The CSI Journal on
Computer Science and Engineering
Vol. 10, No. 2 & 4 (b), 2012
Pages 47-59
Regular Paper

Spectral Modeling Based on Gaussian Conditional Random Field for Statistical Parametric Speech Synthesis

Soheil Khorram Hossein Sameti Fahimeh Bahmaninezhad

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Abstract

This paper proposes an innovative spectral modeling approach based on *Gaussian conditional random field (GCRF)* theory. The proposed method is also incorporated in a *statistical parametric speech synthesis (SPSS)* framework. Conventionally, SPSS systems exploit *hidden Markov model (HMM)*-based spectral modeling technique which suffers from a trivial problem known as state independence assumption. This shortcoming refers to the fact that the distributions of adjacent frames are modeled independently in HMM, whilst they are highly dependent and correlated. The proposed model assumes that spectral trajectories form a left-to-right linear-chain *conditional random field (CRF)* with Gaussian potential functions. Therefore, instead of the inaccurate independence assumption, Markov assumption is established for adjacent frames in a latent state. In order to train the proposed GCRF model a Viterbi algorithm along with a *maximum likelihood (ML)*-based parameter estimation procedure have been applied. The estimation algorithm leads to an optimization problem which is solved numerically through the *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* algorithm. In synthesis phase, an efficient parameter generation algorithm optimizing output probability measure has been derived. The designed parameter generation algorithm has the ability to exploit dynamic features as well as static features. Two sets of experiments are reported to prove the effectiveness of the proposed GCRF. In the first set, GCRF with some heuristic context clusters and ML-based parameter estimation is evaluated in contrast to the predominant HMM-based method. The results of objective and subjective tests confirm that the proposed system using heuristic contextual clusters outperformed the standard HMM in small training databases (i.e. 50, 100 and 200 sentences), but in large datasets HMM performs better. It is mainly due to the inability of the proposed system to adjust the number of model parameters with the size of training database. In the second set of experiments, the performance of GCRF using decision tree-based clusters is investigated. This later model has the ability to change the model complexity according to the size of training database. All evaluation results of this experiment confirm significant improvement of the proposed system over the conventional HMM.

Keywords: Gaussian Conditional Random Field, GCRF, Hidden Markov Model, HMM, HMM-Based Speech Synthesis, Spectral Modeling, State Independence Assumption, Statistical Parametric Speech Synthesis.

1. Introduction

The automatic conversion of written text to speech waveform is commonly called *text-to-speech (TTS)*. TTS systems are generally composed of two main subsystems. The first one converts input text into several language specifications called contextual factors and the second subsystem is known as *speech synthesis* which uses the contextual factors to generate a synthesized waveform [1-3]. This paper introduces a new speech synthesis system.

Among all synthesis systems designed for TTS application, *statistical parametric speech synthesis (SPSS)* has emerged as the most common method during the last decade [1-3]. Overall architecture of a typical SPSS is shown in figure 1. According to this figure, an SPSS system comprises two distinct phases: *training* and *synthesis*. Training phase starts with the extraction of acoustic features and contextual factors for all utterances in the training database. Acoustic features, including spectral, excitation and duration parameters, are extracted by a speech vocoder (e.g. *MELP* [4], *STRAIGHT* [5], *DSM* [6, 7] and *HNM* [8]).

Thereafter, the relationship between extracted acoustic features and contextual factors are captured through context-dependent statistical models [2]. In the synthesis phase, contextual factors are first obtained for a given text using the same natural language processor applied in the training phase. Next, a *parameter generation (PG)* algorithm [9-12] is applied to generate acoustic trajectories. Acoustic parameters are then fed into the same vocoder used during the training phase in order to generate synthesized speech.

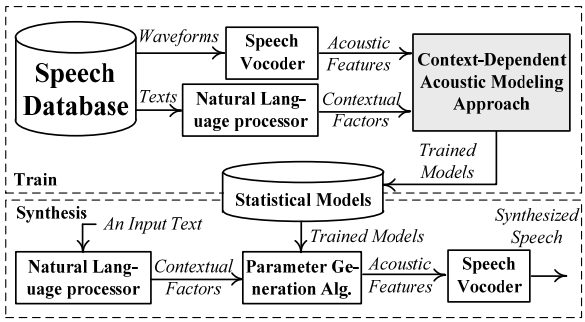


Figure 1. Block diagram of a typical SPSS

Statistical representation of speech in SPSS causes several advantages including high flexibility to modify voice characteristics [13], capability to exploit all speech recognition techniques (e.g. adaptation methods) [14, 15], proper support of multilingual synthesizers [16-18], improved robustness [19], improved coverage of acoustic space, low memory requirement [2]. However, all mentioned advantages are achieved at the expense of one important disadvantage, i.e. reduced quality of synthesized speech [2]. One major reason of this problem is the inefficiency of statistical models. This paper is an attempt to improve the performance of the predominant HMM-based statistical modeling method.

Multi-stream Left-to-right without skip transitions *hidden semi-Markov model (HSMM)* [20] applying *multi-space probability distribution (MSD)* [21] has emerged as the most common method for statistical modeling during the last decade. For the sake of simplicity, this predominant model is simply called HMM in this paper. The focus in this article is hereby on improving the performance of conventional HMM-based modeling [22] by introducing an accurate statistical modeling approach. The improvement is achieved by releasing one inaccurate assumption of HMM-based modeling, namely state independence assumption. In fact, HMM assumes that the output probability distributions of successive frames are independent of each other. This assumption causes HMM to generate piece-wise constant trajectories which are not similar to natural trajectories. In the next section, we will briefly review related works.

1.1. Related Work

To overcome the unfavorable effects caused by state independence assumption of the standard HMM, a variety of sophisticated models have been proposed. These models include HMMs with polynomial regression functions [23-25], hidden dynamic models [26-29], partly hidden Markov models [30], stochastic segment models [31], segmental HMMs [32-34], temporally varying means and precisions [35, 36], frame-correlated HMMs [37-41], buried Markov models [42], switching linear dynamical systems [43, 44], dynamic Bayesian networks [45] and etc. All the above-mentioned methods have been designed for speech recognition application.

However, a smaller number of methods, that improve predominant HMM, have been proposed in the speech synthesis field of research. One of the first techniques is based on *trajectory HMM* [46, 47] which reformulates the HMM by imposing explicit relationships between static and dynamic acoustic features. Trajectory HMM removes the incorrect conditional independence assumption of state output probabilities in HMM structure, at the expense of intensive training procedure. This system is then outperformed by integrating the *global variance (GV)* constraint into its training procedure [48]. Autoregressive HMM [49, 50] is another modeling method that is able to eliminate the mentioned independence assumption with a much more computationally tractable parameter estimation algorithm. *Gaussian process regression (GPR)* [51] is another new technique that releases the incorrect stationarity assumption of the state output distribution in HMM. GPR uses frame-level contextual factors to predict frame-level acoustic trajectories. These frame-level factors are then used as the explanatory variable in a GPR framework.

The fact that classical HMM expresses each frame distribution independent of its adjacent frames leads to an insufficient context generalization as well; because HMM cannot capture cross-correlation between adjacent frames. Also, HMM-based speech synthesis exploits a decision tree-based clustering method to capture the dependencies between acoustic features and contextual factors [52]. This decision tree clustered structure is another reason for inadequate generalization to unseen models [53]. Many efforts have in turn devoted to improve the generalization capabilities of HMM. One of the most notable works is developed based on *deep neural networks (DNNs)* [53]. DNNs are able to predict difficult context dependencies by applying plenty of hidden layers, as opposed to the decision tree structure that is not efficient enough to predict complex dependencies such as XORs or multiplexers [53]. Other deep learning approaches such as *restricted Boltzmann machines (RBMs)* [54] and *deep belief networks (DBNs)* [55] have also been demonstrated to be effective in SPSS. Some other methods also offer superior generalization by replacing the non-overlapped clusters of decision tree with a number of overlapped regions. These methods include *contextual additive model* [56-58], in which acoustic trajectories are assumed to be a superposition of multiple additive components with different decision trees, and *hidden maximum entropy model (HMEM)* [59] which estimates the smoothest distribution preserving statistics of the overlapped clusters.

1.2. Scope of the Paper

As previously stated, classical HMM-based speech synthesis employs a decision tree clustered left-to-right without skip hidden semi-Markov model [20] in the statistical modeling phase. Roughly speaking, this model initially partitions acoustic trajectories into a fixed number of time slices (so-called states) and then the distribution of each state is simply expressed using a context-dependent [60] multi-space probability distribution [21]. This rough explanation shows

that the distributions of successive frames in predominant HMM are modeled independently and the correlations between adjacent frames are completely forgotten. Therefore, HMM is unable to exploit the statistics of training data efficiently, and it suffers from inadequate generalization. In other words, HMM is only able to capture statistics of one individual frame; while, it is possible to design a model capturing mutual statistics of adjacent frames. This paper presents a new modeling method which is designed based on a *Gaussian conditional random field (GCRF).* GCRF is a random field with Markovian property that defines Gaussian potential functions. As it will be described later, GCRF is able to model the dependencies of adjacent frames by defining its potential functions as functions of two succeeding frames.

The rest of the paper is organized as follows. In Section 2, the fundamental theory of GCRF is discussed. Section 3 introduces a context-dependent acoustic modeling method using GCRF. The proposed acoustic modeling is then incorporated into a SPSS system in Section 4. Experimental results are presented in Section 5 and final remarks are given in Section 6.

2. Gaussian Conditional Random Field

In order to introduce GCRF-based speech synthesis, first a brief description of *Markov random field (MRF)* and *conditional random field (CRF)* is given in this section. The definitions presented in this section are minimum prerequisites for our future discussion.

MRF definition: Let $G = (V, E)$ be an undirected graph with node set V and edge set E, $X = (X_v)_{v \in V}$ be a set of random variables indexed by nodes of G, X is modeled by MRF if and only if $\forall A, B \subseteq V,\ P(X_A|X_B) = P(X_A|X_S)$, 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 [61].

CRF definition: (X, C) is a CRF iff for any given set of random variables C, X forms an MRF [61, 62].

In the speech synthesis framework, given an utterance contextual information C, acoustic features of an arbitrary hidden state can be assumed to be conditionally independent of all other features given its adjacent frames; therefore, CRF seems to be a promising structure for modeling the random field formed by acoustic trajectories.

Hammersley-Clifford'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 [61].

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

where $\mathcal{A}$ denotes a set of all maximal cliques of G. It should be noted that clique is a group of nodes that all of them are mutually connected and maximal clique is a clique that cannot be extended by including even one adjacent node. Also, $Z(c)$ is called partition function which ensures that the distribution sums to one. In other words,

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

The theorem also states that for any choice of positive local functions $\{\Psi_a(x)\}$ (called potential functions) a valid CRF is generated. One of the simplest choices of a potential function is Gaussian function.

GCRF definition: CRF with Gaussian potential function is named GCRF.

CRF model has widely been used in variety of speech and signal processing applications such as speech recognition [63-65], speaker verification [66] and gesture recognition [67]. In all of these applications, the distribution of multiple discrete class labels given speech signal is modeled through GCRF framework, while in the application of speech synthesis the distribution of continues speech parameter trajectories given some contextual factors has to be modeled; therefore our final model is completely different from previous CRF modeling schemes.

3. GCRF-Based Spectral Modeling

Speech spectrum envelope is normally parameterized through a number of spectral features, such as *linear prediction coefficients (LPCs)* and *mel-cepstral (mcep)* [68, 69] coefficients. To model these coefficients, standard HMM imposes a quasi-stationary assumption on the spectral trajectories. These trajectories are hereby split into a fixed number of time intervals (so-called states); then independent and identical distributions are trained for state output distribution. Although this quasi-stationary assumption might be valid in some cases (e.g. where signals are recorded in extremely controlled situations), it is not generally satisfied, because spectral parameters are by nature non stationary. As a consequence, HMM is clearly unable to represent intra-state time-dependencies. This study assumes that intra-state time-dependencies of spectral parameters follow the Markovian property and spectral trajectories form an MRF. Based on this assumption a new distribution is derived in this section.

3.1. GCRF Graphical Structure

The statistical dependencies of MRF-based graphical models can be simply shown through a factor graph [61]. Factor graph is an undirected graph with two types of vertices representing random variables and potential functions. Conventionally, circular and square nodes are respectively used for illustrating random variables and potential functions. Edges of the factor graph are just allowed to connect a potential function with a random variable and it means that the potential function requires the random variable as an input argument. Figure 2 compares the factor graph of the standard HMM and the proposed GCRF. In this figure, S_t indicates state label of t-th frame and o_{lt} is the l-th dimension of the acoustic feature extracted for frame t. Also, P_{lt} and Ψ_{lt} are defined as the potential functions of HMM and GCRF associated with l-th dimension in time t. As it is realized from this figure, various dimensions are modeled independently in both HMM and GCRF. Additionally, potential functions of HMM depend only on the current feature; while, in GCRF potential functions are assumed to be functions of both current and previous frames. Therefore, GCRF unlike HMM is able to capture the dynamics of acoustic trajectories.

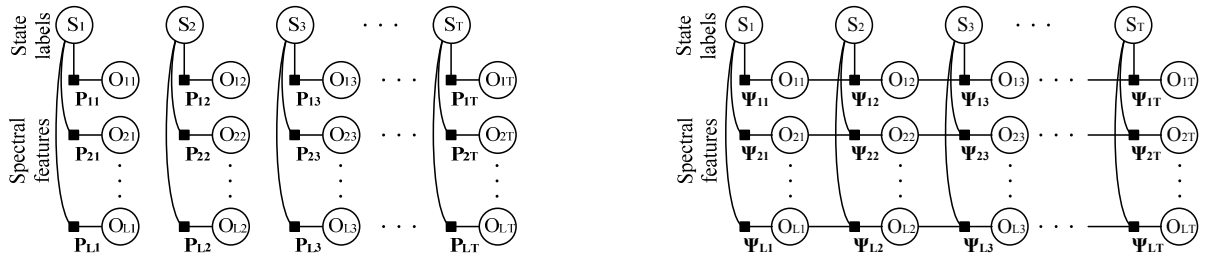


Figure 2. Factor graph of the (a) conventional HMM, (b) proposed GCRF

Let us first explore the distribution of HMM by considering the graphical model of figure 2 (a). The following equation expresses the relationship commonly used for HMM potential function:

$$\mathrm{P}_{lt}(\mathrm{o}_{lt}|\mathrm{s}_\mathrm{t};\lambda) \stackrel{\text{def}}{=} \exp\left\{-\frac{\left(\mathrm{o}_{lt}-\mu_l(\mathrm{s}_\mathrm{t})\right)^T\left(\mathrm{o}_{lt}-\mu_l(\mathrm{s}_\mathrm{t})\right)}{2{\sigma_l}^2(\mathrm{s}_\mathrm{t})}\right\}. \quad (3)$$

This potential function is slightly different for multi-mixture or multi-space output probability distribution HMM, but we confine our discussion to a context-dependent HMM with just one mixture and one space; because it has been proved that for spectral modeling, increasing the number of mixtures or spaces has no tangible effect on the quality of synthesized speech.

To obtain the final distribution of HMM, λ, Eq. (3) has to be replaced with the potential function of Eq. (1). The replacement leads to the final distribution expressed by Eq. (4) which is a well-known multivariate Gaussian distribution.

$$\mathrm{P}_{\mathrm{HMM}}(\mathrm{o}|\mathrm{s};\lambda) = \prod_{\mathrm{t}=1}^{\mathrm{T}}\prod_{l=1}^{\mathrm{L}}\frac{1}{\sigma_l(\mathrm{s}_\mathrm{t})\sqrt{2\pi}}\exp\left\{-\frac{\left(\mathrm{o}_{lt}-\mu_l(\mathrm{s}_\mathrm{t})\right)^T\left(\mathrm{o}_{lt}-\mu_l(\mathrm{s}_\mathrm{t})\right)}{2{\sigma_l}^2(\mathrm{s}_\mathrm{t})}\right\}. \quad (4)$$

In this equations, $\mu_l(\mathrm{s}_\mathrm{t})$ and $\sigma_l^2(\mathrm{s}_\mathrm{t})$ are the l-th dimension of the context-dependent mean and variance which are obtained by traversing HMM decision trees as follows:

$$\begin{aligned}&\mu_l(\mathrm{s}_\mathrm{t}) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})\mu_l^\mathrm{i},\\ &\sigma_l^2(\mathrm{s}_\mathrm{t}) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})\sigma_l^{\mathrm{i}^2},\\ &\mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t}) \stackrel{\text{def}}{=} \begin{cases}1 \text{ if state } \mathrm{s}_\mathrm{t} \in \mathrm{i}^{\mathrm{th}} \text{ cluster}\\ 0 \text{ if state } \mathrm{s}_\mathrm{t} \notin \mathrm{i}^{\mathrm{th}} \text{ cluster}\end{cases},\end{aligned} \quad (5)$$

where I is termed the total number of clusters, μ_l^i and $\sigma_l^{\mathrm{i}^2}$ represent the l-th dimension of mean and variance trained for cluster i, and $\mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})$ is the indicator function of the i-th cluster. Note that the duration and excitation modeling methods of GCRF and HMM are supposed to be the same in this study; hence, they are not shown in Figure 2.

3.2. GCRF probability distribution

Having described the HMM distribution, the goal of this subsection is to investigate the probability distribution factorized by GCRF graphical model. Hammersley-Cliffort's theorem implies the following equality for the graphical model shown in Figure 2 (b).

$$\mathrm{P}_{\mathrm{GCRF}}(\mathrm{o}|\mathrm{s};\lambda) = \frac{1}{\mathrm{Z}(\mathrm{s};\lambda)}\prod_{\mathrm{t}=1}^{\mathrm{T}}\prod_{l=1}^{\mathrm{L}}\Psi_{lt}(\tilde{\mathrm{o}}_{lt},\mathrm{s}_\mathrm{t};\lambda), \quad (6)$$

where λ is the set of all GCRF parameters and $\tilde{\mathrm{o}}_{lt}$ is a two-dimensional vector defined as $\tilde{\mathrm{o}}_{lt} = \left[\mathrm{o}_{l(\mathrm{t}-1)}, \mathrm{o}_{lt}\right]^T$. All other employed notations were described in previous sections. In accordance with the Gaussian potential function of HMM, this paper assumes that the GCRF partition function, Ψ_{lt}, is formulated by Eq. (7) which is also a two-dimensional Gaussian function with parameters $\mathrm{H}_{lt}(\mathrm{s}_\mathrm{t})$ and $\mathrm{u}_{lt}(\mathrm{s}_\mathrm{t})$.

$$\Psi_{lt}(\tilde{\mathrm{o}}_{lt},\mathrm{s}_\mathrm{t};\lambda) \stackrel{\text{def}}{=} \exp\left\{-\frac{1}{2}(\tilde{\mathrm{o}}_{lt}^T\mathrm{H}_{lt}(\mathrm{s}_\mathrm{t})\tilde{\mathrm{o}}_{lt} + \mathrm{u}_{lt}(\mathrm{s}_\mathrm{t})^T\tilde{\mathrm{o}}_{lt})\right\}. \quad (7)$$

The parameters H_{lt} and u_{lt} determine mean vector and covariance matrix of this Gaussian equation as $-0.5\mathrm{H}_{lt}^{-1}\mathrm{u}_{lt}$ and H_{lt}^{-1} respectively. Note that all constant parts of the Gaussian function (the parts which are independent of observation features) are intentionally eliminated, since they have no influence on the final distribution. Moreover, in Eq. (7), Gaussian function is written in terms of its symmetric precision matrix H_{lt} instead of its covariance matrix. It is mainly due to the fact that the distribution is computed through multiplying the Gaussian potential functions and multiplying many Gaussian functions expressed by the precision matrixes leads to a much simpler equation.

In contrast to the conventional HMM that uses two values (i.e. mean and variance) to parameterize the distribution of each state in a certain dimension, GCRF defines a 2-by-2 symmetric precision matrix H_{lt} and a 2-dimentional vector u_{lt} for each dimension of a state. It means that the total number of parameters in GCRF is 2.5 times the number of HMM parameters.

Similar to HMM, it is also possible to cluster GCRF states. Suppose H_l^i and u_l^i denote the parameters of i-th cluster, $\mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})$ is the indicator function of i-th cluster and I is the number of clusters. In this case, the state parameters of context-dependent GCRF can be expressed by:

$$\begin{aligned}&\mathrm{H}_{lt}(\mathrm{s}_\mathrm{t}) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})\mathrm{H}_l^\mathrm{i},\\ &\mathrm{u}_{lt}(\mathrm{s}_\mathrm{t}) = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})\mathrm{u}_l^\mathrm{i},\end{aligned} \quad (8)$$

Note that $\mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})$ can also represent an overlapped contextual cluster. For example it is possible to define multiple decision trees for $\mathrm{f}_\mathrm{i}(\mathrm{s}_\mathrm{t})$, similar to [56, 58, 59], we

also can exploit decision trees with multiple questions in each intermediate node, similar to [57], or define it heuristically [59].

By considering the potential function expressed by Eq. (7) and according to the fundamental theorem of Hammersley and Cliffort the final distribution is given by:

$$P_{GCRF}(o|s;\lambda) = \frac{1}{Z(s;\lambda)} \prod_{l=1}^{L} \exp\left\{-\frac{1}{2}(o_l^T H_l(s) o_l + u_l(s)^T o_l)\right\}, \quad (9)$$

where $o_l = [o_{l1}, o_{l2}, \dots, o_{lT}]^T$, $s = [s_1, s_2, \dots, s_T]^T$, and model parameters are indicated by an L-by-L band-diagonal precision matrix $H_l(s)$ and an L dimentional vector $u_l(s)$. The parameters H_l and u_l are calculated as a sum of overlapping local contributions, H_{lt} and u_{lt}, where successive local contributions are functions of the state at successive frames. Schematically,

$$H_l(s)=\begin{bmatrix} \ddots & & & \\ & H_{l1}(s_1) & & \\ & & H_{l2}(s_2) & \\ & & & H_{l3}(s_3) \\ & & & & \ddots \\ & & & & & H_{lT}(s_T) \end{bmatrix}, \quad u_l(s)=\begin{bmatrix} \vdots \\ u_{l1}(s_1) \\ u_{l2}(s_2) \\ \vdots \\ u_{lT}(s_T) \end{bmatrix} \quad (10)$$

Note that superscript T and subscript T denote the transpose matrix operation and the total number of frames, respectively. Also, Z is the partition function that can be calculated through the integral of Eq. (2). Selecting the potential functions of Eq. (7) makes it possible to find a closed formula for the partition function as:

$$Z(s;\lambda) = (2\pi)^{\frac{LT}{2}} \prod_{l=1}^{L} (\det(H_l^{-1}))^{\frac{1}{2}} \exp\left(\frac{1}{8} u_l^T H_l^{-1} u_l\right). \quad (11)$$

In sum, GCRF handles different dimensions of observation vectors independently. Each observation is simply expressed by a multivariate Gaussian distribution. The mean vector and covariance matrix of the l-th dimension are defined as $-0.5H_l^{-1}u_l$ and H_l^{-1}; where, H_l and u_l are calculated as a sum of multiple overlapping local contributions as it is shown in Eq. (10).

In order to have a more clear and straightforward equations, this paper just covered a simple GCRF capturing the dependencies of two adjacent frames, while in general GCRF is also able to model longer time-dependencies (e.g. the dependencies of three successive frames). To this end, we need to define a parameter J as a depth or order of GCRF.

Then each potential function becomes a function of J successive frames. More precisely, $\tilde{o}_{lt}$ in Eq. (6) has to be defined as $\tilde{o}_{lt} = \left[o_{l(t-J+1)}, \dots, o_{l(t-1)}, o_{lt}\right]^T$. Final distribution of this GCRF with order J is equal to the Eq. (10) in which the Gaussian parameters $H_l(s)$ and $u_l(s)$ have to be computed as an overlapped summation of many J-by-J matrixes and J-dimentional vectors.

An interesting point is that conventional HMM can be considered as a first order GCRF that uses decision tree-based context clusters.

4. GCRF Based Speech Synthesis

Overall architecture of the proposed GCRF-based speech synthesis system is shown in figure 3. Due to the use of GCRF, statistical modeling and parameter generation algorithm are different from classical HMM-based synthesis. Also, Viterbi-type training is selected instead of the *expectation maximization (EM)* approach to handle unknown state indexes. Viterbi-type training requires less computational resources in contrast to EM; thus, it is a better selection for the computationally expensive training procedure of GCRF.

Model training is performed through two successive steps. In the first step, spectral features are trained using a GCRF-based modeling scheme. Other features including excitation and duration are modeled in the second step exploiting context-dependent HMM structure. However, parameter estimation in both GCRF and HMM requires state occupation probabilities or at least state boundaries to handle the nonstationarity of speech trajectories. Only spectral features are involved in determining these state boundaries. To find an efficient set of state boundaries, first, an initial set of boundaries is borrowed from a classical HMM trained on spectral features. Next, these initial boundaries are contributed to parameter estimation of GCRF.

Trained GCRF is then employed to update state boundaries. After each GCRF estimation or Viterbi decoding, likelihood measure computed for spectral parameters is increased. This procedure is repeated until this increase falls below a threshold. Final state boundaries are fed into the HMM-based excitation and duration modeling module. It should be noted that Viterbi decoding is performed completely according to the *conditional random field (CRF)* Viterbi [61]. Three important blocks in the architecture, including parameter generation algorithm, GCRF modeling and Viterbi state decoding are described in the following subsections.

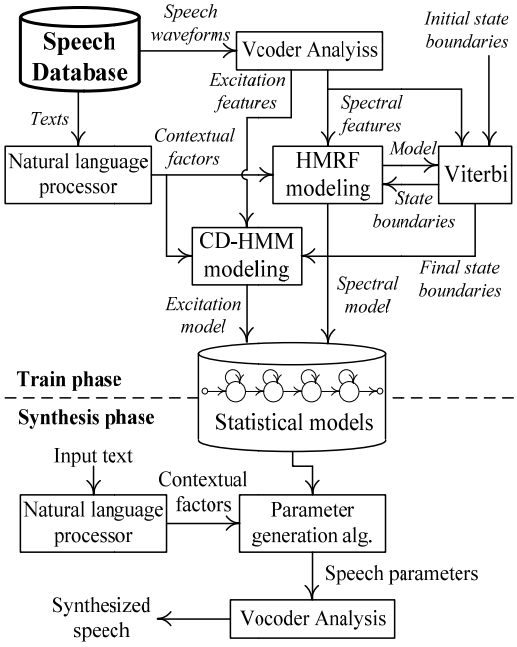


Figure 3. Block diagram of the proposed GCRF-based system

4.1. GCRF Parameter Generation Algorithm

As described before, state duration and excitation parameters are modeled using HMM-based modeling; therefore, these parameters are generated in accordance with standard parameter generation algorithms [9, 10].

This subsection, for a given GCRF-based spectral model, derives an algorithm to estimate the best spectral parameters ($\hat{\mathrm{c}}$) by optimizing the log-likelihood criteria, i.e.

$$\hat{\mathrm{c}} = \mathrm{argmax}_{\mathrm{c}}\, \mathcal{L}(\mathrm{c}, \mathrm{s}; \lambda), \tag{12}$$

where c represents the static spectral trajectory and $\mathcal{L}$ denotes the log-likelihood measure that is computed by taking the logarithm of GCRF distribution given in Eq. (9) as:

$$\mathcal{L}(\mathrm{c}, \mathrm{s}; \lambda) = \sum_{l=1}^{\mathrm{L}} -\frac{1}{2}(\mathrm{o}_l^T \mathrm{H}_l(\mathrm{s})\mathrm{o}_l + \mathrm{u}_l(\mathrm{s})^T \mathrm{o}_l) - \log \mathrm{Z}(\mathrm{s}; \lambda). \tag{13}$$

This log-likelihood function can also be arranged in a unique matrix form:

$$\begin{aligned} &\mathcal{L}(\mathrm{c}, \mathrm{s}; \lambda) = -\frac{1}{2}(\mathrm{o}^T \mathrm{H}(\mathrm{s})\mathrm{o} + \mathrm{u}(\mathrm{s})^T \mathrm{o}) - \log \mathrm{Z}(\mathrm{s}; \lambda), \\ &\mathrm{o} = [\mathrm{o}_1^T, \mathrm{o}_2^T, \dots, \mathrm{o}_{\mathrm{L}}^T]^T, \mathrm{u}(\mathrm{s}) = \\ &[\mathrm{u}_1(s)^T, \mathrm{u}_2(s)^T, \dots, \mathrm{u}_{\mathrm{L}}(s)^T]^T, \\ &\mathrm{H}(\mathrm{s}) = \begin{bmatrix} \mathrm{H}_1(\mathrm{s}) & 0 & \cdots & 0 \\ 0 & \mathrm{H}_2(\mathrm{s}) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{H}_{\mathrm{L}}(\mathrm{s}) \end{bmatrix}, \end{aligned} \tag{14}$$

where o contains all observation features; all u parameters are arranged in an $\mathrm{L} \times \mathrm{T}$ dimensional vector, and H is an $(\mathrm{L} \times \mathrm{T})$-by-$(\mathrm{L} \times \mathrm{T})$ band diagonal precision matrix.

In order to generate the optimum spectral features $\hat{\mathrm{c}}$, log-likelihood function has to be written in terms of static features c. In addition, observation vector o, in most cases [9], is assumed to be a linearly transformed version of c, i.e.

$$\mathrm{o}_{\mathrm{LT}\times 1} = \mathrm{W}_{\mathrm{LT}\times(\mathrm{LT}/3)}\mathrm{c}_{(\mathrm{LT}/3)\times 1}, \tag{15}$$

where $\mathrm{c} = \left[\mathrm{c}_1^T, \mathrm{c}_2^T, \dots, \mathrm{c}_{\mathrm{L}/3}^T\right]^T$, and $\mathrm{c}_l = [\mathrm{c}_{l1}, \mathrm{c}_{l2}, \dots, \mathrm{c}_{l\mathrm{T}}]^T$.

According to the above equations, the final parameter generation algorithm is achieved by solving the following unconstraint optimization problem:

$$\hat{\mathrm{c}} = \mathrm{argmin}_{\mathrm{c}}\, (\mathrm{c}^{\mathrm{T}} \mathrm{W}^T \mathrm{H}(\mathrm{s})\mathrm{W}\mathrm{c} + \mathrm{u}(\mathrm{s})^T \mathrm{W}\mathrm{c}). \tag{16}$$

The result is obtained by computing the partial derivatives of Eq. (16) with respect to c and setting it to zero. This procedure results in the following system of equations:

$$(\mathrm{W}^T \mathrm{H}(\mathrm{s})\mathrm{W})\hat{\mathrm{c}} = -\frac{1}{2}\mathrm{W}^T \mathrm{u}(\mathrm{s}), \tag{17}$$

which can be solved efficiently using *Cholesky decomposition*, since $\mathrm{W}^T \mathrm{H}(\mathrm{s})\mathrm{W}$ is symmetric and positive definite.

4.2. GCRF Parameter Estimation

This section discusses the training procedure of model parameters λ. In model training, we are given a set of N iid training sentences $\{\mathrm{O}^{\mathrm{n}}, \mathrm{S}^{\mathrm{n}}\}_{\mathrm{n}=1}^{\mathrm{N}}$, and the goal is to estimate the best set of parameters, $\hat{\lambda}$, which maximizes the following log-likelihood measure:

$$\begin{aligned} &\hat{\lambda} = \mathrm{argmin}_{\lambda} \mathcal{L}_{\mathrm{ML}}(\lambda), \\ &\mathcal{L}_{\mathrm{ML}}(\lambda) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{l=1}^{\mathrm{L}} \left\{ -\frac{1}{2}\left(\mathrm{o}_l^{\mathrm{n}T} \mathrm{H}_l(\mathrm{s}^{\mathrm{n}})\mathrm{o}_l^{\mathrm{n}} + \mathrm{u}_l(\mathrm{s}^{\mathrm{n}})^T \mathrm{o}_l^{\mathrm{n}}\right) - \log \mathrm{Z}(\mathrm{s}^{\mathrm{n}}; \lambda) \right\}. \end{aligned} \tag{18}$$

Therefore, the *maximum likelihood (ML)* criterion is optimized during the training procedure. Model likelihood is denoted by $\mathcal{L}_{\mathrm{ML}}$ and n is an index defined for the utterance number. Replacing $\mathrm{Z}(\mathrm{s}; \lambda)$ with Eq. (11) gives:

$$\begin{aligned} \mathcal{L}_{\mathrm{ML}}(\lambda) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{l=1}^{\mathrm{L}} -\frac{1}{2} \Big\{ &\mathrm{o}_l^{\mathrm{n}T} \mathrm{H}_l(\mathrm{s}^{\mathrm{n}})\mathrm{o}_l^{\mathrm{n}} + \mathrm{u}_l(\mathrm{s}^{\mathrm{n}})^T \mathrm{o}_l^{\mathrm{n}} - \log \det\left(\mathrm{H}_l(\mathrm{s}^{\mathrm{n}})\right) + \\ &\frac{1}{4}\mathrm{u}_l(\mathrm{s}^{\mathrm{n}})^T \mathrm{H}_l(\mathrm{s}^{\mathrm{n}})^{-1} \mathrm{u}_l(\mathrm{s}^{\mathrm{n}}) \Big\}. \end{aligned} \tag{19}$$

As it is described in Eq. (8), GCRF model parameters λ include a set of 2-by-2 matrixes $\mathrm{H}_l^{\mathrm{i}}$ and a set of 2-dimensional vectors $\mathrm{u}_l^{\mathrm{i}}$ that are defined for every dimension $1 \le l \le \mathrm{L}$ and every decision tree cluster $1 \le i \le \mathrm{I}$. Therefore, our goal is to find the best values for all $\mathrm{H}_l^{\mathrm{i}}$ and $\mathrm{u}_l^{\mathrm{i}}$. To achieve this goal, partial derivatives of $\mathcal{L}_{\mathrm{ML}}$ with respect to these parameters has to be computed and then by setting them to zero the optimum parameters will be estimated.

Partial derivatives are given by Eq. (20).

$$\begin{aligned} &\frac{\partial \mathcal{L}_{\mathrm{ML}}}{\partial \mathrm{u}_l^{\mathrm{i}}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} -\frac{1}{2}\mathrm{F}_{\mathrm{i}}(\mathrm{s}^{\mathrm{n}})^T \left[\frac{1}{2}\mathrm{H}_l(\mathrm{s}^{\mathrm{n}})^{-1}\mathrm{u}_l(\mathrm{s}^{\mathrm{n}}) + \mathrm{o}_l^{\mathrm{n}}\right], \\ &\frac{\partial \mathcal{L}_{\mathrm{ML}}}{\partial \mathrm{H}_l^{\mathrm{i}}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} -\frac{1}{2}\mathrm{F}_{\mathrm{i}}(\mathrm{s}^{\mathrm{n}})^T \Big[\mathrm{o}_l^{\mathrm{n}}\mathrm{o}_l^{\mathrm{n}T} - \mathrm{H}_l(\mathrm{s}^{\mathrm{n}})^{-1} - \\ &\frac{1}{4}\mathrm{H}_l(\mathrm{s}^{\mathrm{n}})^{-1}\mathrm{u}_l(\mathrm{s}^{\mathrm{n}})\mathrm{u}_l(\mathrm{s}^{\mathrm{n}})^T\mathrm{H}_l(\mathrm{s}^{\mathrm{n}})^{-1}\Big] \mathrm{F}_{\mathrm{i}}(\mathrm{s}^{\mathrm{n}}). \end{aligned} \tag{20}$$

F is a 2-by-T^{n} binary matrix defined as:

$$\begin{aligned} \mathrm{F}_{\mathrm{i}}(\mathrm{s}^{\mathrm{n}}) &= \frac{\partial \mathrm{u}_l(\mathrm{s}^{\mathrm{n}})}{\partial \mathrm{u}_l^{\mathrm{i}}} \\ &= \begin{bmatrix} 0 & \mathrm{f}_{\mathrm{i}}(\mathrm{s}_1^{\mathrm{n}}) & \mathrm{f}_{\mathrm{i}}(\mathrm{s}_2^{\mathrm{n}}) & \cdots & \mathrm{f}_{\mathrm{i}}(\mathrm{s}_{\mathrm{T}^{\mathrm{n}}-1}^{\mathrm{n}}) \\ \mathrm{f}_{\mathrm{i}}(\mathrm{s}_1^{\mathrm{n}}) & \mathrm{f}_{\mathrm{i}}(\mathrm{s}_2^{\mathrm{n}}) & \mathrm{f}_{\mathrm{i}}(\mathrm{s}_3^{\mathrm{n}}) & \cdots & \mathrm{f}_{\mathrm{i}}(\mathrm{s}_{\mathrm{T}^{\mathrm{n}}}^{\mathrm{n}}) \end{bmatrix}^T \end{aligned} \tag{21}$$

where T^{n} denotes the total number of frames in the n-th utterance, and $\mathrm{f}_{\mathrm{i}}(\mathrm{s}_{\mathrm{t}}^{\mathrm{n}})$ determines whether the t-th frame of utterance n belongs to the cluster i or not. $\mathrm{f}_{\mathrm{i}}(\mathrm{s}_{\mathrm{t}}^{\mathrm{n}})$ is defined precisely in Eq. (5).

Setting all partial derivatives of Eq. (20) to zero, leads to a system of equations that has not a solution in closed formula and has to be solved iteratively. This paper proposes applying the well-known and efficient *Broyden-Fletcher-Goldfarb-Shanno (BFGS)* algorithm [70] in order to solve the above system of equations.

BFGS needs the first partial derivatives of the likelihood function given by Eq. (20).

4.3. GCRF Viterbi Algorithm

In this section, we will briefly explain the Viterbi algorithm employed to determine state boundaries in GCRF. Viterbi is a dynamic programming algorithm for finding the most probable sequence of hidden states (Viterbi path). Given an observation vector o, our goal is to find $\hat{s}$ such that:

$$\hat{\mathrm{s}} = \mathrm{argmax}_{\mathrm{s}} \mathrm{P}(\mathrm{s}|\mathrm{o};\lambda) = \mathrm{argmax}_{\mathrm{s}} \frac{\mathrm{P}(\mathrm{o},\mathrm{s};\lambda)}{\mathrm{P}(\mathrm{o};\lambda)}, \quad (22)$$

where $\mathrm{P}(\mathrm{o};\lambda)$ is independent of state sequence s and can be eliminated from maximization problem; therefore,

$$\hat{\mathrm{s}} = \mathrm{argmax}_{\mathrm{s}} \mathrm{P}(\mathrm{s}|\mathrm{o};\lambda) = \mathrm{argmax}_{\mathrm{s}} \mathrm{P}(\mathrm{o},\mathrm{s};\lambda). \quad (23)$$

According to the architecture given in figure 3, state boundaries are determined by considering spectral features; thus, in equations (22) and (23), o represents the spectral features and other features are not incorporated in that. Moreover, based on the prime assumption of this paper, ordered pair of (o,s) forms an MRF. The distribution of this MRF is factorized by Eq. (6); therefore, $\mathrm{P}(\mathrm{o},\mathrm{s};\lambda)$ can be written as:

$$\mathrm{P}(\mathrm{o},\mathrm{s};\lambda) = \frac{1}{\mathrm{Z}} \prod_{\mathrm{t}=1}^{\mathrm{T}} \prod_{l=1}^{\mathrm{L}} \Psi_{lt}(\tilde{\mathrm{o}}_{lt}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (24)$$

where the partition function, Z, is independent of both o and s, and is calculated through Eq. (2). Replacing $\mathrm{P}(\mathrm{o},\mathrm{s};\lambda)$ with Eq. (24) and eliminating the constant value Z, leads to

$$\hat{\mathrm{s}} = \mathrm{argmax}_{\mathrm{s}} \prod_{\mathrm{t}=1}^{\mathrm{T}} \prod_{l=1}^{\mathrm{L}} \Psi_{lt}(\tilde{\mathrm{o}}_{lt}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (25)$$

For the sake of clarity in next equations, we also define another potential function ϕ_{t} by multiplying Ψ_{lt} over all dimensions,

$$\phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda) = \prod_{l=1}^{\mathrm{L}} \Psi_{lt}(\tilde{\mathrm{o}}_{lt}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (26)$$

This new potential function simplifies our problem as:

$$\hat{\mathrm{s}} = \mathrm{argmax}_{\mathrm{s}} \prod_{\mathrm{t}=1}^{\mathrm{T}} \phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (27)$$

Now, suppose the given utterance is made up of J successive states. For example, if we design a five-state GCRF for each phoneme, this J will be calculated by multiplying the number of phonemes by five. Also, assume the last frame of j-th state is denoted by τ_{j} and $\tau_0 = 1$. Then the optimization problem can be rewritten as the following equation:

$$\hat{\mathrm{s}} = \mathrm{argmax}_{\mathrm{s}} \prod_{\mathrm{j}=1}^{\mathrm{J}} \prod_{\mathrm{t}=\tau_{\mathrm{j}-1}}^{\tau_{\mathrm{j}}} \phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (28)$$

Similar to HMM Viterbi, in order to break down the complex problem of state labeling into simpler sub-problems, we need to define two auxiliary variables δ, τ^*. δ is a function of two variables t and j, and computes the maximum value of the above optimization function, when the number of states is j and the utterance has just t frames.

$$\delta_j(t) = \max_{\tau_1,\tau_2,\ldots,\tau_{j-1}} \left(\prod_{\mathrm{j}=1}^{j-1} \prod_{\mathrm{t}=\tau_{\mathrm{j}-1}}^{\tau_{\mathrm{j}}} \phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda) \right) \times \prod_{\mathrm{t}=\tau_{j-1}}^{t} \phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (29)$$

This auxiliary variable can be simply computed through the following recursion:

$$\delta_j(t) = \max_{\tau_{j-1}} \delta_{j-1}(\tau_{j-1}) \times \prod_{\mathrm{t}=\tau_{j-1}}^{t} \phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (30)$$

Other auxiliary variable $\tau_j^*(t)$ stores the frame index that maximizes the above equation. In other words,

$$\tau_j^*(t) = \mathrm{argmax}_{\tau_{j-1}} \delta_{j-1}(\tau_{j-1}) \times \prod_{\mathrm{t}=\tau_{j-1}}^{t} \phi_{\mathrm{t}}(\mathrm{o}, \mathrm{s}_{\mathrm{t}};\lambda). \quad (31)$$

GCRF Viterbi is performed through two steps, namely forward and backward steps. In the forward step, $\delta_j(t)$ and $\tau_j^*(t)$ are computed for all values of $1 \le j \le \mathrm{J}$ and $1 \le t \le \mathrm{T}$ using the above recursions. Thereafter the backward step finds the optimum state boundaries. Clearly, the optimum and only possible value for τ_{J} is the total number of frames, i.e. T. Thus, this τ_{J} is used for initializing backward step. Other optimum state boundaries can also be computed through the following backward iterations:

$$\tau_{\mathrm{j}} = \tau_{\mathrm{j}+1}^*(\tau_{\mathrm{j}+1}). \quad (32)$$

It should be noted that the auxiliary variable $\delta_j(t)$ gets smaller and smaller values by increasing the frame index t. As a result, for long sentences with a great number of states, the δ may become so small that cannot be stored correctly as a positive value in computers. In this case, the forward and backward procedure fails and returns an incorrect Viterbi path. As a solution to this problem, we can optimize the logarithm of $\mathrm{P}(\mathrm{s}|\mathrm{o};\lambda)$ in Eq. (23) and derive all the above equations for this new objective function.

5. Experiments

This section aims to evaluate GCRF-based acoustic modeling in contrast to the classical HMM-based method. To achieve this goal, we have conducted two sets of evaluations. The first one examines the performance of GCRF with some heuristic and overlapped context clusters. The second experiment investigates the impact of decision tree-based context clusters on the quality of GCRF spectral modeling. The results of these two sets of experiments are presented in subsections 5.3 and 5.4. Also, the next subsection provides the details of experimental conditions.

5.1. Experimental Conditions

Experiments were wholly conducted on a standard Persian speech database designed originally for single-speaker speech synthesis application [59]. This database is prepared completely in accordance with the CMU Arctic speech databases [71] and has been carefully recorded under studio conditions. It consists of approximately 1000 phonetically balanced Persian utterances with an average duration of eight seconds for each utterance. Furthermore, it covers most frequent Persian words, most frequent syllables, all bi-letters, and all possible bi-phones combinations. The recorded waveforms are packaged with contextual information required for building a single-speaker speech synthesis. The contextual information includes phoneme labels and

boundaries, syllable and word boundaries, part-of-speech (POS) and "Ezafe"[1] tags, ToBI accentual and intonational phrase information, and the stress level of all syllables. More than 64 segmental and suprasegmental contextual factors are extracted for this database.

All speech signals are sampled at 16 kHz sampling rate, windowed by a 25-ms Blackman window with 5-ms shift. Spectral and excitation features were extracted by STRAIGHT vocoder [5]. These acoustic features include 25 mel-cepstral coefficients [68, 69], five band pass aperiodicity parameters [72] related to five sub-bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz), and a fundamental frequency. Above features along with their derivatives were used as the speech parameters in implemented systems.

A 5-state multi-stream single-mixture left-to-right with no skip path HMM was trained as the baseline system using the publicly available HTS toolkit [73]. To handle undefined F0 values in unvoiced frames, F0 stream was modeled by a multi-space probability distribution [20]. Also, state duration probabilities are explicitly represented by a Gaussian distribution using hidden semi-Markov modeling framework [21]. Prevalent maximum likelihood-based decision tree construction algorithm used to tie HMM states, and MDL criterion was used to determine the size of the decision trees. Similar to [49], MDL tuning factor was set to be $1/3$ for all decision trees. Additionally, the first algorithm proposed in [9] which maximizes output probability was adopted to generate parameters from trained models.

During HMM training for the baseline synthesis system, the stream weight for the aperiodicity was set to zero. Thus, the forward-backward procedure depended only on the spectral and F0 features. In other words, model parameters of aperiodicity components were trained in the normal way, but they do not contribute to the calculation of forward and backward variables.

Experiments were conducted on 4 different training sets with 50, 100, 200, and 400 utterances. Additionally, a fixed set of 200 utterances, not included in the training sets, was used for testing.

5.2. Employed Contextual Factors

In our experiments, contextual factors contained several levels, including phonetic, syllable, word, phrase and sentence level factors. In each of these levels both general and detailed factors were taken into account. Features such as phoneme identity, syllable stress pattern or word part of speech tag are examples of general features and a question like the position of the current phoneme is a sample of detailed one. Specific information with regard to contextual features is presented in this subsection.

- Phonetic-level features
  - ✓ Phoneme identity of the two preceding, preceding, current, succeeding, and two succeeding phoneme.
  - ✓ Position of the current phoneme in the current syllable (forward and backward).
  - ✓ Whether this phoneme is "Ezafe" or not.
- Syllable-level features
  - ✓ Stress level of this syllable (5 different stress levels are defined for our speech database).
  - ✓ Position of the current syllable in the current word and phrase (forward and backward).
  - ✓ Type of the current syllable (syllables in Persian language are structured as CV, CVC, or CVCC, where C and V denote consonants and vowels, respectively).
  - ✓ Number of the stressed syllables before and after the current syllable in the current phrase.
  - ✓ Number of syllables from the previous stressed syllable to the current syllable.
  - ✓ Vowel identity of the current syllable.
- Word-level features
  - ✓ Part of speech (POS) tag of the preceding, current and succeeding word.
  - ✓ Position of the current word in the current sentence (forward and backward).
  - ✓ Whether the current word contains "Ezafe" or not.
  - ✓ Whether this word is the last word in the sentence or not.
- Phrase-level features
  - ✓ Number of syllables in the preceding, current, and succeeding phrase.
  - ✓ Position of the current phrase in the current sentence (forward and backward).
- Sentence-level features
  - ✓ Number of syllables, words, and phrases in the current sentence.
  - ✓ Type of the current sentence.

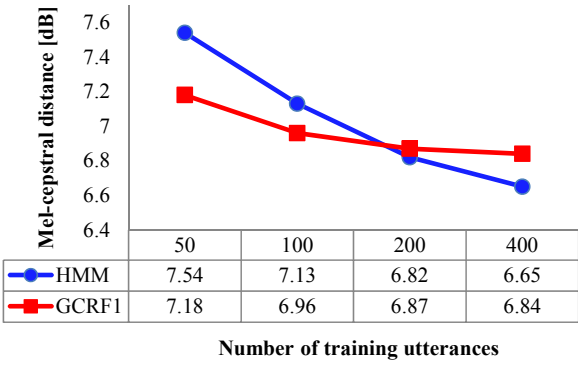
5.3. Evaluation Results of GCRF with Heuristic Contextual Regions

In this experiment, a synthesis system named GCRF1 was developed based on the proposed approach. This system adopted 150 highly overlapped contextual regions that were designed carefully. The indicator functions $f_i(s)$ in equations (20) and (22) were defined based on these 150 contextual regions. To define the indicator functions, first, a set of 64 initial contextual factors were extracted for each segment (phoneme) of the Persian database. Some of these initial factors are mentioned in the previous subsection. Then, from these contextual factors, a set of approximately 8000 contextual questions were designed and the baseline HMM-based system was trained using them. Each question can form two regions; therefore, these 8000 questions can be converted to 16000 regions. For each stream of GCRF1 a set including 150 contextual regions that seem to be more important for that stream were selected and GCRF1 was trained using them. Regions of GCRF1 were selected based on the linguistic knowledge of Persian language.

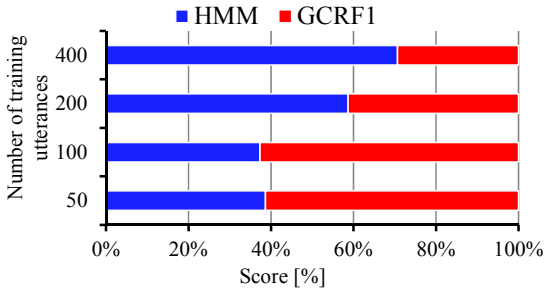
Both subjective and objective tests have been conducted to assess the performance of the proposed GCRF-based spectral modeling in contrast to the conventional HMM-based method. As a relevant objective measure, average mel-cepstral distortion (MCD) [74, 75] between generated and natural spectral trajectories was calculated. This MCD measure between two corresponding mel-cepstral vectors is defined by the following equation:

$$\mathrm{MCD} = \frac{10}{\ln(10)} \sqrt{2 \sum_{i=1}^{25} \left(mc_i^t - mc_i^p\right)^2}, \tag{33}$$

Where mc_i^t and mc_i^p are respectively termed target and predicted i-th mel-cepstral component.

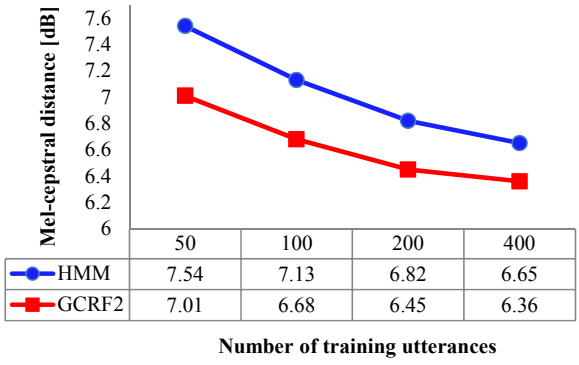


(a)

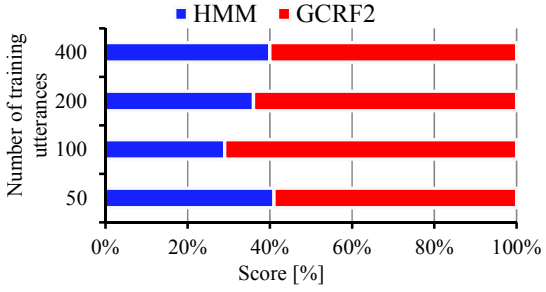


(b)

Figure 4. evaluation results comparing GCRF1 with HMM using (a) MCD measure (b) preference score.



(a)



(b)

Figure 5. the results of objective and subjective tests evaluating the performance of GCRF2 in contrast to the HMM. (a) MCD measure result, (b) preference score test.

For calculating the average mel-cepstral distance, the length of natural and generated trajectories has to be exactly the same. In order to equalize the length of generated and natural trajectories, we first conducted a Viterbi algorithm to obtain the most likely state durations of the natural trajectories and then GCRF trajectories were generated in accordance with the obtained durations.

We also employed preference score measure [76] to compare the proposed and HMM-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.

Figures 4 (a) and 4 (b) show the results of objective and subjective evaluations. Remarkably, the GCRF1 system is noticed to be of a great interest when the training data is limited. The superiority of GCRF1 over HMM is clear in the training sets containing 50 and 100 utterances. Gradually, as the number of utterances in the training set increases, HMM surpasses GCRF1. It is mainly because GCRF with some heuristic contextual regions is unable to balance the model complexity with the size of training data. In other words, for larger databases, it is expected to have a model with larger number of parameters, but GCRF1 applies the same number of parameters for all databases. Therefore, we need to think about a procedure that balances the GCRF model complexity against the size of training data. Next section gives a naïve idea to overcome this issue.

5.4. Evaluation Results of GCRF with Decision Tree-Based Contextual Clusters

A fundamental solution to the problem of aligning model complexity with the amount of training data is to conduct clustering methods such as decision tree construction algorithms over GCRF states. However, clustering GCRF states using an optimum structure (e.g. maximum likelihood clustering [60]) leads to an extremely complicated procedure which is computationally impossible to implement. Therefore, this section proposes borrowing the decision tree-based clusters of conventional HMM. GCRF using HMM contextual clusters is named GCRF2 in this subsection.

Both objective and subjective tests were conducted in this subsection as well. MCD measure and preference score are selected for the objective and subjective evaluations. For the preference score, 10 native participants were asked to listen to 30 randomly chosen pairs of synthesized speech samples generated by two different systems (HMM and GCRF2). Figures 5 (a) and 5 (b) show the results of objective and subjective tests. As it can be seen in the figures, both objective and subjective tests confirm the superiority of the proposed method over the conventional HMM in all databases.

6. Conclusion

The goal of this paper is to address the important problem of state independence assumption in HMM and propose a new spectral modeling approach that relaxes this inaccurate assumption using the capabilities of Gaussian conditional random fields. The proposed GCRF-based spectral modeling has also been incorporated in a new statistical parametric speech synthesis system. In the training phase of this new speech synthesis system, a Viterbi-type framework that maximizes the likelihood measure has been developed. Synthesis phase has been performed through a maximum output probability parameter generation algorithm followed by a STRAIGHT vocoder. Additionally, two methods have

been proposed to capture the context dependencies in the proposed GCRF-based spectral modeling. The first one was based on some heuristic contextual regions. Conducted subjective and objective evaluations confirmed that the heuristic contextual regions are just effective for small databases. The second method has been designed based on decision tree-based regions which are noticed to be effective in all databases. Reported subjective and objective tests prove the effectiveness of the second method of capturing contextual dependencies in all databases.

All of the above advantages are achieved at the expense of extremely high computational complexity parameter estimation algorithm. Indeed, the proposed BFGS optimization procedure requires large processing resources which prevent us to derive and implement an optimum context clustering technique. Our future works are devoted to: (i) introduce a more efficient parameter estimation algorithm requiring less computational resources. (ii) incorporate the proposed algorithm in an optimum context clustering technique and express the relationship between acoustic features and contextual factors directly based on GCRF modeling scheme.

References

[1] S. King, "An Introduction to Statistical Parametric Speech Synthesis," *Journal of Sadhana*, vol. 36, no. 5, pp. 837-852, 2011.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Journal of Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[3] A. W. Black, H. Zen, and K. Tokuda, "Statistical Parametric Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 1219-1229, 2007.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed Excitation for HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Speech Communication and Technology*, pp. 2263-2266, 2001.

[5] H. Kawahara, I. Masuda-Katsuse, and A. 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," *Journal of Speech Communication*, vol. 27, no. 3, pp. 187-207, 1999.

[6] T. Drugman, and T. Dutoit, "The Deterministic plus Stochastic Model of the Residual Signal and its Applications," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 968-981, 2012.

[7] T. Drugman, G. Wilfart, and T. Dutoit, "A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis," *Proc, IEEE Int'l Conf. Speech Communication and Technology*, pp. 1779-1782, 2009.

[8] Y. Stylianou, "Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21-29, 2001.

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, vol. 3, no. 2, pp. 1315-1318, 2000.

[10] T. Toda, and K. Tokuda, "Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis," *IEEE Trans. Information and Systems*, vol. 9, no. 5, pp. 816-824, 2007.

[11] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter Generation Methods with Rich Context Models for High-quality and Flexible Text-to-speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 135-138, 2013.

[12] M. Shannon, and W. Byrne, "Fast, Low-artifact Speech Synthesis Considering Global Variance," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 7869-7873, 2013.

[13] J. Yamagishi, and T. Kobayashi, "Average-voice-based Speech Synthesis Using HSMM-based Speaker Adaptation and Adaptive Training," *IEEE Trans. Information and Systems*, vol. 90, no. 2, pp. 533-543, 2007.

[14] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust Speaker-adaptive HMM-based Text-to-speech Synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208-1230, 2009.

[15] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66-83, 2009.

[16] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State Mapping based Method for Cross-lingual Speaker Adaptation in HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 528-531, 2009.

[17] H. Liang, J. Dines, and L. Saheer, "A Comparison of Supervised and Unsupervised Cross-lingual Speaker Adaptation Approaches for HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 4598-4601, 2010.

[18] M. Gibson, T. Hirsimaki, R. Karhila, M. Kurimo, and W. Byrne, "Unsupervised Cross-lingual Speaker Adaptation for HMM-based Speech Synthesis Using Two-pass Decision Tree Construction," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 4642-4645, 2010.

[19] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 581-584, 2008.

[20] H. Zen, T. Keiichi, T. Masuko, T. Kobayashi, and T. Kitamura, "A Hidden Semi-markov Model-based Speech Synthesis System," *IEEE Trans. Information and Systems*, vol. 90, no. 5, pp. 825-834, 2007.

[21] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space Probability Distribution HMM," *IEEE Trans. Information and Systems*, vol. 85, no. 3, pp. 455-464, 2002.

[22] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and T. Keiichi, "The HMM-based Speech Synthesis System (HTS) Version 2.0," *Proc, IEEE Int'l Workshop on Speech Synthesis*, pp. 294-299. 2007.

[23] L. Deng, M. Aksmanovic, X. Sun, and J. Wu, "Speech Recognition Using Hidden Markov Models with Polynomial Regression Functions as Nonstationary States," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 507-520, 1994.

[24] L. Deng, and M. Aksmanovic, "Speaker-independent Phonetic Classification Using Hidden Markov Models with Mixture of Trend Functions," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 4, pp. 319-324, 1997.

[25] H. Gish, and K. Ng, "Parametric Trajectory Models for Speech Recognition," *Proc, IEEE Int'l Conf. Spoken Language Processing*, vol. 1, no. 3, pp. 466-469, 1996.

[26] L. Deng, "A Dynamic, Feature-based Approach to the Interface between Phonology and Phonetics for Speech Modeling and Recognition," *Journal of Speech Communication*, vol. 24, no. 4, pp. 299-323, 1998.

[27] H. Richards, and J. Bridle, "The HDM: a Segmental Hidden Dynamic Model of Co-articulation," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 357-360. 1999.

[28] J. Ma, and L. Deng, "Target-directed Mixture Linear Dynamic Models for Spontaneous Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 1, pp. 47-58, 2004.

[29] J. L. Zhou, F. Seide, and L. Deng, "Co-articulation Modeling by Embedding a Target-directed Hidden Trajectory Model into HMM-model and Training," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 744-747, 2003.

[30] T. Kobayashi, K. Masumitsu, and J. Furuyama, "Partly Hidden Markov Model and its Application to Speech Recognition," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 121-124, 1999.

[31] M. Ostendorf, and S. Roukos, "A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1857-1869, 1989.

[32] M. Russel, "A Segmental HMM for Speech Pattern Modeling," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 499-502, 1993.

[33] M. Gales, and S. Young, *The Theory of Segmental Hidden Markov Models*, Technical Report, Cambridge University, Cambridge, United Kingdom, 1993.

[34] W. Holmes, and M. Russel, "Probabilistic-trajectory Segmental HMMs," *Journal of Computer, Speech and Language*, vol. 13, no. 1, pp. 3-37, 1999.

[35] K. C. Sim, and M. Gales, *Precision Matrix Modeling for Large Vocabulary Continuous Speech Recognition*, Technical Report, Cambridge University, Cambridge, United Kingdom, 2004.

[36] K. C. Sim, and M. Gales, "Temporally Varying Model Parameters for Large Vocabulary Continuous Speech Recognition," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 2137-2140, 2005.

[37] C. Wellekens, "Explicit Correlation in Hidden Markov Model for Speech Recognition," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 383-386, 1987.

[38] P. Brown, *The Acoustic Modeling Problem in Automatic Speech Recognition*, Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, United States, 1987.

[39] K. Paliwal, "Use of Temporal Correlation between Successive Frames in Hidden Markov Model based Speech Recognizer," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 215-218, 1993.

[40] G. Qing, Z. Fang, W. Jian, and W. Wenhu, "A New Method Used in HMM for Modeling Frame Correlation," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 169-172, 1999.

[41] K. Paliwal, "Use of Temporal Correlation between Successive Frames in Hidden Markov Model based Speech Recognizer," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 215-218, 1993.

[42] J. Bilmes, "Buried Markov Models: a Graphical Modeling Approach for Automatic Speech Recognition," *Journal of Computer, Speech and Language*, vol. 17, no. 2, pp. 213-231, 2003.

[43] A. Rosti, and M. Gales, *Switching Linear Dynamical Systems for Speech Recognition*, Technical Report, Cambridge University, Cambridge, United Kingdom, 2003.

[44] L. Lee, H. Attias, L. Deng, and P. Fieguth, "A Multi-modal Variation Approach to Learning and Inference in Switching State Space Models," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 505-508, 2004.

[45] G. Zweig, *Speech Recognition Using Dynamic Bayesian Networks,* Ph.D. Dissertation, University of California, Berkeley, United States, 1998.

[46] H. Zen, T. Keiichi, and T. Kitamura, "Reformulating the HMM as a Trajectory Model by Imposing Explicit Relationships between Static and Dynamic Feature Vector Sequences," *Journal of Computer, Speech and Language*, vol. 21, no. 1, pp. 153-173, 2007.

[47] H. Zen, K. Tokuda, and T. Kitamura, "An Introduction of Trajectory Model into HMM-based Speech Synthesis,"

Proc, IEEE Int'l Workshop on Speech Synthesis, pp. 84-88, 2004.

[48] T. Toda, and S. Young, "Trajectory Training Considering Global Variance for HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 4025-4028, 2009.

[49] M. Shannon, H. Zen, and W. Byrne, "Autoregressive Models for Statistical Parametric Speech Synthesis," *IEEE Trans. Audio, Speech, Language Processing*, vol. 21, no. 3, pp. 587-597, 2013.

[50] C. Quillen, "Autoregressive HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 4021-4024, 2012.

[51] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical Parametric Speech Synthesis based on Gaussian Process Regression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 99, pp. 1-11, 2013.

[52] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modeling," *Proc, IEEE Int'l Workshop on Human Language Technology, Association for Computational Linguistics*, pp. 307-312, 1994.

[53] H. Zen, A. Senior, and M. Schuster, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 7962-7966, 2013.

[54] Z. H. Ling, L. Deng, and D. Yu, "Modeling Spectral Envelopes Using Restricted Boltzmann Machines for Statistical Parametric Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 7825-7829, 2013.

[55] S. Kang, X. Qian, and H. Meng, "Multi-distribution Deep Belief Network for Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 8012-8016, 2013.

[56] S. Takaki, Y. Nankaku, and K. Tokuda, "Spectral Modeling with Contextual Additive Structure for HMM-based Speech Synthesis," *Proc, IEEE Int'l Workshop on Speech Synthesis*, pp. 100-105, 2010.

[57] S. Takaki, Y. Nankaku, and K. Tokuda, "Contextual Partial Additive Structure for HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 7878-7882, 2013.

[58] H. Zen, and N. Braunschweiler, "Context-dependent Additive log f0 Model for HMM-based Speech Synthesis," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 2091-2094, 2009.

[59] S. Khorram, H. Sameti, F. Bahmaninezhad, S. King, and T. Drugman, "Context-dependent Acoustic Modeling based on Hidden Maximum Entropy Model for Statistical Parametric Speech Synthesis," *Journal on Audio, Speech, and Music Processing*, vol. 1, no. 12, pp. 12-24, 2014.

[60] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition,* Ph.D. Dissertation, Cambridge University, Cambridge, United Kingdom, 1995.

[61] C. Sutton, and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning," *Journal of Statistical Relational Learning*, vol. 2, no. 1, pp. 64-76, 2006.

[62] J. Lafferty, M. C. Andrew, and C. P. Fernando, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,* Technical Report, University of Pennsylvania, PA, United States, 2001.

[63] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 1117-1120, 2005.

[64] D. Yu, L. Deng, and A. Acero, "Hidden Conditional Random Field with Distribution Constraints for Phone Classification," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 676-679, 2009.

[65] Y. H., Sung, and D., Jurafsky, "Hidden Conditional Random Fields for Phone Recognition," *Proc, IEEE Int'l Conf. Automatic Speech Recognition and Understanding*, pp. 107-112, 2009.

[66] W. T. Hong, "Speaker Identification Using Hidden Conditional Random Field-based Speaker Models," *IEEE Trans. Machine Learning and Cybernetics*, vol. 6, no. 2, pp. 628-634, 2010.

[67] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic Prosody Prediction and Detection with Conditional Random Field Models," *Proc, IEEE Int'l Conf. Chinese Spoken Language Processing*, pp. 135-138, 2010.

[68] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An Adaptive Algorithm for Mel-cepstral Analysis of Speech," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 137-140, 1992.

[69] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized Cepstral Analysis-a Unified Approach to Speech Spectral Estimation," *Proc, IEEE Int'l Conf. Spoken Language Processing*, pp. 100-106, 1994.

[70] J. Nocedal, and J. W. Stephen, *Numerical Optimization,* Springer, 1999.

[71] J. Kominek, and A. W. Black, "The CMU Arctic Speech Databases," *Proc, IEEE Int'l Workshop on Speech Synthesis*, pp. 200-204, 2004.

[72] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity Extraction and Control Using Mixed Mode Excitation and Group Delay Manipulation for a High Quality Speech Analysis, Modification and Synthesis System STRAIGHT,"

Proc, IEEE Int'l Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, pp. 13-15, 2001.

[73] HMM-based speech synthesis system (HTS), http://hts.sp.nitech.ac.jp/, May 2003.

[74] R. Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," *IEEE Journal of Communications, Computers and Signal Processing*, vol. 1, no. 3, pp. 125-128, 1993.

[75] A. W. Black, "CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling," *Proc, IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, pp. 1762-1765, 2006.

[76] J. Vepa, and S. King, "Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1763-1771, 2006.

Soheil Khorram received the B.Sc. degree from Shahid Bahonar University of Kerman, Iran, in 2006 and the M.Sc. degree from Sharif University of Technology, Tehran, Iran, in 2009. He is currently a Ph.D. student in Artificial Intelligence at Sharif University of Technology. His research interests include acoustic modeling, speech and natural language processing.
E-mail: khorram@ce.sharif.edu

Hossein Sameti was born in Tehran, Iran, in 1961. He received his Ph.D. degree in electrical engineering from University of Waterloo, Canada, in 1994. In 1995, he joined the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, where he is an Associate Professor now. There, he has also served as the head of the Artificial Intelligence group and the Department Chair. He founded Speech Processing Lab (SPL) at the Department of Computer Engineering in 1998 and is the supervisor of the lab. SPL has developed Nevisa, the first Persian continuous speech recognition engine, which is a commercial product now. SPL has done considerable research projects on Persian language and holds numerous patents on different aspects of Persian NLP. SPL has resulted in a few spin off's for commercial high tech speech processing products. Dr. Sameti's current research interests include speech and language processing, automatic speech recognition, speech synthesis, speech enhancement, spoken dialogue systems, spoken language understanding, speaker identification and verification, and spoken term detection.
E-mail: sameti@sharif.edu

Fahimeh Bahmaninezhad received the M.Sc. degree in Artificial Intelligence from the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, in 2012. Her main research interest is speech and natural language processing.
E-mail: fahime.bahmany@gmail.com

Paper Handling Data:

Submitted: 26.05.2014
Received in revised form: 02.10.2014
Accepted: 26.10.2014
Corresponding author: Dr. Hossein Sameti,
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.

[1] "Ezafe" is a special feature in Persian, normally pronounced as a short vowel "e" and relates two words together. Ezafe is not written but is pronounced and has a profound effect on intonation.