STATISTICAL APPROACH TO VOCAL TRACT TRANSFER FUNCTION ESTIMATION BASED ON FACTOR ANALYZED TRAJECTORY HMM

Tomoki Toda[†∗], *Keiichi Tokuda*[‡∗]

[†] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
[‡] Graduate School of Engineering, Nagoya Institute of Technology (Nitech), Japan
∗ National Institute of Information and Communications Technology (NiCT), Japan
tomoki@is.naist.jp, tokuda@nitech.ac.jp

ABSTRACT

In this paper, we describe a novel statistical approach to the vocal tract transfer function (VTTF) estimation of a speech signal based on a factor analyzed trajectory hidden Markov model (HMM). Because speech is a quasi-periodic signal, there are many missing frequency components between adjacent F_0 harmonics. The proposed method determines a time-varying VTTF sequence based on the maximum a posteriori (MAP) estimation considering not only harmonic components observed at each analyzed frame but also those at other frames for stochastically interpolating the missing frequency parts.

Index Terms— speech analysis, vocal tract transfer function, factor analysis, trajectory HMM, MAP

1. INTRODUCTION

The estimation of the vocal tract transfer function (VTTF) for a speech signal is an essential problem in speech processing. Because the speech signal results from a convolution of the VTTF and a quasi-periodic excitation signal, an observed spectrum at a voiced frame consists of line-spectra on which only harmonic components at frequencies corresponding to integral multiples of an F_0 are basically helpful for estimating the VTTF. Therefore, many missing frequency components between adjacent harmonic components make it indeed hard to extract the accurate VTTF.

Many sophisticated frame-by-frame spectral analysis methods have been studied. Itakura and Saito [1] proposed maximum likelihood estimation of a speech spectral envelope. This method determines the all-pole spectral envelope minimizing the Itakura-Saito distance that evaluates a matching error sensitively around peaks of spectral densities, i.e., around harmonic components. Tokuda et al. [2] extended this analysis method to mel-generalized cepstral analysis for treating various spectral representations including all-pole and cepstrum on a warped frequency scale in a unified framework. In these methods, it is necessary to adjust an analysis order to keep the estimated VTTF from capturing periodic components of the excitation signal. To alleviate this problem, Kawahara et al. [3] proposed STRAIGHT analysis that explicitly uses F_0 information for removing the periodic components from the estimated VTTF. These conventional spectral analysis methods basically interpolate missing frequency components considering neighboring harmonic components based on a parametric spectral envelope modeling process or an F_0 adaptive smoothing process on a time-frequency region.

Several offline spectral analysis methods for statistically extracting the averaged VTTF from multiple frames have also been studied particularly in the area of speech synthesis. This framework assumes that additional information such as phoneme transcriptions is basically available for selecting frames at which the VTTFs are presumed similar to each other. Akamine and Kagoshima [4] proposed closed loop training (CLT) for extracting a VTTF sequence for each diphone unit so that an error between natural and re-synthesized diphone waveforms is minimized. Shiga and King [5] proposed the VTTF determination based on the minimization of an error of harmonic components for multiple acoustic frames at which simultaneously recorded articulatory parameters are similar to each other. These methods basically estimate missing frequency components from harmonic components observed at other frames for determining the common VTTF for those frames.

In this paper, we describe a novel statistical approach to the offline VTTF estimation based on the maximum a posteriori (MAP) estimation. To model harmonic components observed over an utterance, we propose a factor analyzed trajectory hidden Markov model (HMM). It enables the estimation of a time-varying VTTF sequence considering not only harmonic components at each analyzed frame but also those at other frames to interpolate the missing frequency components in a probabilistic manner. We conduct a simulation experiment for demonstrating the effectiveness of the proposed method.

2. BASIC IDEA OF PROPOSED VTTF ESTIMATION

Purpose of the VTTF analysis is to extract a spectral parameter sequence $\boldsymbol{c}$ capturing time-varying VTTFs from a speech signal,

$$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top \quad (1)$$

where $\boldsymbol{c}_t$ is a D-dimensional spectral parameter vector at frame t. In this paper, we employ mel-cepstrum, which is one of effective spectral parameters having a good property to model speech signals. A sequence of observed harmonic components $\boldsymbol{s}$ is shown as

$$\boldsymbol{s} = \left[\boldsymbol{s}_1^\top, \boldsymbol{s}_2^\top, \cdots, \boldsymbol{s}_t^\top, \cdots, \boldsymbol{s}_T^\top\right]^\top \quad (2)$$

$$\boldsymbol{s}_t = \left[s_t(f_t^{(1)}), s_t(f_t^{(2)}), \cdots, s_t(f_t^{(h)}), \cdots, s_t(f_t^{(H_t)})\right]^\top \quad (3)$$

where at frame t the number of harmonics is H_t and the frequency of the h^{th} harmonic component is $f_t^{(h)}$. The total number of dimensions of $\boldsymbol{s}$ is H ($= \sum_{t=1}^{T} H_t$).

A basic idea of the proposed approach is to estimate the spectral parameter sequence that maximizes its posterior probability density given the observed harmonic components as follows:

$$\hat{\boldsymbol{c}} = \operatorname*{argmax}_{\boldsymbol{c}} P(\boldsymbol{c}|\boldsymbol{s}, \boldsymbol{\lambda}) = \operatorname*{argmax}_{\boldsymbol{c}} P(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{\lambda}) P(\boldsymbol{c}|\boldsymbol{\lambda}) \quad (4)$$

where $\boldsymbol{\lambda}$ is a parameter set of a context-dependent model trained using harmonic components at multiple other frames at which VTTFs seem similar to each other. **Fig. 1** shows an example of the estimated VTTFs. If estimating the VTTF from only the harmonic components observed at an analyzed frame in a manner such as the conventional frame-by-frame analysis methods, which is approximately related to

The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

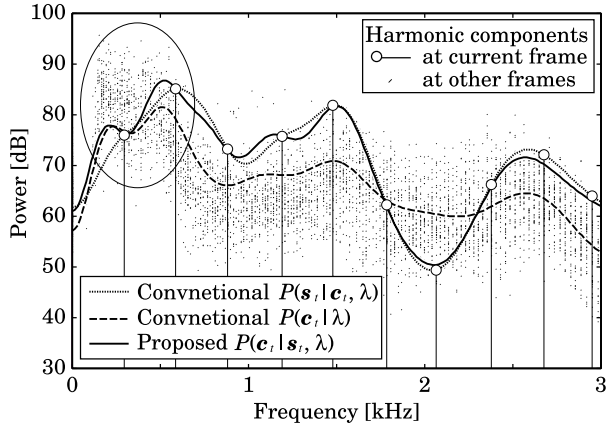

Fig. 1. An example of estimated VTTFs.

the maximization process of $P(\boldsymbol{s}_t|\boldsymbol{c}_t, \boldsymbol{\lambda})$, it is essentially difficult to estimate spectral peaks between adjacent harmonic components. It is observed from the averaged VTTF for harmonic components at multiple frames, which is determined based on $P(\boldsymbol{c}_t|\boldsymbol{\lambda})$ trained by one of the conventional offline analysis methods, that those components are very helpful for estimating such missing peaks. In the MAP estimation (i.e., based on $P(\boldsymbol{c}_t|\boldsymbol{s}_t, \boldsymbol{\lambda})$), the VTTF is determined from harmonic components observed at the analyzed frame (i.e., based on $P(\boldsymbol{s}_t|\boldsymbol{c}_t, \boldsymbol{\lambda})$) while missing frequency components are stochastically interpolated from those at multiple frames (i.e., based on $P(\boldsymbol{c}_t|\boldsymbol{\lambda})$). We can see that the MAP estimation determines the VTTF having peaks between adjacent F_0 harmonics (even a peak at lower frequency than an F_0) as shown in the area surrounded by an ellipse.

3. MODELING OF HARMONIC COMPONENT SEQUENCE BY FACTOR ANALYZED TRAJECTORY HMM

In order to realize the proposed estimation process, we need to define the probability densities $P(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{\lambda})$ and $P(\boldsymbol{c}|\boldsymbol{\lambda})$. Note that only the harmonic component sequence $\boldsymbol{s}$ is observed and the mel-cepstral sequence $\boldsymbol{c}$ should be considered as a hidden variable. This framework is described by factor analysis. We assume that the harmonic component vector $\boldsymbol{s}_t$ at frame t is modeled as follows:

$$\boldsymbol{s}_t = \boldsymbol{B}_t\boldsymbol{c}_t + \boldsymbol{n}_t \tag{5}$$

where $\boldsymbol{B}_t$ is a time-varying factor loading matrix, which is specifically an H_t-by-D DFT matrix to convert mel-cepstral coefficients into log-scaled power spectra at individual harmonic frequencies varying according to an F_0. The noise vector $\boldsymbol{n}_t$ is distributed according to a Gaussian probability density with zero mean and a diagonal covariance matrix diag$[\boldsymbol{v}_t]$ whose diagonal elements are given by

$$\boldsymbol{v}_t = \left[v_t(f_t^{(1)}), v_t(f_t^{(2)}), \cdots, v_t(f_t^{(h)}), \cdots, v_t(f_t^{(H_t)})\right]^\top. \tag{6}$$

The hidden variable, i.e., mel-cepstrum, is modeled in a state space. It is well known that dynamic characteristics of mel-cepstrum vary according to phonemic environments. The trajectory HMM [6] has a good property to model such a parameter sequence. By combining these two powerful techniques, we propose a factor analyzed trajectory HMM that effectively models the probability density of the harmonic component sequence $P(\boldsymbol{s}|\boldsymbol{\lambda})$. In this model, the spectral extraction process and the spectral modeling process are simultaneously optimized for the given observed harmonic components. This framework is similar to the structured speech modeling [7] in terms of using a state space model for modeling an observation sequence.

3.1. Factor Analyzed Trajectory HMM

Definition of $P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda})$: Let a spectral feature vector sequence be $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \cdots, \boldsymbol{o}_t^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$ where $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top]^\top$ includes not only static but also dynamic features. In the conventional HMM, the probability density of $\boldsymbol{o}$ given an HMM state sequence $\boldsymbol{q} = [q_1, q_2, \cdots, q_t, \cdots, q_T]$ is written as

$$P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\boldsymbol{q}}, \boldsymbol{U}_{\boldsymbol{q}}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{q_t}, \boldsymbol{U}_{q_t}) \tag{7}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{U})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{U}$, and

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \boldsymbol{\mu}_{q_2}^\top, \cdots, \boldsymbol{\mu}_{q_t}^\top, \cdots, \boldsymbol{\mu}_{q_T}^\top\right]^\top \tag{8}$$

$$\boldsymbol{U}_{\boldsymbol{q}} = \text{diag}\left[\left[\boldsymbol{U}_{q_1}^\top, \boldsymbol{U}_{q_2}^\top, \cdots, \boldsymbol{U}_{q_t}^\top, \cdots, \boldsymbol{U}_{q_T}^\top\right]^\top\right]. \tag{9}$$

By imposing an explicit relationship between static and dynamic features, which is given by $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$ where $\boldsymbol{W}$ is a conversion matrix to append dynamic features, the conventional HMM is reformed as the trajectory HMM [6]. For given $\boldsymbol{q}$, the probability density of $\boldsymbol{c}$ in the trajectory HMM is written as

$$P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c}; \overline{\boldsymbol{c}_{\boldsymbol{q}}}, \boldsymbol{P}_{\boldsymbol{q}}) = \frac{1}{Z_{\boldsymbol{q}}} P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}) \tag{10}$$

where

$$\overline{\boldsymbol{c}_{\boldsymbol{q}}} = \boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}} \tag{11}$$

$$\boldsymbol{P}_{\boldsymbol{q}}^{-1} = \boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{W} \tag{12}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top\boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}} \tag{13}$$

$$Z_{\boldsymbol{q}} = \frac{\sqrt{(2\pi)^{DT}|\boldsymbol{P}_{\boldsymbol{q}}|}}{\sqrt{(2\pi)^{3DT}|\boldsymbol{U}_{\boldsymbol{q}}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{q}}^\top\boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{r}_{\boldsymbol{q}}^\top\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}})\right). \tag{14}$$

Note that in the trajectory HMM the mean vector $\overline{\boldsymbol{c}_{\boldsymbol{q}}}$ varies within states and inter-frame correlation is modeled by the temporal covariance matrix $\boldsymbol{P}_{\boldsymbol{q}}$ that is generally full even if using the same number of model parameters as in the conventional HMM.

Definition of $P(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{\lambda})$: The conditional probability density of $\boldsymbol{s}$ given $\boldsymbol{c}$ and $\boldsymbol{q}$ is modeled as follows:

$$P(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{s}; \boldsymbol{B}\boldsymbol{c}, \boldsymbol{V}_{\boldsymbol{q}}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{B}_t\boldsymbol{c}_t, \text{diag}\,[\boldsymbol{v}_{q_t}]) \tag{15}$$

where

$$\boldsymbol{B} = \left[\tilde{\boldsymbol{B}}_1, \tilde{\boldsymbol{B}}_2, \cdots, \tilde{\boldsymbol{B}}_t, \cdots \tilde{\boldsymbol{B}}_T\right] \tag{16}$$

$$\tilde{\boldsymbol{B}}_t = \left[\boldsymbol{0}_{H_1\times D}^\top, \boldsymbol{0}_{H_2\times D}^\top, \cdots, \boldsymbol{B}_t^\top, \cdots, \boldsymbol{0}_{H_T\times D}^\top\right]^\top \tag{17}$$

$$\boldsymbol{V}_{\boldsymbol{q}} = \text{diag}\left[\left[\boldsymbol{v}_{q_1}^\top, \boldsymbol{v}_{q_2}^\top, \cdots, \boldsymbol{v}_{q_t}^\top, \cdots, \boldsymbol{v}_{q_T}^\top\right]^\top\right]. \tag{18}$$

Note that the dimension of $\boldsymbol{s}_t$ varies frame by frame.

Definition of $P(\boldsymbol{s}|\boldsymbol{\lambda})$: For given $\boldsymbol{q}$, the probability density of $\boldsymbol{s}$ is written as

$$P(\boldsymbol{s}|\boldsymbol{q}, \boldsymbol{\lambda}) = \int P(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{\lambda})P(\boldsymbol{c}|\boldsymbol{q}, \boldsymbol{\lambda})d\boldsymbol{c} = \mathcal{N}(\boldsymbol{s}; \overline{\boldsymbol{s}_{\boldsymbol{q}}}, \boldsymbol{O}_{\boldsymbol{q}}) \tag{19}$$

where

$$\overline{\boldsymbol{s}_{\boldsymbol{q}}} = \boldsymbol{B}\overline{\boldsymbol{c}_{\boldsymbol{q}}} \tag{20}$$

$$\boldsymbol{O}_{\boldsymbol{q}} = \boldsymbol{V}_{\boldsymbol{q}} + \boldsymbol{B}\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{B}^\top. \tag{21}$$

The intra/inter-frame correlation between any harmonic component pair over an utterance is modeled by the covariance matrix $\boldsymbol{O_q}$ that is generally full even if using diagonal covariance matrices in Eq. (9). Consequently, the likelihood function of the factor analyzed trajectory HMM for a harmonic component sequence is given by

$$P(\boldsymbol{s}|\boldsymbol{\lambda}) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{\lambda})P(\boldsymbol{s}|\boldsymbol{q},\boldsymbol{\lambda}). \quad (22)$$

To reduce the computation complexity, we approximate the likelihood function using a single HMM state sequence as follows:

$$P(\boldsymbol{s}|\boldsymbol{\lambda}) \simeq P(\boldsymbol{q}|\boldsymbol{\lambda})P(\boldsymbol{s}|\boldsymbol{q},\boldsymbol{\lambda}). \quad (23)$$

In this paper, the state sequence $\boldsymbol{q}$ is determined with Viterbi algorithm so that the likelihood $P(\boldsymbol{o},\boldsymbol{q}|\boldsymbol{\lambda})$ is maximized for a melcepstrum sequence extracted by a conventional analysis method.

3.2. Estimation of Model Parameters

Model parameters $\boldsymbol{\lambda}$ of the factor analyzed trajectory HMM are estimated so that the likelihood is maximized as follows:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\text{argmax}}\, P(\boldsymbol{s}|\boldsymbol{q},\boldsymbol{\lambda}), \quad (24)$$

where $\boldsymbol{\lambda}$ consists of

$$\boldsymbol{m} = \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \cdots, \boldsymbol{\mu}_N^\top\right]^\top \quad (25)$$

$$\boldsymbol{\Sigma}^{-1} = \left[\boldsymbol{U}_1^{-1\top}, \boldsymbol{U}_2^{-1\top}, \cdots, \boldsymbol{U}_N^{-1\top}\right]^\top \quad (26)$$

$$\boldsymbol{\sigma} = \left[\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top, \cdots, \boldsymbol{v}_{N_s}^\top\right]^\top. \quad (27)$$

Note that the transition probabilities are not updated because the state sequence $\boldsymbol{q}$ is fixed as mentioned above. The other parameters are estimated with EM algorithm.

Auxiliary function: The auxiliary function is written as

$$Q(\boldsymbol{\lambda},\hat{\boldsymbol{\lambda}}) = \int P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda}) \log P(\boldsymbol{s},\boldsymbol{c}|\boldsymbol{q},\hat{\boldsymbol{\lambda}}) d\boldsymbol{c}. \quad (28)$$

The posterior probability density function $P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda})$ in the RHS of Eq. (28) is given by

$$P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{c};\overline{\boldsymbol{c}'_{\boldsymbol{q}}},\boldsymbol{P}'_{\boldsymbol{q}}) = \frac{1}{Z'_{\boldsymbol{q}}} P(\boldsymbol{s}|\boldsymbol{c},\boldsymbol{q},\boldsymbol{\lambda}) P(\boldsymbol{W}\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda}) \quad (29)$$

where

$$\overline{\boldsymbol{c}'_{\boldsymbol{q}}} = \boldsymbol{P}'_{\boldsymbol{q}}\boldsymbol{r}'_{\boldsymbol{q}} \quad (30)$$

$$\boldsymbol{P}'^{-1}_{\boldsymbol{q}} = \boldsymbol{R}'_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}} + \boldsymbol{R}^{(s)}_{\boldsymbol{q}} \quad (31)$$

$$\boldsymbol{r}'_{\boldsymbol{q}} = \boldsymbol{r}_{\boldsymbol{q}} + \boldsymbol{r}^{(s)}_{\boldsymbol{q}} \quad (32)$$

$$\boldsymbol{R}^{(s)}_{\boldsymbol{q}} = \boldsymbol{B}^\top \boldsymbol{V}_{\boldsymbol{q}}^{-1} \boldsymbol{B} \quad (33)$$

$$\boldsymbol{r}^{(s)}_{\boldsymbol{q}} = \boldsymbol{B}^\top \boldsymbol{V}_{\boldsymbol{q}}^{-1} \boldsymbol{s} \quad (34)$$

$$Z'_{\boldsymbol{q}} = \frac{\sqrt{(2\pi)^{DT}|\boldsymbol{P}'_{\boldsymbol{q}}|}}{\sqrt{(2\pi)^H|\boldsymbol{V}_{\boldsymbol{q}}|(2\pi)^{3DT}|\boldsymbol{U}_{\boldsymbol{q}}|}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{q}}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}} + \boldsymbol{s}^\top \boldsymbol{V}_{\boldsymbol{q}}^{-1}\boldsymbol{s} - \boldsymbol{r}'^\top_{\boldsymbol{q}} \boldsymbol{P}'_{\boldsymbol{q}} \boldsymbol{r}'_{\boldsymbol{q}})\right). \quad (35)$$

The log-scaled joint probability density function $\log P(\boldsymbol{s},\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda})$ in the RHS of Eq. (28) is given by

$$\log P(\boldsymbol{s},\boldsymbol{c}|\boldsymbol{q},\boldsymbol{\lambda}) = -\frac{1}{2}\Big\{\log|\boldsymbol{P}_{\boldsymbol{q}}| + \boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{r}_{\boldsymbol{q}} - 2\boldsymbol{r}_{\boldsymbol{q}}^\top \boldsymbol{c} + tr(\boldsymbol{R}_{\boldsymbol{q}}\boldsymbol{c}\boldsymbol{c}^\top) + \log|\boldsymbol{V}_{\boldsymbol{q}}| + \boldsymbol{s}^\top \boldsymbol{V}_{\boldsymbol{q}}^{-1}\boldsymbol{s} - 2\boldsymbol{r}_{\boldsymbol{q}}^{(s)\top}\boldsymbol{c} + tr(\boldsymbol{R}^{(s)}_{\boldsymbol{q}}\boldsymbol{c}\boldsymbol{c}^\top)\Big\}. \quad (36)$$

In these Eqs., parameter sequences over an utterance are given by

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{A}_{\boldsymbol{q}}\boldsymbol{m} \quad (37)$$

$$\boldsymbol{U}_{\boldsymbol{q}}^{-1} = \text{diag}\left[\boldsymbol{A}_{\boldsymbol{q}}\boldsymbol{\Sigma}^{-1}\right] \quad (38)$$

$$\boldsymbol{V}_{\boldsymbol{q}} = \text{diag}\left[\boldsymbol{A}^{(\sigma)}_{\boldsymbol{q}}\boldsymbol{\sigma}\right] \quad (39)$$

where $\boldsymbol{A}_{\boldsymbol{q}}$ and $\boldsymbol{A}^{(\sigma)}_{\boldsymbol{q}}$ are a state assignment matrix and a state and frequency-bin assignment matrix, respectively.

E-step: The following statistics are calculated,

$$\int P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda})\boldsymbol{c}d\boldsymbol{c} = \overline{\boldsymbol{c}'_{\boldsymbol{q}}} \quad (40)$$

$$\int P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda})\boldsymbol{c}\boldsymbol{c}^\top d\boldsymbol{c} = \boldsymbol{P}'_{\boldsymbol{q}} + \overline{\boldsymbol{c}'_{\boldsymbol{q}}}\,\overline{\boldsymbol{c}'_{\boldsymbol{q}}}^\top. \quad (41)$$

M-step: Updated noise variance vectors $\hat{\boldsymbol{\sigma}}$ are written as

$$\hat{\boldsymbol{\sigma}} = (\boldsymbol{A}^{(\sigma)\top}_{\boldsymbol{q}}\boldsymbol{A}^{(\sigma)}_{\boldsymbol{q}})^{-1}\boldsymbol{A}^{(\sigma)\top}_{\boldsymbol{q}} \cdot\text{on-diag}\left[\boldsymbol{B}\boldsymbol{P}'_{\boldsymbol{q}}\boldsymbol{B}^\top + (\boldsymbol{s} - \boldsymbol{B}\overline{\boldsymbol{c}'_{\boldsymbol{q}}})(\boldsymbol{s} - \boldsymbol{B}\overline{\boldsymbol{c}'_{\boldsymbol{q}}})^\top\right], \quad (42)$$

where on-diag[·] uses only diagonal elements of a square matrix. Updated mean vectors $\hat{\boldsymbol{m}}$ are written as

$$\hat{\boldsymbol{m}} = \left(\boldsymbol{A}_{\boldsymbol{q}}^\top \boldsymbol{W}\boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{W}^\top \boldsymbol{A}_{\boldsymbol{q}}\text{diag}\left[\boldsymbol{\Sigma}^{-1}\right]\right)^{-1} \boldsymbol{A}_{\boldsymbol{q}}^\top \boldsymbol{W}\overline{\boldsymbol{c}'_{\boldsymbol{q}}}. \quad (43)$$

Covariance matrices $\boldsymbol{\Sigma}$ are updated using the following gradient,

$$\frac{\partial Q}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2}\boldsymbol{A}_{\boldsymbol{q}}^\top \text{on-diag}\left[\boldsymbol{W}(\boldsymbol{P}_{\boldsymbol{q}} + \hat{\overline{\boldsymbol{c}_{\boldsymbol{q}}}}\,\hat{\overline{\boldsymbol{c}_{\boldsymbol{q}}}}^\top)\boldsymbol{W}^\top - 2\hat{\boldsymbol{\mu}}_{\boldsymbol{q}}\hat{\overline{\boldsymbol{c}_{\boldsymbol{q}}}}^\top \boldsymbol{W}^\top - \boldsymbol{W}(\boldsymbol{P}'_{\boldsymbol{q}} + \overline{\boldsymbol{c}'_{\boldsymbol{q}}}\,\overline{\boldsymbol{c}'_{\boldsymbol{q}}}^\top)\boldsymbol{W}^\top + 2\hat{\boldsymbol{\mu}}_{\boldsymbol{q}}\overline{\boldsymbol{c}'_{\boldsymbol{q}}}^\top \boldsymbol{W}^\top\right] \quad (44)$$

where

$$\hat{\overline{\boldsymbol{c}_{\boldsymbol{q}}}} = \boldsymbol{P}_{\boldsymbol{q}}\boldsymbol{W}^\top \boldsymbol{U}_{\boldsymbol{q}}^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{q}}. \quad (45)$$

It is straightforward to extend this training algorithm to the multiple observation sequences.

3.3. MAP Estimation of Spectral Parameter Sequence

Based on the trained factor analyzed trajectory HMMs, a spectral parameter sequence is determined by maximizing the posterior probability density function $P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda})$ given by Eq. (29) as follows:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\text{argmax}}\, P(\boldsymbol{c}|\boldsymbol{s},\boldsymbol{q},\boldsymbol{\lambda}) = \overline{\boldsymbol{c}'_{\boldsymbol{q}}}. \quad (46)$$

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We conducted a simulation experiment based on the VTTF estimation for re-synthesized speech samples. First, we analyzed natural speech samples using STRAIGHT analysis method. And then, we re-synthesized speech waveforms from the extracted spectra and varied F_0s using STRAIGHT synthesis. A manipulation ratio of F_0 was set to 2^{-1}, $2^{-0.5}$, 1, $2^{0.5}$, 2, and $2^{1.5}$ (e.g., re-synthesized F_0s were twice as large as the original F_0s when the manipulation ratio was set to 2). We employed the STRAIGHT mixed excitation and a full representation of the extracted STRAIGHT spectra with no compression to re-synthesize speech as accurately as possible. We used 50 phonetically balanced sentences uttered by a Japanese male speaker (MHT). Sampling frequency was 16 kHz.

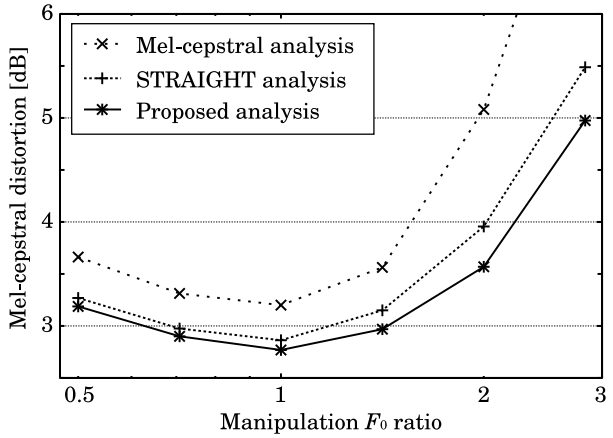

Fig. 2. Mel-cepstral distortion as a function of manipulation F_0 ratio.

We evaluated the VTTF estimation accuracy of mel-cepstral analysis [2], STRAIGHT analysis [3], and the proposed analysis by comparing the VTTFs estimated from the re-synthesized speech samples with *true VTTFs*, i.e., the spectra extracted from natural speech samples, which were used in re-synthesis. Frame shift was set to 5 ms. Mel-cepstral distortion with the first through 24^{th} mel-cepstral coefficients was employed as an objective measure. In mel-cepstral analysis, the analysis order was set to 24, i.e., the 0^{th} through 24^{th} mel-cepstral coefficients were determined. The proposed method also used those coefficients as a spectral parameter to model the VTTF.

In the proposed method, we employed the continuous HMM (5 state left-to-right with no skips) of which each state output probability density was modeled by a single Gaussian distribution with a diagonal covariance matrix. In order to determine the initial parameters of the factor analyzed trajectory HMMs, we first trained the initial monophone HMMs using mel-cepstra obtained by STRAIGHT analysis. And then, we constructed tied-state triphone HMMs using a decision-tree based context clustering technique adopting the minimum description length (MDL) criterion [8]. The total number of resulting HMM states was 149 (i.e., $N = 149$ in Eqs. (25) and (26)). Using the resulting HMMs, the HMM state sequence was determined for each utterance with Viterbi algorithm. As for the noise variance, we employed a single variance vector tied over all HMM states (i.e., $N_s = 1$ in Eq. (27)). The initial noise variance values were determined based on errors between harmonic components represented by the STRAIGHT mel-cepstra and the observed harmonic components. After these initialization processes, we iteratively updated parameters of the factor analyzed trajectory HMMs. In order to alleviate F_0 interference in observing harmonic components, the pitch synchronous analysis method [3] in STRAIGHT was employed. Based on the trained factor analyzed trajectory HMMs, a mel-cepstral sequence to model the time-varying VTTFs for each utterance was determined with the MAP estimation.

4.2. Experimental results

Fig. 2 shows mel-cepstral distortion between the estimated VTTF and the *true VTTF* as a function of the manipulation F_0 ratio. An example of the estimated VTTF sequences is shown in **Fig. 3**. The estimation accuracy rapidly degrades as F_0s relatively increase because the number of observed harmonic components decreases. The estimation accuracy of mel-cepstral analysis is worse than the others because the estimated VTTF captures F_0 harmonic components (see around 0.3 [s] in **Fig. 3**). STRAIGHT realizes more robust VTTF estimation by using F_0 information to remove its influence. However, it is essentially difficult to estimate peaks of the VTTF between

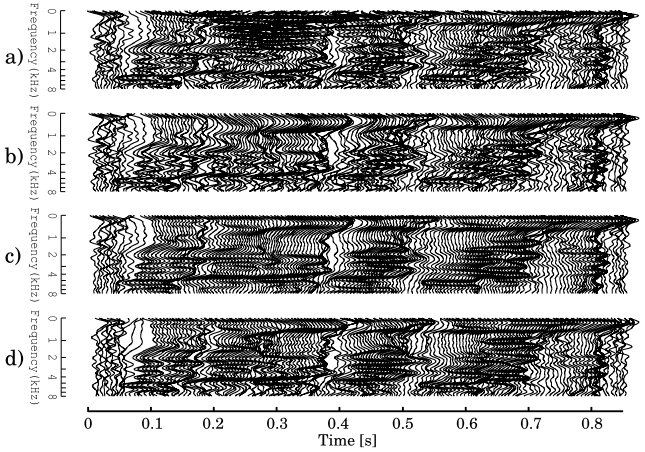

Fig. 3. An example of VTTF sequences extracted using a) mel-cepstral analysis, b) STRAIGHT analysis, and c) proposed analysis. The *true VTTF* sequence is shown as d). Manipulation F_0 ratio is set to two. Every VTTF is liftered by the 24^{th} mel-cepstral coefficients.

F_0 harmonics. The best VTTF estimation accuracy is attained by the proposed method. As shown in **Fig. 3**, the proposed method makes it possible to estimate the VTTF sequence exhibiting similar spectral structures and their dynamic characteristics to the true ones.

5. CONCLUSIONS

We proposed a statistical method for estimating the vocal tract transfer function (VTTF) from a speech signal. The proposed method realized the maximum a posteriori estimation of the VTTF sequence based on a factor analyzed trajectory hidden Markov model to effectively model harmonic components observed over an utterance. An experimental result showed that the proposed method is very effective particularly when an F_0 of the analyzed speech is high. It is worthwhile to evaluate the effectiveness of the proposed framework as a training algorithm in the HMM-based speech synthesis system.

6. REFERENCES

[1] F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *IEICE Trans.*, Vol. J53-A, No. 1, pp. 35–42, 1970.

[2] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis – a unified approach to speech spectral estimation. *Proc. ICSLP*, pp. 1043–1045, Yokohama, Japan, Sep. 1994.

[3] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[4] M. Akamine and T. Kagoshima. Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS Drive TTS). *Proc. ICSLP*, pp. 1927–1930, Sydney, Australia, Dec. 1998.

[5] Y. Shiga and S. King. Estimating the spectral envelope of voiced speech using multi-frame analysis. *Proc. EUROSPEECH*, pp. 1737–1740, Geneva, Switzerland, Sep. 2003.

[6] H. Zen, K. Tokuda, and T. Kitamura. Reformulating the HMM as a trajetory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, Vol. 21, pp. 153-173, 2007.

[7] L. Deng, D. Yu, and A. Acero. Structured speech modeling. *IEEE Trans. Audio, Speech and Language Processing*, Vol. 14, No. 5, pp. 1492–1504, 2006.

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