	Covariance
BASELINE	808	808	2015	2015
PROPOSED1	808	1	2015	1
PROPOSED2	1311	1	2210	1

were used for training speaker-dependent models. The remaining 2 sets of 53 sentences were used for evaluation. Ten subjects participated in this listening test. Twenty sentences were randomly selected from 106 sentences for each subject. The **BASELINE** system with $\alpha = 1.0$, the **PROPOSED1** system with $\alpha = 1.0$, and the **PROPOSED2** systems with $\alpha = 1.0$ and 2.0 were compared. Figure 8 shows the subjective results. All proposed methods had a smaller footprint than the **BASELINE** system while maintaining the quality of the synthesized speech. The **PROPOSED2** system with $\alpha = 1.0$ had better subjective scores than the **PROPOSED1** system with $\alpha = 1.0$. Table 1 shows the number of leaf nodes of each system with $\alpha = 1.0$. In the **PROPOSED2** system, the number of mean parameters can be increased even when the total number of parameters is decreased. It is supposed that the balance between model complexities of mean parameters and covariance parameters can be adjusted by using the proposed context-clustering technique, which constructs the appropriate mean vector structure while tying all covariance matrices.

6. Conclusion

A technique for reducing the footprints of HMM-based speech synthesis systems by tying all covariance matrices is described. Experimental results showed that the proposed technique efficiently reduced the footprints of an HMM-based speech synthesis system to less than half of its original size while retaining the quality of the synthesized speech. Future work includes using a separated clustering technique for mean vectors and covariance matrices.

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), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

References

[1] A.W. Black and P. Taylor, "CHATR: A generic speech synthesis system," Proc. COLING94, pp.983–986, 1994.

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

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

[4] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. ICASSP 1996, pp.389–392, 1996.

[5] 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 1999, pp.2347–2350, 1999.

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

[7] K.F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Trans. Acoust. Speech Signal Process., vol.38, no.4, pp.599–609, 1990.

[8] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," Proc. ICASSP'92, pp.573–576, 1992.

[9] 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.

[10] M.Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," Proc. ICASSP 1993, pp.311–314, 1993.

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

[12] J.J. Odell, "The use of context in large vocabulary speech recognition," PhD dissertation, Cambridge University, 1995.

[13] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, 2000.

[14] W. Chou and W. Reichl, "Decision tree state tying based on penalized bayesain information criterion," Proc. ICASSP'99, pp.345–348, 1999.

[15] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Training of shared states in hidden Markov model based on bayesian approach," IEICE Technical Report, SP2002-14, 2002.

[16] T. Kato, S. Kuroiwa, T. Shimizu, and N. Higuchi, "Tree-based clustering for gaussian mixture HMMs," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J83-D-II, no.11, pp.2128–2136, Nov. 2000.

[17] H.J. Nock, "Context clustering for triphone-based speech recognition," Master Thesis, Cambridge University, 1996.

[18] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE Trans. Speech Audio Process., vol.7, no.3, pp.272–281, 1999.

[19] 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 Commun., vol.9, pp.357–363, 1990.

[20] K. Tokuda, T. Kobayashi, T. Chiba, and S. Imai, "Spectral estimation of speech by Mel-Generalized cepstral analysis," IEICE Trans. Fundamentals (Japanese Edition), vol.J75-A, no.7, pp.1124–1134, July 1992.

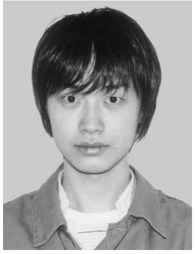
[21] H. Kawahara, M.K. Ikuyo, and A. Cheneigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., vol.27, pp.187–207, 1999.

[22] H. Zen, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.825–834, May 2007.

[23] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP 1999, vol.1, pp.229–232, 1999.

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

Keiichiro Oura was born in 1982. He receieved the B.E., and M.E. degrees in computer science, and computer science and Engineering from the Nagoya Institute of technology, Nagoya, Japan in 2005, and 2007, respectively. He was an intern researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan from September 2007 to December 2007. From April to May 2009, he was an intern/co-op researcher in the Centre of Speech Technology Research, Edinburgh, UK. He is currently a Doctor's candidate at the Nagoya Institute of technology. His research interests include statistical speech recognition and synthesis. He received the best student paper award at ISCSLP in 2008. He is a student member of the Acoustical Society of Japan.



Heiga Zen was born in Osaka, Japan, on March 4, 1979. He received the A.E. degree in electronic and information engineering from Suzuka National College of Technology, Suzuka, Japan, in 1999, and the B.E., M.E., and Dr.Eng. degrees in computer science, electrical and computer engineering, and computer science and engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2001, 2003, and 2006, respectively. During 2003, he was an intern researcher at ATR Spoken Language Translation Research Laboratories (ATR-SLT), Kyoto, Japan. From June 2004 to May 2005, he was an intern/co-op researcher in the Human Language Technology group at IBM T.J. Watson Research Center, Yorktown Heights, NY. He is currently a postdoctoral fellow at Nagoya Institute of Technology. His research interests include statistical speech recognition and synthesis. He received the Awaya and Itakura Awards from the Acoustical Society of Japan (ASJ) in 2006 and 2008, respectively, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2008. He is a member of ASJ and ISCA.

Yoshihiko Nankaku received the B.E. degree in Computer Science, and the M.E. and Dr.Eng. degrees in the Department of Electrical and Electronic Engineering from the Nagoya, Institute of Technology, Nagoya Japan, in 1999, 2001, and 2004 respectively. He is currently an Assistant Professor of Nagoya Institute of Technology. His research interests include image recognition, speech recognition and synthesis and multimodal interface. He is a member of the Acoustical Society of Japan.

Akinobu Lee was born in Kyoto, Japan, on December 19, 1972. He received the B.E. and M.E. degrees in information science, and the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 1996, 1998, and 2000, respectively. During 2000–2005, he was an Assistant Professor in Nara Institute of Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan. He is now an Associate Professor of Nagoya Institute of Technology. His research interests include speech recognition and spoken language understanding. He is a member of the IEE, ISCA, IPSJ and ASJ.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, in 1984 and the M.E. and Dr.Eng. degrees in 2003 information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1986 and 1989, respectively. From 1989 to 1996, he was a Research Associate in the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004, he was an Associate Professor in the Department of Computer Science, Nagoya Institute of Technology, where he is currently a Professor. He is also an Invited Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan, and was a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning. Prof. Tokuda is a corecipient of the Paper Award and the Inose Award, both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001 and 2008. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is a member the ISCA, IPSJ, ASJ, and JSAI.