Speech Synthesis Based on Gaussian Conditional Random Fields

Soheil Khorram, Fahimeh Bahmaninezhad, Hossein Sameti

Department of Computer Engineering, Sharif University of Technology, Tehran-Iran
khorram@ce.sharif.edu, bahmaninezhad@ce.sharif.edu,
sameti@sharif.edu

Abstract. Hidden Markov Model (HMM)-based synthesis (HTS) has recently been confirmed to be the most effective method in generating natural speech. However, it lacks adequate context generalization when the training data is limited. As a solution, current study provides a new context-dependent speech modeling framework based on the Gaussian Conditional Random Field (GCRF) theory. By applying this model, an innovative speech synthesis system has been developed which can be viewed as an extension of Context-Dependent Hidden Semi Markov Model (CD-HSMM). A novel Viterbi decoder along with a stochastic gradient ascent algorithm was applied to train model parameters. Also, a fast and efficient parameter generation algorithm was derived for the synthesis part. Experimental results using objective and subjective criteria have shown that the proposed system outperforms HSMM substantially in limited speech databases. Moreover, Mel-cepstral distance of the spectral parameters has been reduced considerably for any size of training database.

Keywords: Gaussian conditional random field, statistical parametric speech synthesis, HSMM extension.

1 Introduction

Statistical Parametric Speech Synthesis (SPSS) has reportedly been a dominant research area due to its peculiarities since the last decade [1, 2]. Modeling in the domain of SPSS is of prime importance and it is naïve to assume unnecessary simplifying assumptions in modeling as it may reduce the quality of synthetic speech. This work extends Hidden Semi Markov Model (HSMM) synthesis [3] by eliminating some of its simplifying assumptions. In the next subsection we will briefly discuss related works.

1.1 Related Work

Many research activities have already been performed to improve the quality of basic HTS. The progresses such as Hidden Semi Markov Model (HSMM) [3], Trajectory HMM [4] and Multi-Space Distribution HMM [5] have made HTS the most powerful

statistical approach. However, these systems do not lead to an acceptable quality with limited databases (less than 30 minutes). This deficiency is a direct result of applying decision-tree-based context clustering which cannot exploit contextual information efficiently, because each training sample is associated in modeling only one context cluster. This study is an attempt to improve SPSS quality even for limited training data.

The rest of the paper is organized as follows. In Section 2, GCRF is introduced. Sections 3 & 4 propose a context-dependent model for speech using GCRF and its application in speech synthesis. Experimental results are presented in Section 5 and final remarks are given in Section 6.

2 Gaussian Conditional Random Field

To define GCRF, first a brief description of Markov Random Field (MRF) and Conditional Random Field (CRF) is given.

Definition 1. Let $G = (V, E)$ be an undirected graph, $X = (X_v)_{v \in V}$ be a set of random variables indexed by nodes of G, X is modeled by MRF iff $\forall A, B \subseteq V,\ P(X_A|X_B) = P(X_A|X_S)$, where S is a border subset of A such that every path from a node in A to a node in B passes through S [6].

Definition 2. (X, C) is a CRF iff for any given set of random variables C, X forms an MRF [6].

In the speech synthesis framework, given an utterance contextual information C, sufficient statistics of speech (acoustic features) can be considered as an MRF.

Hammersley-Cliffort's Theorem. Suppose (x, c) is an arbitrary realization of a CRF (X, C) defined based on a graph G with positive probability, then $P(x|c)$ can be factorized by the following Gibbs distribution [7].

$$P(x|c) = \frac{1}{Z(c)} \prod_{\mathcal{A}} \Psi_a(x, c), \tag{1}$$

where $\mathcal{A}$ denotes a set of all maximal cliques of G. $Z(c)$ is called partition function which ensures that the distribution sums to one. In other words,

$$Z(c) = \iint_x \prod_{\mathcal{A}} \Psi_a(x, c). \tag{2}$$

The theorem also states that for any choice of positive local functions $\{\Psi_a(x)\}$ (potential functions) a valid CRF is gene-rated. One of the simplest choices of a potential function is Gaussian function. CRF with Gaussian potential function is named GCRF which is introduced in the next section.

3 Context-Dependent Speech Modeling Using GCRF

For modeling speech, the proposed system primarily splits each segment into a fixed number of states. Then, acoustic and binary contextual features (sufficient statistics)

are extracted for each state. The goal is to model and generate acoustic features provided that contextual features are present. The following notations are taken into account henceforth.

L,I:	Total number of acoustic and linguistic features.
$\mathcal{J}$:	Total number of states for the current utterance.
V:	All acoustic parameters. (Extracted from frame samples)
x_{lj}:	l-th acoustic feature of state j. (Extracted from V)
x_l:	l-th acoustic feature vector, $x_l \stackrel{\text{def}}{=} [x_{l1}, \dots, x_{l\mathcal{J}}]^T$.
X:	All acoustic features, $X \stackrel{\text{def}}{=} [x_1, \dots, x_L]$.
c_{ji}:	i-th binary linguistic feature of state j.
c_j:	Linguistic feature vector of state j, $c_j \stackrel{\text{def}}{=} [c_{j1}, \dots, c_{jI}]^T$.
C:	All linguistic features, $C \stackrel{\text{def}}{=} [c_1, \dots, c_{\mathcal{J}}]$.

3.1 GCRF Graphical Structure

Factor graph [8] of the proposed GCRF (with order one) is depicted in Figure 1. As it is obvious in the figure, GCRF is a set of L linear chain CRF [8] (with order one) which are independent when C is given. Each rectangular node (Ψ_{lj}) represents a potential function describing the effect of a maximal clique ($x_{lj}, x_{l(j-1)}, c_j$) in the random field distribution. This figure can be extended to higher order linear chain CRFs. As a result, if GCRF extends with order o, Ψ_{lj} becomes a function of ($x_{lj}, \dots, x_{l(j-o)}, c_j$).

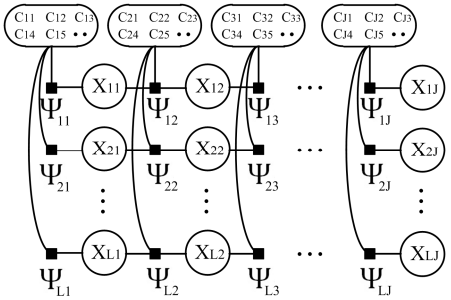

Fig. 1. Factor graph of the first order GCRF.

3.2 GCRF Distribution

Having described the graphical model, this subsection investigates the probability distribution provided by GCRF. Markov property of MRFs implies the following equality.

$$P(X|C;\theta) = \prod_{l=1}^{L} P(x_l|C;\theta), \tag{3}$$

where θ is the set of all model parameters. This paper assumes that the partition function, Ψ_{lj}, is formulated by Equation 4 which is a Gaussian function with parameters H_{lji} and u_{lji}.

$$\Psi_{lj} \stackrel{\text{def}}{=} \exp\left\{-\frac{1}{2}\sum_{i=1}^{I}\left[\left(x_l^T H_{lji} x_l + u_{lji}^T x_l\right)c_{ji}\right]\right\}. \tag{4}$$

In this equation, H_{lji} has to be a symmetric and positive definite matrix. If H_{lji} is not restricted to a positive definite matrix, the distribution may be realized by a number greater than one. Thus, considering positive definite condition seems to be necessary. Moreover, in GCRF with order o, H_{lij} and u_{lij} contain only $(o+1)\times(o+1)$ and $(o+1)$ nonzero elements respectively. The overall form of model parameters is shown as follows.

$$H^{lij} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & h^{lij}_{(j-o)(j-o)} & \cdots & h^{lij}_{(j-o)j} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ 0 & h^{lij}_{j(j-o)} & \cdots & h^{lij}_{jj} & 0 & \cdots \\ 0 & 0 & \dots & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \end{bmatrix}, u^{lij} = \begin{bmatrix} 0 \\ u^{lij}_{j-o} \\ \vdots \\ u^{lij}_{j} \\ 0 \\ \vdots \end{bmatrix}. \tag{5}$$

By considering defined potential function and according to the fundamental theorem of Hammersley and Cliffort the final expression for $P(x_l|C;\theta_l)$ is given by

$$P(x_l|C;\theta_l) = \frac{1}{Z_l(C;\theta_l)}\exp\left\{-\frac{1}{2}(x_l^T H_l x_l + u_l^T x_l)\right\}, \tag{6}$$

where $H_l = \sum_{j=1}^{J}\sum_{i=1}^{I} c_{ji} H_{lji}$ and $u_l = \sum_{j=1}^{J}\sum_{i=1}^{I} c_{ji} u_{lji}$.
Z_l is the partition function and is computed by Equation 2. Fortunately, for Gaussian distribution of Equation 4 there is a closed formula for the partition function as:

$$Z_l(C;\theta_l) = (2\pi)^{\frac{J}{2}}(\det(H_l^{-1}))^{\frac{1}{2}}\exp\left(\frac{1}{8}u_l^T H_l^{-1} u_l\right). \tag{7}$$

A marvelous point is that conventional CD-HSMM can be considered as a type of GCRF with order zero and mutually exclusive contextual features.

4 Speech Synthesis Based on GCRF

Figure 2 shows an overview of the proposed GCRF-based speech synthesis system. All blocks in the figure are identical to classical SPSS [1], except the three further blocks added with a different color. In the training part, acoustic sufficient statistics or features (X) are extracted according to both speech parameters (V) and state boundaries ($\mathcal{T}$). State boundaries are latent and the added Viterbi block is employed to train

them in an unsupervised manner. It should be noted that only sufficient statistics are modeled in the training phase; therefore synthesis phase has to generate them first. After generating features, speech parameters and speech signal are successively synthesized.

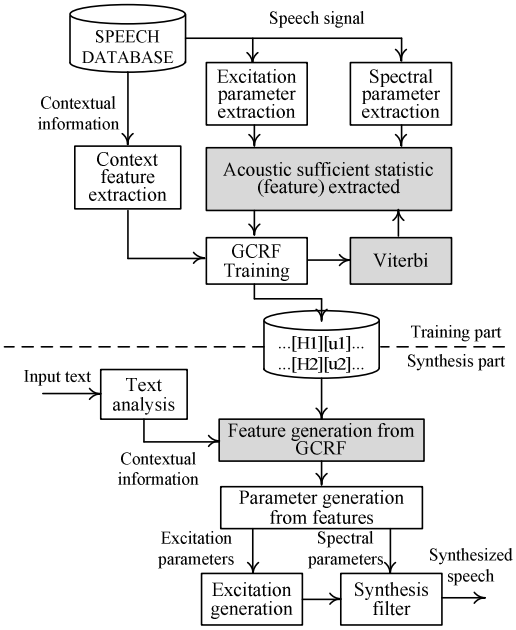

Fig. 2. An overview of the proposed architecture.

4.1 Estimation of Model Parameters

In this section, we discuss how to train model parameters θ. We are given a set of T iid training data $\{X^t, C^t\}_{t=1}^T$, the goal is to find the best set of parameters, $\hat{\theta}$, which maximizes the conditional log likelihood:

$$\hat{\theta} = \text{argmax}_\theta \, L(\theta), \tag{8}$$

$$L(\theta) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \log P(X^t|C^t;\theta). \tag{9}$$

The problem is that, acoustic feature Matrix X^t, wholly depends on the state boundaries which are latent. Hence, it is impossible to compute $L(\theta)$. A correct solution for this problem that converges to the Maximum Likelihood (ML)-estimate is given by the Expectation Maximization (EM) algorithm; however, EM is computationally expensive. Another commonly used method which is computationally efficient and works well in practice is to compute first X^t and then $L(\theta)$ on the Viterbi path. Appling this approach and substituting $P(X^t|C^t;\theta)$ with Equation 6 gives

$$\mathrm{L}(\theta) = -\frac{1}{2\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\sum_{l=1}^{\mathrm{L}}\{\mathrm{L}_l^{\mathrm{t}}(\theta_l)\}, \tag{10}$$

$$\mathrm{L}_l^{\mathrm{t}}(\theta_l) \stackrel{\text{def}}{=} \mathrm{x}_l^{\mathrm{t}T}\mathrm{H}_l^{\mathrm{t}}\mathrm{x}_l^{\mathrm{t}} + \mathrm{u}_l^{\mathrm{t}T}\mathrm{x}_l^{\mathrm{t}} + \mathcal{J}\log 2\pi - \log\det\mathrm{H}_l^{\mathrm{t}} + \frac{1}{4}\mathrm{u}_l^{\mathrm{t}T}\mathrm{H}_l^{\mathrm{t}-1}\mathrm{u}_l^{\mathrm{t}}. \tag{11}$$

In general, this function cannot be maximized in closed form, therefore numerical optimization is used. The partial derivatives of $\mathrm{L}(\theta)$ are calculated as follows.

$$\frac{\partial\mathrm{L}(\theta)}{\partial\mathrm{u}_{l\mathrm{ij}}} = -\frac{1}{2\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\frac{\partial\mathrm{L}_l^{\mathrm{t}}(\theta_l)}{\partial\mathrm{u}_{l\mathrm{ij}}}, \tag{12}$$

$$\frac{\partial\mathrm{L}_l^{\mathrm{t}}(\theta)}{\partial\mathrm{u}_{l\mathrm{ij}}} = \left[\left(\mathrm{x}_l^{\mathrm{t}} + \frac{1}{2}\mathrm{H}_l^{\mathrm{t}-1}\mathrm{u}_l^{\mathrm{t}}\right)\mathrm{c}_{\mathrm{ji}}^{\mathrm{t}}\right] \star \mathbb{b}(\mathcal{J},\mathrm{j},\mathrm{o}), \tag{13}$$

$$\frac{\partial\mathrm{L}(\theta)}{\partial\mathrm{H}_{l\mathrm{ij}}} = -\frac{1}{2\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\frac{\partial\mathrm{L}_l^{\mathrm{t}}(\theta_l)}{\partial\mathrm{H}_{l\mathrm{ij}}}, \tag{14}$$

$$\frac{\partial\mathrm{L}_l^{\mathrm{t}}(\theta_l)}{\partial\mathrm{H}_{l\mathrm{ij}}} = \left[\left(\mathrm{x}_l^{\mathrm{t}}\mathrm{x}_l^{\mathrm{t}\mathrm{T}} - \mathrm{H}_l^{\mathrm{t}-1} - \frac{1}{4}\mathrm{H}_l^{\mathrm{t}-1}\mathrm{u}_l^{\mathrm{t}}\mathrm{u}_l^{\mathrm{t}\mathrm{T}}\mathrm{H}_l^{\mathrm{t}-1}\right)\mathrm{c}_{\mathrm{ji}}^{\mathrm{t}}\right] \star \mathbb{B}(\mathcal{J},\mathrm{j},\mathrm{o}). \tag{15}$$

where o denotes the order of model, $\star$ denotes element-by-element product operator and $\mathbb{B}$ ($\mathbb{b}$) is a $\mathcal{J}$-by-$\mathcal{J}$ ($\mathcal{J}$) Boolean matrix (vector) defined by an indicator function I as:

$$\mathbb{b}(\mathcal{J},\mathrm{j},\mathrm{o}) \stackrel{\text{def}}{=} [\mathbb{b}_{\mathrm{m}}(\mathcal{J},\mathrm{j},\mathrm{o})]_{\mathcal{J}\times 1}, \tag{16}$$

$$\mathbb{b}_{\mathrm{m}}(\mathcal{J},\mathrm{j},\mathrm{o}) \stackrel{\text{def}}{=} \mathrm{I}(\mathrm{j}-\mathrm{o} \le \mathrm{m} \le \mathrm{j}),$$

$$\mathbb{B}(\mathcal{J},\mathrm{j},\mathrm{o}) \stackrel{\text{def}}{=} [\mathbb{B}_{\mathrm{mn}}(\mathcal{J},\mathrm{j},\mathrm{o})]_{\mathcal{J}\times 1}, \tag{17}$$

$$\mathbb{B}_{\mathrm{mn}}(\mathcal{J},\mathrm{j},\mathrm{o}) \stackrel{\text{def}}{=} \mathrm{I}\big((\mathrm{j}-\mathrm{o} \le \mathrm{m} \le \mathrm{j})\ \&(\mathrm{j}-\mathrm{o} \le \mathrm{n} \le \mathrm{j})\big).$$

A common solution of this optimization problem is to take entire training samples into account and update model parameters using an optimization algorithm such as *BFGS*. Unfortunately, this in turn leads to large computational complexity. This paper proposes the application of *stochastic gradient ascent* [9] method which is faster than above-mentioned algorithm by orders of magnitude. This method has proven to be effective [9]. Following equations express its updating rule:

$$\mathrm{u}_{l\mathrm{ij}}^{\mathrm{t}} = \mathrm{u}_{l\mathrm{ij}}^{\mathrm{t}-1} - \alpha^t\left.\frac{\partial\mathrm{L}_l^{\mathrm{k}}(\theta_l)}{\partial\mathrm{u}_{l\mathrm{ij}}}\right|_{\mathrm{u}_{l\mathrm{ij}}^t,\mathrm{H}_{l\mathrm{ij}}^t}, \tag{18}$$

$$\mathrm{H}_{l\mathrm{ij}}^{\mathrm{t}} = \mathrm{H}_{l\mathrm{ij}}^{\mathrm{t}-1} - \alpha^t\left.\frac{\partial\mathrm{L}_l^{\mathrm{t}}(\theta_l)}{\partial\mathrm{H}_{l\mathrm{ij}}}\right|_{\mathrm{u}_{l\mathrm{ij}}^{\mathrm{t}},\mathrm{H}_{l\mathrm{ij}}^{\mathrm{t}}}. \tag{19}$$

A variable step size algorithm described by [10] is utilized in our experiments.

4.2 Viterbi Algorithm for GCRF

Given a sequence of acoustic parameters (V), sentence contextual features (C) and a trained GCRF parameters (θ), this section presents an algorithm to find the most likely state boundaries ($\hat{\mathcal{T}}$). Thus the aim is to estimate $\hat{\mathcal{T}}$ such that

$$\hat{\mathcal{T}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\mathcal{T}} \mathrm{P}(\mathcal{T}|\mathrm{V},\mathrm{C};\theta) = \operatorname{argmax}_{\mathcal{T}} \mathrm{P}(\mathrm{X}(\mathcal{T},\mathrm{V})|\mathrm{V},\mathrm{C};\theta). \tag{20}$$

From Equation 6 we have

$$\hat{\mathcal{T}} = \operatorname{argmin}_{\mathcal{T}} \sum_{\mathrm{j}=1}^{\mathcal{J}} \phi_{\mathrm{j}}(\mathcal{T},\mathrm{V},\mathrm{C},\theta), \tag{21}$$

where $\phi_{\mathrm{j}}(\mathcal{T},\mathrm{V},\mathrm{C},\theta) \stackrel{\text{def}}{=} \sum_{\mathrm{l}=1}^{\mathrm{L}} \sum_{\mathrm{i}=1}^{\mathrm{I}} \left(\mathrm{x}_{\mathrm{lj}}^{\mathrm{T}} \mathrm{H}_{\mathrm{lij}} \mathrm{x}_{\mathrm{lj}} + \mathrm{b}_{\mathrm{lij}}^{\mathrm{T}} \mathrm{x}_{\mathrm{lj}}\right) \mathrm{c}_{\mathrm{ji}}$.
Let t_{j} be the j-th state boundary (j-th element of $\mathcal{T}$), then for a GCRF with order o, ϕ_{j} becomes a function of $t_{\mathrm{j-o-1}}, \ldots, t_{\mathrm{j}}$ instead of entire elements of $\mathcal{T}$. This fact gives us an ability to exploit dynamic programming for performing a complete search on $\mathcal{T}$. Inspired by the other Viterbi algorithms, we need to define an auxiliary variable δ_{j}.

$$\delta_{\mathrm{j}}(t_{\mathrm{j-o}}, \ldots, t_{\mathrm{j}}) \stackrel{\text{def}}{=} \min_{t_1,\ldots,t_{\mathrm{j-o-1}}} \sum_{j=1}^{\mathrm{j}} \phi_{\mathrm{j}}(t_{j-\mathrm{o}-1}, \ldots, t_j). \tag{22}$$

δ_{j} can be calculated from $\delta_{\mathrm{j-1}}$ by following recursion.

$$\delta_{\mathrm{j+1}}(t_{\mathrm{j-o+1}}, \ldots, t_{\mathrm{j+1}}) = \min_{t_{\mathrm{j-o}}} \left[\delta_{\mathrm{j}}(t_{\mathrm{j-o}}, \ldots, t_{\mathrm{j}}) + \phi_{\mathrm{j+1}}(t_{\mathrm{j-o}}, \ldots, t_{\mathrm{j+1}})\right]. \tag{23}$$

Using this recursion, it is straightforward to obtain Viterbi algorithm.

4.3 Parameter Generation Algorithm

This section, for a given GCRF, derives an algorithm to estimate the best synthesized speech parameters ($\hat{\mathrm{V}}$) by maximizing the likelihood criteria, i.e.

$$\hat{\mathrm{V}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\mathrm{V}} \mathrm{P}(\mathrm{V}|\theta) = \operatorname{argmax}_{\mathrm{V}} \sum_{\mathcal{T}} \mathrm{P}(\mathrm{X}(\mathrm{V},\mathcal{T})|\theta). \tag{24}$$

The synthesis part needs to respond quickly, however, solving this problem directly is challenging. Hence, the algorithm derived from Equation 24 is not practical.
A two-step algorithm is proposed here which approximates $\hat{\mathrm{V}}$ fast.
Step 1. For a given θ, compute the ML-estimate of X:

$$\hat{\mathrm{X}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\mathrm{X}} \mathrm{P}(\mathrm{X}|\theta). \tag{25}$$

Step 2. For a given X, compute the ML-estimate of V:

$$\hat{\mathrm{V}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\mathrm{V}} \mathrm{P}(\mathrm{V}|\mathrm{X}). \tag{26}$$

The first step is simply obtained by considering the distribution discussed in section 3. Since different acoustic features are statistically independent (given in Equation 3), the algorithm can generate features independently, i.e.

$$\hat{x}_l = \text{argmax}_{x_l} P(x_l|C; \theta_l). \quad (27)$$

Optimizing the Gaussian distribution $P(x_l|C; \theta_l)$, expressed by Equation 6, results in the set of linear equations below:

$$H^l \hat{x}_l = -\frac{1}{2} b^l. \quad (28)$$

H^l is symmetric and positive definite, so Equation 28 can be efficiently solved using the Cholesky decomposition.
Second step depends heavily on the selected acoustic features. For the set of acoustic features extracted in our system, Tokoda et al. [11] algorithm was used in this step.

5 EXPERIMENTS

5.1 Experimental Conditions

To evaluate the proposed system, a Persian speech database [12] consisting of 1000 utterances with an average length of 8 seconds was employed. Experiments were conducted on a fixed test set of 200 utterances and 5 different training sets with remaining 50, 100, 200, 400 and 800 utterances. It should be noted that the average length of each utterance is about 8 seconds. Speech parameters including mel-cepstral coefficients, bandpass aperiodicity and fundamental frequency were extracted by STRAIGHT [13]. Sample mean and variance of each static and dynamic parameter, in addition to the voicing probability and duration are computed as the acoustic state features. For contextual state features a set of 150 well designed binary questions are employed. Following subsections evaluate the proposed method in contrast to the HSMM-based technique.

5.2 Objective Evaluation

As Figure 3 shows, three objective measures were calculated to evaluate the proposed and HSMM-based systems, namely the average mel-cepstral distortion (expressed in dB) [14], the Root-Mean-Square (RMS) error of fundamental frequency logarithm (expressed in cent) and the RMS error of phoneme durations (expressed in terms of number of frames). Computing the first and second measures needs an assumption about state boundaries that was estimated here using the Viterbi algorithm. Since F0 value is not observed in unvoiced regions, only voiced frames of speech were taken into account for the second measure.

From Figure 3, it is noticeable that GCRF always outperforms HSMM in generating mel-cepstral and duration parameters, but HSMM is superior in synthesizing fundamental frequency when the number of training data is larger than 200 utterances. This drawback is a result of weak estimation of F0 parameters during the training process. Table 1 compares the accuracy of voiced/unvoiced detection in proposed system with its counterpart in HSMM-based synthesis.

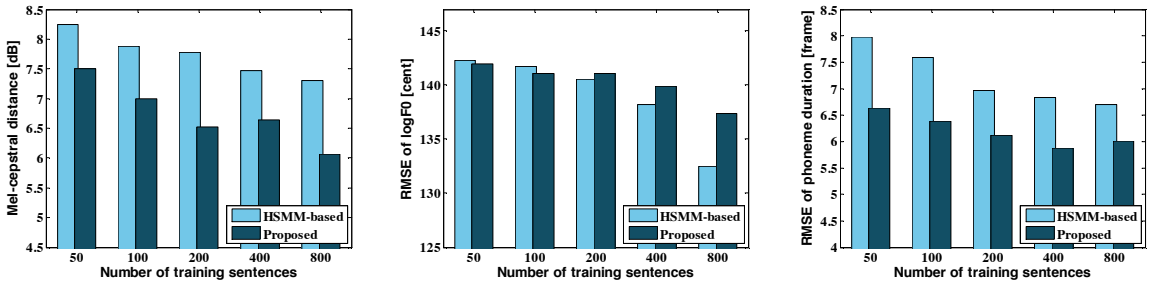

Fig. 3. Objective evaluation of HSMM-based and proposed speech synthesis systems. (Left) Mel-cepstral distance [dB]; (Middle) RMSE of log F_0 [cent]; (Right) RMSE of phoneme duration [frame].

5.3 Subjective Evaluation

We conducted preference score measure to compare the proposed and HSMM-based systems subjectively. 20 subjects were presented with 10 randomly chosen pairs of synthesized speech from the two models and then asked for their preference.

Figure 4 shows the average preference score. The result confirms that the synthetic speech generated by proposed system has been favorable when training data are limited.

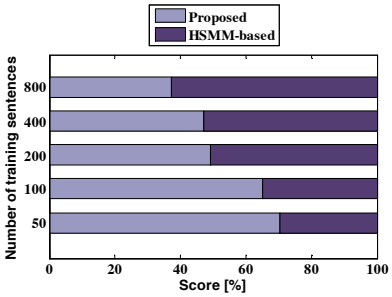

Fig. 4. Subjective evaluation of HSMM and proposed systems using preference score.

Table 1. Accuracy of Voiced/Unvoiced Detector.

# train data	Proposed accuracy	HSMM accuracy
50	0.9184	0.8851
100	0.9241	0.8828
200	0.9157	0.8903
400	0.9104	0.8783
800	0.9037	0.8809

6 Conclusion

This paper improves HSMM-based synthesis in the following ways:

1. The independence assumption of states distribution in HTS is removed.
2. In contrast to HMM, the proposed model does not limit its potential functions to be a probability distribution.

3. CD-HMM uses decision-tree-based context clustering that does not provide efficient generalization in limited training data, because each speech parameter vector is associated in modeling of only one context cluster. In contrast, our method contributes each training vector in many clusters to offer an efficient generalization.

Despite the advantages, which made our system to outperform in small training data, a drawback such as difficult training procedure is noticed in large databases.

7 References

1. A.W. Black, H. Zen, and K. Tokuda, "Statistical Parametric Speech Synthesis", *ICASSP'2007*, Honolulu, Hawai'i, USA, pp. IV-1229-IV-1232, 2007.
2. H. Zen, K. Tokuda, and A.W. Black, "Statistical Parametric Speech Synthesis", *Speech Communication Elsevier*, Vol. 51, Issue 11, Nov., 2009.
3. H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden Semi-Markov Model Based Speech Synthesis", *Interspeech'2004*, Jeju Island, Korea, pp. 1393-1396, October 4-8, 2004.
4. H. Zen, K. Tokuda, and T. Kitamura, "An Introduction of Trajectory Model into HMM-based Speech Synthesis", *SSW5*, Carnegie Mellon University, pp. 191-196, June, 2004.
5. K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi "Multi-Space Probability Distribution HMM", *IEICE Transaction on Information and Systems*, Vol. E85-D, No.3, pp. 455-464, 2002.
6. J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282-289, 2001.
7. G.R. Grimmett, "A Theorem about Random Fields", *Bulletin of the London Mathematical Society*, Vol. 5, pp. 81-84, 1973.
8. C. Sutton, and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning", In Lise Getoor and Ben Taskar, editors, *Introduction to statistical Relational Learning*, MIT Press, 2006.
9. W.A. Gardner, "Learning Characteristics of Stochastic-Gradient-Descent Algorithms: A General Study, Analysis and Critique", *Signal Processing* 6, No. 2, pp. 113-133, 1984.
10. M.N. Vrahatis, G.S. Androulakis, J.N. Lambrinos, and G.D. Magoulas, "A Class of Gradient Unconstrained Minimization Algorithms with Adaptive Stepsize", *Journal of Computational and Applied Mathematics* 114, No. 2, pp. 367-386, 2000.
11. K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", *ICASSP'2000*, Vol.3, Istanbul, pp. 1315- 1318, June, 2000.
12. M. Bijankhan, J. Sheikhzadegan, M.R. Roohani, Y. Samareh, C. Lucas, and M. Tebiani, "The Speech Database of Farsi Spoken Language", *Proceedings of 5th Australian International Conference on Speech Science and Technology (SST'94)*, pp. 826-831, 1994.
13. H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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*, Vol. 27, No. 3-4, pp. 187-207, 1999.
14. R. F. Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment", *in Proc. IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, 1993, pp. 125-128.