STATISTICAL F_0 PREDICTION FOR ELECTROLARYNGEAL SPEECH ENHANCEMENT CONSIDERING GENERATIVE PROCESS OF F_0 CONTOURS WITHIN PRODUCT OF EXPERTS FRAMEWORK

Kou Tanaka[1], Hirokazu Kameoka[2], Tomoki Toda[3], and Satoshi Nakamura[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]Information Technology Center, Nagoya University, Japan
{ko-t, s-nakamura}@is.naist.jp, kameoka.hirokazu@lab.ntt.co.jp, tomoki@icts.nagoya-u.ac.jp

ABSTRACT

We have previously proposed a statistical fundamental frequency (F_0) prediction method that makes it possible to predict the underlying F_0 contour of electrolaryngeal (EL) speech from its spectral feature sequence. Although this method was shown to contribute to improving the naturalness of EL speech as a whole, the predicted F_0 contour was still unnatural compared with that in normal speech. One possible solution to improve the naturalness of the predicted F_0 contours would be to take account of the physical mechanism of vocal phonation. Recently a statistical model of voice F_0 contours was formulated by constructing a stochastic counterpart of the Fujisaki model, a well-founded mathematical model representing the control mechanism of vocal fold vibration. This paper proposes a Product-of-Experts model to incorporate this generative model of voice F_0 contours into the statistical F_0 prediction model. Based on the constructed model, we derive algorithms for parameter training and F_0 prediction. Experimental results revealed that the proposed method successfully outperformed our previously proposed method in terms of the naturalness of the predicted F_0 contours.

Index Terms— Electrolaryngeal speech enhancement, F_0 prediction, Generative model, Product of Experts

1. INTRODUCTION

Speech is one of the most common tools in human communication. Since speech is produced by the vocal apparatus, the produced sounds are physically constrained by the conditions of human body. Unfortunately, there are many people with disabilities that prevent them from producing speech freely, leading to communication barriers. Those who are unable to produce speech freely involve laryngectomees, who have undergone an operation to remove the larynx including the vocal folds for such reasons as injury or laryngeal cancer. The ability by these people to generate excitation sounds is severely impaired because they no longer have their vocal folds. One alternative means of producing voice for these patients involves the use of electrolaryngeal (EL) speech, which is produced by using the excitation signals mechanically generated from an electrolarynx. EL speech is reasonably intelligible, but somewhat unnatural particularly due to the mechanical sounding of the excitation signals.

To address this problem, we have previously proposed methods that aim to convert EL speech to normal-sounding speech, by predicting the fundamental frequency (F_0) contour from the spectrum sequence of the EL speech based on Gaussian Mixture Models (GMMs) followed by synthesizing the speech waveforms according to the predicted acoustic parameters [1–3]. These methods were shown to contribute to improving the naturalness of EL speech [1, 2] and also preserving its intelligibility [3]. However, the F_0 contours predicted using these methods still sounded unnatural compared with that in normal speech. This was because the predicted F_0 contours were not necessarily guaranteed to satisfy the physical constraint of the actual control mechanism of the thyroid cartilage, even though they were optimal in a statistical sense. In this regard, these methods still had a plenty of room for improvement. One possible solution to improve the naturalness of the F_0 contours of the converted speech would be to incorporate a generative model of voice F_0 contours into the statistical F_0 prediction model to take account of the physical mechanism of vocal phonation.

One of the authors previously proposed a statistical model of voice F_0 contours [4–6], formulated by constructing a stochastic counterpart of the Fujisaki model [7], a well-founded mathematical model representing the control mechanism of vocal fold vibration. The Fujisaki model [7] assumes that an F_0 contour on a logarithmic scale is the superposition of a phrase component, an accent component and a base value. The phrase and accent components are considered to be associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times. The model proposed in [4–6] has made it possible to estimate the underlying parameters of the Fujisaki model that best explain the given F_0 contour, by using powerful statistical inference techniques.

To incorporate the generative F_0 contour model into the statistical F_0 prediction framework, this paper proposes a Product-of-Experts (PoE) model [8] combining the abovementioned two models. Since the PoE model is obtained by multiplying the densities of different models, it usually becomes complicated due to the renormalization term. To avoid this, we introduce a latent trajectory model proposed in [9] to reformulate the prediction model so that it can be smoothly combined with the generative F_0 contour model.

2. GMM-BASED STATISTICAL F_0 PREDICTION

We briefly review our statistical F_0 prediction method [1–3], which exploits the idea of statistical voice conversion techniques [10, 11]. The aim of this method is to predict F_0 contours from the spectral parameters of EL speech. As with voice conversion methods, it consists of training and prediction processes.

In the training process, the parameters $\boldsymbol{\lambda}_G$ of the joint probability density $p((\boldsymbol{x}[k]^\mathsf{T}, \boldsymbol{o}[k]^\mathsf{T})^\mathsf{T}|\boldsymbol{\lambda}_G)$ described as a Gaussian mixture model (GMM) are trained, where $^\mathsf{T}$ de-

This work was supported in part of JSPS KAKENHI 26280060.

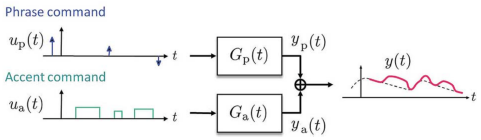

Fig. 1. Original Fujisaki model [7].

notes transposition, and $\boldsymbol{x}[k]$ and $\boldsymbol{o}[k]$ denote a source feature and a target feature at time frame k, respectively. The corresponding joint feature vectors can be obtained by performing automatic frame alignment with Dynamic Time Warping. As a source feature, the spectral segment feature of EL speech is extracted on a frame-by-frame basis from the mel-cepstra at multiple frames around the current frame k [12]. The target feature $\boldsymbol{o}[k] = (y[k], \Delta y[k])^\mathsf{T}$ consists of the static and delta (time derivative) components of the log-scaled F_0 value $y[k]$, extracted on a frame-by-frame basis from the target normal speech. Note that to improve prediction accuracy, we interpolate unvoiced frames of F_0 patterns by using spline interpolation and remove micro-prosody [13].

In the prediction process, given the spectral segment sequence $\boldsymbol{x} = (\boldsymbol{x}[1]^\mathsf{T}, \ldots, \boldsymbol{x}[K]^\mathsf{T})^\mathsf{T}$ of EL speech, the most likely F_0 sequence $\boldsymbol{y} = (y[1], \ldots, y[K])^\mathsf{T}$ can be obtained as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\lambda}_G) \;\; \text{subject to } \boldsymbol{o} = \boldsymbol{W}\boldsymbol{y}, \tag{1}$$

where $\boldsymbol{o} = (\boldsymbol{o}[1]^\mathsf{T}, \ldots, \boldsymbol{o}[K]^\mathsf{T})^\mathsf{T}$ denotes the joint static and dynamic feature vector sequence, $\boldsymbol{W}$ is a constant matrix that transforms the static feature vector sequence $\boldsymbol{y}$ to $\boldsymbol{o}$. Namely, each row of $\boldsymbol{W}$ consists of the coefficients of an identity mapping operator or time differential operator. $p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\lambda}_G)$ is the GMM with the trained parameters, which we approximate as

$$\begin{aligned} p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\lambda}_G) &= \sum_{\boldsymbol{m}} p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\lambda}_G) p(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{\lambda}_G) \\ &\simeq p(\boldsymbol{o}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}_G) p(\hat{\boldsymbol{m}}|\boldsymbol{x}, \boldsymbol{\lambda}_G), \end{aligned} \tag{2}$$

with $\hat{\boldsymbol{m}} = \operatorname{argmax}_m p(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{\lambda}_G)$. Here, $\boldsymbol{m} = (m_1, \ldots, m_K)$ indicates a sequence of mixture indices. $p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\lambda}_G)$ is given as the product of $p(\boldsymbol{o}[k]|\boldsymbol{x}[k], m_k, \boldsymbol{\lambda}_G) = \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{e}_{m_k}^{(o|x)}, \boldsymbol{D}_{m_k}^{(o|x)})$ over k where $\boldsymbol{e}_{m_k}^{(o|x)}$ and $\boldsymbol{D}_{m_k}^{(o|x)}$ are the mean vector and covariance matrix of m_k-th mixture component, respectively. Thanks to this approximation, the solution to Eq. (1) is given explicitly as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\mathsf{T} \boldsymbol{D}^{(o|x)^{-1}} \boldsymbol{W})^{-1} \boldsymbol{W}^\mathsf{T} \boldsymbol{D}^{(o|x)^{-1}}_{\hat{\boldsymbol{m}}} \boldsymbol{e}^{(o|x)}_{\hat{\boldsymbol{m}}}, \tag{3}$$

where $\boldsymbol{e}^{(o|x)}_{\hat{\boldsymbol{m}}}$ is a stacked vector of the mean vectors $\boldsymbol{e}^{(o|x)}_{\hat{m}_1}, \ldots, \boldsymbol{e}^{(o|x)}_{\hat{m}_K}$ and $\boldsymbol{D}^{(o|x)}_{\hat{\boldsymbol{m}}}$ is a block diagonal matrix where each block is $\boldsymbol{D}^{(o|x)}_{\hat{m}_1}, \ldots, \boldsymbol{D}^{(o|x)}_{\hat{m}_K}$.

3. GENERATIVE MODEL OF VOICE F_0 CONTOURS

The generative model of F_0 contours proposed in [4–6] is a stochastic counterpart of a discrete-time version of the Fujisaki model [7].

The Fujisaki model (shown in Fig. 1) assumes that a log-scaled F_0 contour $y(t)$ is the superposition of a phrase component $y_\mathrm{p}(t)$, an accent component $y_\mathrm{a}(t)$ and a base value μ_b. The phrase and accent components are assumed to be the outputs of different second-order critically damped filters, excited with Dirac deltas $u_\mathrm{p}(t)$ (phrase commands) and rectangular pulses $u_\mathrm{a}(t)$ (accent commands), respectively. Here,

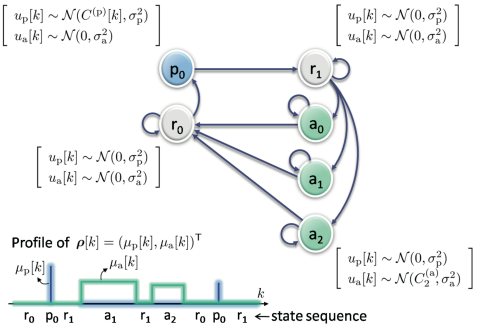

Fig. 2. Command function modeling with HMM.

it must be noted that the phrase and accent commands do not usually overlap each other. The base value is a constant value related to the lower bound of the speaker's F_0, below which no regular vocal fold vibration can be maintained. The log F_0 contour, $y(t)$, is thus expressed as

$$y(t) = y_\mathrm{p}(t) + y_\mathrm{a}(t) + \mu_b, \tag{4}$$

where

$$y_\mathrm{p}(t) = g_\mathrm{p}(t) * u_\mathrm{p}(t), \tag{5}$$

$$y_\mathrm{a}(t) = g_\mathrm{a}(t) * u_\mathrm{a}(t). \tag{6}$$

Here, $*$ denotes convolution over time. $g_\mathrm{p}(t)$ and $g_\mathrm{a}(t)$ are the impulse responses of the two second-order systems, which are known to be almost constant within an utterance as well as across utterances for a particular speaker.

A key idea in the proposed model [4–6] is that the sequence of the phrase and accent command pair (i.e., the underlying parameters of the Fujisaki model) is modeled as a path-restricted hidden Markov model (HMM) with Gaussian emission densities (shown in Fig. 2) so that estimating the state transition of the HMM directly amounts to estimating the Fujisaki-model parameters.

We hereafter use k to indicate the discrete time index. Given a state sequence $\boldsymbol{s} = (s_1, \ldots, s_K)$ of the above HMM, the conditional distributions of the phrase command sequence $\boldsymbol{u}_\mathrm{p} = (u_\mathrm{p}[1], \ldots, u_\mathrm{p}[K])^\mathsf{T}$ and the accent command sequence $\boldsymbol{u}_\mathrm{a} = (u_\mathrm{a}[1], \ldots, u_\mathrm{a}[K])^\mathsf{T}$ are given as

$$p(\boldsymbol{u}_\mathrm{p}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \mathcal{N}(\boldsymbol{u}_\mathrm{p}; \boldsymbol{\mu}_\mathrm{p}, \boldsymbol{\Sigma}_\mathrm{p}), \tag{7}$$

$$p(\boldsymbol{u}_\mathrm{a}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \mathcal{N}(\boldsymbol{u}_\mathrm{a}; \boldsymbol{\mu}_\mathrm{a}, \boldsymbol{\Sigma}_\mathrm{a}), \tag{8}$$

respectively, where $\boldsymbol{\lambda}_F$ denotes the parameters of the HMM. $\boldsymbol{\mu}_\mathrm{p}$ and $\boldsymbol{\mu}_\mathrm{a}$ denote the mean sequences of the state emission densities and $\boldsymbol{\Sigma}_\mathrm{p}$ and $\boldsymbol{\Sigma}_\mathrm{a}$ are diagonal matrices whose diagonal elements correspond to the variances of the state emission densities. From Eqs. (5) and (6), the relationships between $\boldsymbol{y}_\mathrm{p} = (y_\mathrm{p}[1], \ldots, y_\mathrm{p}[K])^\mathsf{T}$ and $\boldsymbol{u}_\mathrm{p}$ and between $\boldsymbol{y}_\mathrm{a} = (y_\mathrm{a}[1], \ldots, y_\mathrm{a}[K])^\mathsf{T}$ and $\boldsymbol{u}_\mathrm{a}$ can be written as

$$\boldsymbol{G}_\mathrm{p}\boldsymbol{u}_\mathrm{p} = \boldsymbol{y}_\mathrm{p}, \tag{9}$$

$$\boldsymbol{G}_\mathrm{a}\boldsymbol{u}_\mathrm{a} = \boldsymbol{y}_\mathrm{a}, \tag{10}$$

where $\boldsymbol{G}_\mathrm{p}$ and $\boldsymbol{G}_\mathrm{a}$ are Toeplitz matrices where each row is a shifted copy of the convolution kernels $g_\mathrm{p}[1], \ldots, g_\mathrm{p}[K]$ and $g_\mathrm{a}[1], \ldots, g_\mathrm{a}[K]$. By using $\boldsymbol{u}_\mathrm{b}$ to denote the baseline component, the log F_0 sequence $\boldsymbol{y}$ is given as $\boldsymbol{y} = \boldsymbol{y}_\mathrm{p} + \boldsymbol{y}_\mathrm{a} + \boldsymbol{u}_\mathrm{b} + \boldsymbol{n}$ where $\boldsymbol{n}$ is an additive noise component corresponding to micro prosody. If we assume that $\boldsymbol{n}$ follows a Gaussian distribution with mean $\boldsymbol{0}$ and covariance $\boldsymbol{\Gamma}$, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u} = (\boldsymbol{u}_\mathrm{p}^\mathsf{T}, \boldsymbol{u}_\mathrm{a}^\mathsf{T}, \boldsymbol{u}_\mathrm{b}^\mathsf{T})^\mathsf{T}$ is defined as

$$p(\boldsymbol{y}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{G}_{\mathrm{p}}\boldsymbol{u}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{u}_{\mathrm{a}} + \boldsymbol{u}_{\mathrm{b}}, \boldsymbol{\Gamma}). \tag{11}$$

We further assume that $\boldsymbol{u}_{\mathrm{b}}$ follows a Gaussian distribution with mean $\mu_{\mathrm{b}}\mathbf{1}$ and covariance $\boldsymbol{\Sigma}_{\mathrm{b}}$. Then, from Eqs. (7), (8) and (11), the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{s}$ is given as

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\Lambda}_F) &= \textstyle\int p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) \mathrm{d}\boldsymbol{u} \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F), \end{aligned} \tag{12}$$

where $\boldsymbol{\mu}_F = \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\mu}_{\mathrm{a}} + \mu_{\mathrm{b}}\mathbf{1}$ and $\boldsymbol{\Sigma}_F = \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\Sigma}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{p}}^{\mathsf{T}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\Sigma}_{\mathrm{a}}\boldsymbol{G}_{\mathrm{a}}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathrm{b}} + \boldsymbol{\Gamma}$.

4. PROPOSED MODEL

4.1. Product-of-Experts Strategy

PoE [8] is a general technique to model a complicated distribution of data by combining relatively simpler distributions (experts). Since the distribution is obtained by multiplying the densities of the experts, the way the experts are combined is somewhat similar to an "and" operation. In this section, we construct a PoE model by treating the two models introduced in the previous sections as the experts.

Training a PoE model by maximizing the likelihood of the data usually becomes difficult since it is hard even to approximate the derivatives of the renormalization term. By contrast, we propose an elegant formulation that allows the use of the EM algorithm for both parameter training and F_0 prediction. To do so, we first introduce a latent trajectory model proposed in [9] to reformulate the GMM-based statistical F_0 prediction model, which plays a key role in making this possible.

4.2. Latent-Trajectory-GMM-Based F_0 Prediction

We reformulate the GMM-based statistical F_0 prediction model presented in Sec. 2 by employing the idea proposed in [9]. Instead of treating $\boldsymbol{o}$ as a function of $\boldsymbol{y}$, we treat $\boldsymbol{o}$ as a latent variable to be marginalized out, that is related to $\boldsymbol{y}$ through a soft constraint $\boldsymbol{o} \simeq \boldsymbol{W}\boldsymbol{y}$. The relationship $\boldsymbol{o} \simeq \boldsymbol{W}\boldsymbol{y}$ can be expressed through the conditional distribution $p(\boldsymbol{y}|\boldsymbol{o})$

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{o}) &\propto \exp\left\{-\tfrac{1}{2}(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{o})^{\mathsf{T}}\boldsymbol{\Lambda}(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{o})\right\} && (13) \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{H}\boldsymbol{o}, \boldsymbol{V}), && (14) \end{aligned}$$

where $\boldsymbol{H} = (\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\Lambda}$ and $\boldsymbol{V} = (\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{W})^{-1}$. $\boldsymbol{\Lambda}$ is a constant positive definite matrix that can be set arbitrarily. As with Sec. 2, the joint distribution $p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{\lambda}_G)$ is modeled as a GMM. Namely, given mixture indices $\boldsymbol{m}$, the conditional distribution $p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\Lambda}_G)$ is defined as a Gaussian distribution. Thus, the joint distribution of $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{o}$, and $\boldsymbol{m}$ can be described using the distributions defined above

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{m}|\boldsymbol{\lambda}_G) = p(\boldsymbol{y}|\boldsymbol{o}) p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\lambda}_G) p(\boldsymbol{m}|\boldsymbol{\lambda}_G), \tag{15}$$

where $p(\boldsymbol{m}|\boldsymbol{\lambda}_G)$ is the product of mixture weights of the GMM. By marginalizing $\boldsymbol{o}$ and $\boldsymbol{m}$ out, we can readily obtain the joint distribution $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\lambda}_G)$, which can be used as a criterion to train $\boldsymbol{\lambda}_G$ and predict optimal $\boldsymbol{y}$ in a consistent manner, unlike the method presented in Sec. 2. We use this model to construct our PoE model in the next subsection.

4.3. Deriving PoE

In the same way as Eq. (15), we write the model presented in Sec. 3 in the form of a joint distribution

$$p(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{s}|\boldsymbol{\lambda}_F) = p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) p(\boldsymbol{s}|\boldsymbol{\lambda}_F), \tag{16}$$

where $p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ is given as the product of state emission densities and $p(\boldsymbol{s}|\boldsymbol{\lambda}_F)$ the product of the state transition probabilities given a state sequence $\boldsymbol{s}$. We consider constructing a PoE model by combining Eqs. (15) and (16) followed by marginalization, rather than by simply combining the marginal distributions $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\lambda}_G)$ and $p(\boldsymbol{y}|\boldsymbol{\lambda}_F)$, which makes the parameter training and F_0 prediction problems excessively hard. To do so, we first combine the densities of $p(\boldsymbol{y}|\boldsymbol{o})$ and $p(\boldsymbol{y}|\boldsymbol{u})$ to obtain $p(\boldsymbol{y}|\boldsymbol{o}, \boldsymbol{u})$. Since both of these distributions are Gaussians, the product of their distributions can be easily obtained by completing the square of the exponent

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{o}, \boldsymbol{u}) &\propto \mathcal{N}(\boldsymbol{y}; \boldsymbol{H}\boldsymbol{o}, \boldsymbol{V}) \cdot \mathcal{N}(\boldsymbol{y}; \boldsymbol{G}\boldsymbol{u}, \boldsymbol{\Gamma}) \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{y|o,u}, \boldsymbol{\Sigma}_{y|o,u}), && (17) \\ \boldsymbol{\mu}_{y|o,u} &= (\boldsymbol{V}^{-1} + \boldsymbol{\Gamma}^{-1})^{-1}(\boldsymbol{V}^{-1}\boldsymbol{H}\boldsymbol{o} + \boldsymbol{\Gamma}^{-1}\boldsymbol{G}\boldsymbol{u}), && (18) \\ \boldsymbol{\Sigma}_{y|o,u} &= (\boldsymbol{V}^{-1} + \boldsymbol{\Gamma}^{-1})^{-1}, && (19) \end{aligned}$$

where $\boldsymbol{G} = [\boldsymbol{G}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{a}}\boldsymbol{I}]$ and $\boldsymbol{u} = (\boldsymbol{u}_{\mathrm{p}}^{\mathsf{T}}, \boldsymbol{u}_{\mathrm{a}}^{\mathsf{T}}, \boldsymbol{u}_{\mathrm{b}}^{\mathsf{T}})^{\mathsf{T}}$. From Eqs. (15), (16) and (17), the joint distribution of $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{o}$, $\boldsymbol{u}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ can be constructed as

$$\begin{aligned} &p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{u}, \boldsymbol{m}, \boldsymbol{s}|\boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F) && (20) \\ &= \underbrace{p(\boldsymbol{y}|\boldsymbol{o}, \boldsymbol{u}) p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\lambda}_G) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)}_{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{u}|\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F)} p(\boldsymbol{m}|\boldsymbol{\lambda}_G) p(\boldsymbol{s}|\boldsymbol{\lambda}_F). \end{aligned}$$

This can be used as the complete data likelihood for parameter training and F_0 prediction as explained later. By marginalizing $\boldsymbol{o}$ and $\boldsymbol{u}$ out, we can readily obtain the joint distribution $p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{m}, \boldsymbol{s}|\boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F)$, which can be used as a criterion to train $\boldsymbol{\lambda}_G$ and $\boldsymbol{\lambda}_F$ and predict optimal $\boldsymbol{y}$ in a consistent manner.

Since both $p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\lambda}_G)$ and $p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ are Gaussians, let us write them as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\lambda}_G) &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{o}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_o\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{xx} & \boldsymbol{P}_{xo}\\ \boldsymbol{P}_{ox} & \boldsymbol{P}_{oo}\end{bmatrix}^{-1}\right), && (21) \\ p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) &= \mathcal{N}(\boldsymbol{u}; \boldsymbol{\mu}_u, \boldsymbol{P}_u^{-1}). && (22) \end{aligned}$$

Then, from Eqs. (17), (21) and (22), it can be shown that $p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{u}|\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F)$ is given as

$$\begin{aligned} &p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{u}|\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F) \\ &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\\ \boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix}; \begin{bmatrix}\boldsymbol{A}_{11}\boldsymbol{b}_1 + \boldsymbol{A}_{12}\boldsymbol{b}_2\\ \boldsymbol{A}_{21}\boldsymbol{b}_1 + \boldsymbol{A}_{22}\boldsymbol{b}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix}\right), && (23) \end{aligned}$$

where

$$\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{y|o,u}^{-1} & O & -\boldsymbol{V}^{-1}\boldsymbol{H} & -\boldsymbol{\Gamma}^{-1}\boldsymbol{G} \\ O & \boldsymbol{P}_{xx} & -\boldsymbol{P}_{xo} & O \\ -\boldsymbol{H}^{\mathsf{T}}\boldsymbol{V}^{-\mathsf{T}} & -\boldsymbol{P}_{ox} & \boldsymbol{P}_{oo} & O \\ -\boldsymbol{G}^{\mathsf{T}}\boldsymbol{\Gamma}^{-\mathsf{T}} & O & O & \boldsymbol{P}_u \end{bmatrix}^{-1}, \tag{24}$$

$$\begin{bmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\end{bmatrix} = \begin{bmatrix} O \\ \boldsymbol{P}_{xx}\boldsymbol{\mu}_x - \boldsymbol{P}_{xo}\boldsymbol{\mu}_o \\ \boldsymbol{P}_{oo}\boldsymbol{\mu}_o - \boldsymbol{P}_{ox}\boldsymbol{\mu}_x \\ \boldsymbol{P}_u\boldsymbol{\mu}_u \end{bmatrix}, \tag{25}$$

by completing the square of the exponent. Note that $\boldsymbol{A}_{11}$, $\boldsymbol{A}_{12}$, $\boldsymbol{A}_{21}$ and $\boldsymbol{A}_{22}$ can be written explicitly using the blockwise inversion formula.

4.4. Parameter Training and F_0 Prediction

The problems of parameter training and F_0 prediction can be formulated as the following optimization problems:

$$\{\hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F, \hat{\boldsymbol{m}}, \hat{\boldsymbol{s}}\} = \underset{\boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F, \boldsymbol{m}, \boldsymbol{s}}{\operatorname{argmax}} \log p(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}, \boldsymbol{m}, \boldsymbol{s}|\boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F), \tag{26}$$

$$\{\hat{\boldsymbol{y}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{s}}\} = \underset{\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{s}}{\operatorname{argmax}} \log p(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{m}, \boldsymbol{s}|\hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F). \tag{27}$$

where $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{x}}$ denote the observed F_0 contour extracted from normal speech and the observed spectral sequence extracted from non-larynx speech. Both of these problems can be solved using the EM algorithm by treating $\boldsymbol{o}$ and $\boldsymbol{u}$ as latent variables. Owing to space limitations, here we only derive an algorithm for solving Eq. (27).

The likelihood of $\boldsymbol{y}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ given the complete data $\{\tilde{\boldsymbol{x}}, \boldsymbol{o}, \boldsymbol{u}\}$ is given by Eq. (20). By taking the conditional expectation of $\log p(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{o}, \boldsymbol{u}|\boldsymbol{m}, \boldsymbol{s}, \hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F)$ with respect to $\boldsymbol{o}$ and $\boldsymbol{u}$ given $\tilde{\boldsymbol{x}}$, $\boldsymbol{y} = \boldsymbol{y}'$, $\boldsymbol{m} = \boldsymbol{m}'$ and $\boldsymbol{s} = \boldsymbol{s}'$ and then adding $\log p(\boldsymbol{m}|\hat{\boldsymbol{\lambda}}_G)p(\boldsymbol{s}|\hat{\boldsymbol{\lambda}}_F)$, we obtain an auxiliary function

$$Q(\theta, \theta') = \mathbb{E}_{\boldsymbol{o},\boldsymbol{u}|\tilde{\boldsymbol{x}},\boldsymbol{y}',\boldsymbol{m}',\boldsymbol{s}'}\left[\log p(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{o}, \boldsymbol{u}|\boldsymbol{m}, \boldsymbol{s}, \hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F)\right] + \log p(\boldsymbol{m}|\hat{\boldsymbol{\lambda}}_G) + \log p(\boldsymbol{s}|\hat{\boldsymbol{\lambda}}_F), \quad (28)$$

where $\theta = \{\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{s}\}$. From Eq. (23), we obtain

$$\mathbb{E}\left[\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{y}'\\ \tilde{\boldsymbol{x}}\end{bmatrix}, \boldsymbol{m}', \boldsymbol{q}'\right] = \boldsymbol{A}_{21}\boldsymbol{b}_1 + \boldsymbol{A}_{22}\boldsymbol{b}_2 + \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\left(\begin{bmatrix}\boldsymbol{y}'\\ \tilde{\boldsymbol{x}}\end{bmatrix} - \boldsymbol{A}_{11}\boldsymbol{b}_1 - \boldsymbol{A}_{12}\boldsymbol{b}_2\right) =: \begin{bmatrix}\bar{\boldsymbol{o}}\\ \bar{\boldsymbol{u}}\end{bmatrix}, \quad (29)$$

$$\mathbb{E}\left[\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix}\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix}^{\top} \middle| \begin{bmatrix}\boldsymbol{y}'\\ \tilde{\boldsymbol{x}}\end{bmatrix}, \boldsymbol{m}', \boldsymbol{q}'\right] = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} + \begin{bmatrix}\bar{\boldsymbol{o}}\\ \bar{\boldsymbol{u}}\end{bmatrix}\begin{bmatrix}\bar{\boldsymbol{o}}\\ \bar{\boldsymbol{u}}\end{bmatrix}^{\top}, \quad (30)$$

which are the values to be computed at the "E-step" by substituting θ into θ'. At the "M-step", we compute

$$\{\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{s}\} \leftarrow \underset{\boldsymbol{y},\boldsymbol{m},\boldsymbol{s}}{\operatorname{argmax}}\, Q(\theta, \theta'). \quad (31)$$

The update equations are omitted owing to space limitations.

5. EXPERIMENTAL EVALUATION

5.1. Experimental Conditions

We conducted objective and subjective evaluation experiments to evaluate the performance of the proposed method. For the objective evaluation, we evaluated the F_0 correlation coefficients between the predicted and target F_0 contours. We also subjectively evaluated the naturalness of the F_0 contour of converted speech.

The source speech was EL speech uttered by one male laryngectomee, and the target speech was normal speech uttered by a professional female speaker. Each speaker uttered about 50 sentences in the ATR phonetically balanced sentence set [15]. We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance pairs were used for evaluation. The sampling frequency was set at 16 kHz. The settings of HMM in the generative F_0 contour model were the same as reported in [4–6].

For initialization of the mixture index sequence $\boldsymbol{m}$ and state sequence $\boldsymbol{s}$, we performed the conventional GMM-based and Fujisaki-model-based methods. Note that the Fujisaki-model-based method refers to a post-processing method that consists of first applying the GMM-based method and then fitting the Fujisaki model to the predicted F_0 contour using the method of [5, 6], which is similar to a post-processing method for HMM-based speech synthesis [16]. Note that in this experiment, we implemented a simplified and approximated version of the proposed method, in which the E-step procedure is replaced with the conventional Fujisaki-model-based and GMM-based methods. Therefore, the convergence of the algorithm implemented for the current evaluation is not strictly guaranteed. The speech used for evaluation were synthesized using STRAIGHT [17] given the mel-cepstrum sequence and F_0 contour. The methods selected for comparison were:

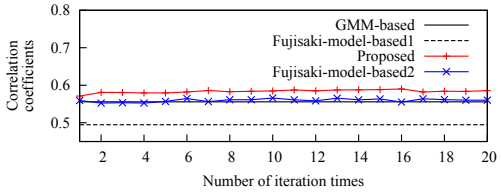

Fig. 3. F_0 correlation coefficients between predicted F_0 patterns from EL speech and extracted F_0 patterns from normal speech.

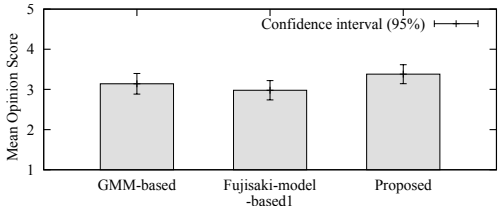

Fig. 4. Result of opinion test on naturalness.

- *GMM-based*: Predict F_0 contours with the GMM-based method.
- *Fujisaki-model-based1*: Fit the Fujisaki model to the predicted F_0 contours obtained with the *GMM-based* method.
- *Proposed*: Predict F_0 contours with an approximated version of the proposed method, in which the E-step is replaced with the *GMM-based* and *Fujisaki-model-based2* methods.
- *Fujisaki-model-based2*: Fit the Fujisaki model to the predicted F_0 contours obtained with *Proposed*.

5.2. Experimental Results

As Fig. 3 shows, *Proposed* obtained the highest prediction accuracy because of eq. (18) meaning an "and" operation for eq. (11) and (14). Therefore, we found that it is effectiveness to construct a PoE model. Futhermore, since *Fujisaki-model-based2* has higher correration coefficients than *Fujisaki-model-based1*, we found that the predicted F_0 contours by *Proposed* were given good influences by considering not only the GMM-based method but also the Fujisaki model. Note that we used the predicted F_0 contours obtained with the *GMM-based* method as the input for the *Fujisaki-model-based1* method in this experiment. To make a more fair comparison, it would be necessary to modify the Fujisaki-model-based method so as not to depend on the GMM-based method. In additional, as for proposed, there is no large difference of correration coefficients in each iteration. To make exactly evaluation, we have to conduct the PoE model not replaced E-step with the conventional methods.

As Fig. 4 shows, *Proposed* outperformed the conventional methods, *GMM-based* and *Fujisaki-model-based1*. This result is reasonable since *Proposed* obtained the highest prediction accuracy as in Fig. 3.

6. CONCLUSIONS

In this paper, to improve F_0 prediction performance in electrolaryngeal speech enhancement, we proposed a Product-of-Experts model that combined two conventional methods, a statistical F_0 prediction method and a statistical F_0 contour modeling method based on its generative process. Experimental results revealed that the proposed method successfully outperformed our previously proposed method in terms of the naturalness of the predicted F_0 contours.

7. REFERENCES

[1] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," in *Proc. Speech Communication,* vol. 54, pp. 134– 146, January 2012.

[2] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *Audio, Speech, and Language Processing, IEEE/ACM Transactionson,* vol. 22, no. 1, pp. 172–183, January 2014.

[3] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems,* vol. E97-D, no. 6, pp. 1429–1437, June 2014.

[4] H. Kameoka, J. Le Roux, Y. Ohishi, "A statistical model of speech F_0 contours," *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition,* pp. 43–48, September 2010.

[5] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Statistical approach to fujisaki-model parameter estimation from speech signals and its quantitative evaluation," *Proc. Speech Prosody,* pp. 175–178, May 2012.

[6] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *Audio, Speech, and Language Processing, IEEE/ACM Transactionson,* vol. 23, no. 6, pp. 1042–1053, June 2015.

[7] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal Fold Physiology: Voice Production, Mechanisms and Functions,* pp. 347–355, 1998.

[8] G. E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation,* vol. 14, no. 8, pp. 1771–1800, August 2002.

[9] H. Kameoka, "Modeling speech parameter sequences with latent trajectory hidden markov model," in *Proc. The 25th IEEE International Workshop on Machine Learning for Signal Processing,* September 2015.

[10] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on,* vol. 6, no. 2, pp. 131–142, March 1998.

[11] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, November 2007.

[12] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 9, pp. 2505–2517, November 2012.

[13] K. J. Kohler, "Macro and micro F_0 in the synthesis of intonation," *Papers in Laboratory Phonology I,* pp. 115–138, 1990.

[14] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 509–5012, 2002.

[15] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database," *ATR Technical Report,* TR-I-0166, September 1990.

[16] T. Matsuda, K. Hirose, and N. Minematsu, "Applying generation process model constraint to fundamental frequency contours generated by hidden-Markov-model-based speech synthesis," *Acoustical Science and Technology,* Vol. 33, No. 4, pp. 221–228, 2012.

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "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," in *Proc. Speech Communication,,* Vol. 27, No. 3-4, pp. 187–207, April 1999.