INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

DNN-SPACE: DNN-HMM-based Generative Model of Voice F_0 Contours for Statistical Phrase/Accent Command Estimation

Nobukatsu Hojo[1], *Yasuhito Ohsugi*[2], *Yusuke Ijima*[1], *Hirokazu Kameoka*[3]

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]The University of Tokyo, Japan
[3]NTT Communication Science Laboratories, NTT Corporation, Japan

hojo.nobukatsu@lab.ntt.co.jp

Abstract

This paper proposes a method to extract prosodic features from a speech signal by leveraging auxiliary linguistic information. A prosodic feature extractor called the statistical phrase/accent command estimation (SPACE) has recently been proposed. This extractor is based on a statistical model formulated as a stochastic counterpart of the Fujisaki model, a well-founded mathematical model representing the control mechanism of vocal fold vibration. The key idea of this approach is that a phrase/accent command pair sequence is modeled as an output sequence of a path-restricted hidden Markov model (HMM) so that estimating the state transition amounts to estimating the phrase/accent commands. Since the phrase and accent commands are related to linguistic information, we may expect to improve the command estimation accuracy by using them as auxiliary information for the inference. To model the relationship between the phrase/accent commands and linguistic information, we construct a deep neural network (DNN) that maps the linguistic feature vectors to the state posterior probabilities of the HMM. Thus, given a pitch contour and linguistic information, we can estimate phrase/accent commands via state decoding. We call this method "DNN-SPACE." Experimental results revealed that using linguistic information was effective in improving the command estimation accuracy.

Index Terms: the Fujisaki model, SPACE, linguistic information, DNN

1. Introduction

The fundamental frequency (F_0) contours in speech contain various types of non-linguistic information such as the speaker's identity, emotions and intentions. Modeling the F_0 contours of speech utterances can thus be potentially useful for many speech applications, including speech recognition, speaker recognition, speech synthesis and dialog systems.

The Fujisaki model [1, 2] is a well-founded mathematical model that describes an F_0 contour as the sum of two contributions. This model approximates actual F_0 contours of speech fairly well when the model parameters are appropriately chosen, and its validity has been demonstrated for many typologically diverse languages [1, 3, 4, 5, 6, 7, 8, 9]. Since prosodic features in speech are predominantly characterized by the levels and timings of the phrase and accent components, one important challenge is to solve the inverse problem of estimating the Fujisaki model parameters automatically from a raw F_0 contour.

For estimating the Fujisaki model parameters from a raw F_0 contour, several methods have been developed [2, 10, 11, 12, 13, 14, 15]. Among these methods, the statistical phrase/accent command estimation (SPACE) [16, 17, 15] method is noteworthy because it is based on a statistical model formulated as a stochastic counterpart of the Fujisaki model. The key idea of this approach is that a phrase/accent command pair sequence is modeled as an output sequence of a path-restricted hidden Markov model (HMM) so that estimating the state transition amounts to estimating the phrase/accent commands. There are two benefits for this statistical reformulation: one is to derive an efficient parameter inference algorithm utilizing powerful statistical methods, and the other is to obtain an automatically trainable version of the Fujisaki model. Owing to these benefits, some SPACE-based applications have already been proposed such as F_0 generation for text-to-speech synthesis [18] or F_0 prediction for electrolaryngeal speech enhancement [19]. Therefore, if we could improve the parameter estimation accuracy of the SPACE method, we expect that it will also be able to improve the basic performance of such SPACE-based applications.

An important approach to improve Fujisaki model parameter estimation is leveraging auxiliary linguistic information [20, 21, 22, 23]. Since the phrase/accent commands are closely associated with the linguistic information such as breath group or accent nucleus, it is expected that the command estimation accuracy can be improved by using them as auxiliary information for inferences. In a previous study, Hirose et al. [21] proposed to use linguistic information to obtain a first approximation of the location of the commands which is then adjusted by an iterative analysis-by-synthesis process. Torres et al. [22] also proposed to use genetic algorithms to estimate Fujisaki model parameters considering linguistic aspects. It is expected that the parameter estimation accuracy of the SPACE method will also improve by utilizing the linguistic information.

Motivated in a way similar to these previous studies, we aimed to improve the parameter estimation accuracy of the SPACE method by extending the statistical model so that it would incorporate a linguistic information model. More specifically, we constructed a deep neural network (DNN) that maps the linguistic feature vectors to the state posterior probabilities of the HMM. We formulated a generative model of the F_0 contours by combining the DNN with the SPACE model in such a way that a parameter estimation algorithm could be derived on the basis of a DNN-HMM [24] framework and an auxiliary function method. We call this proposed method "DNN-SPACE" in this paper. The key of the formulation is to incorporate a linguistic information model to the conventional SPACE method without changing its basic parameter optimization algorithms.

The rest of this paper is organized as follows. Section 2 briefly reviews the original Fujisaki model and a discrete-time stochastic counterpart to the Fujisaki model (SPACE). Section 3 formulates the proposed method (DNN-SPACE) to extend SPACE to integrate the linguistic information models. Section 4 presents experimental evaluations obtained for the proposed method. Section 5 concludes the paper with a summary of key points and a mention of future work.

http://dx.doi.org/10.21437/Interspeech.2017-719

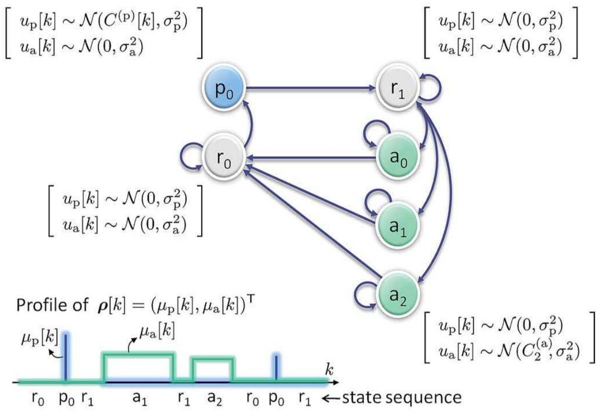
$u_p[k] \sim \mathcal{N}(C^{(p)}[k], \sigma_p^2)$
$u_a[k] \sim \mathcal{N}(0, \sigma_a^2)$
$u_p[k] \sim \mathcal{N}(0, \sigma_p^2)$
$u_a[k] \sim \mathcal{N}(0, \sigma_a^2)$
p0
r1
r0
a0
$u_p[k] \sim \mathcal{N}(0, \sigma_p^2)$
$u_a[k] \sim \mathcal{N}(0, \sigma_a^2)$
a1
a2
Profile of $\rho[k] = (\mu_p[k], \mu_a[k])^T$
$\mu_p[k]$
$\mu_a[k]$
$u_p[k] \sim \mathcal{N}(0, \sigma_p^2)$
$u_a[k] \sim \mathcal{N}(C_2^{(a)}, \sigma_a^2)$
k
r0 p0 r1 a1 r1 a2 r0 p0 r1 ←state sequence

Figure 1: *Command function modeling with an HMM.*

2. Generative model of speech F_0 contours

2.1. Original Fujisaki model

The Fujisaki model [1, 2] assumes an F_0 contour on a logarithmic scale, $y(t)$, where t is time, is the superposition of three components: a phrase component $y_p(t)$, an accent component $y_a(t)$, and a base component y_b:

$$y(t) = y_p(t) + y_a(t) + y_b. \quad (1)$$

The phrase component $y_p(t)$ consists of the major-scale pitch variations over the duration of the prosodic units, and the accent component $y_a(t)$ consists of the smallerscale pitch variations in accented syllables. These two components are modeled as the outputs of second-order critically damped filters, one being excited with a command function $u_p(t)$ consisting of Dirac deltas (phrase commands), and the other with $u_a(t)$ consisting of rectangular pulses (accent commands):

$$x_p(t) = G_p(t) * u_p(t), \quad (2)$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \quad (3)$$

$$x_a(t) = G_a(t) * u_a(t), \quad (4)$$

$$G_a(t) = \begin{cases} \beta^2 t e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (5)$$

where $*$ denotes convolution over time. The baseline component y_b 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. α and β are natural angular frequencies 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. It has been shown that $\alpha = 3$ rad/s and $\beta = 20$ rad/s can be used as default values.

It is interesting to note that the phrase and accent commands, which we will henceforth refer to as the Fujisaki model parameters, can be interpreted as quantities related to linguistic information. In the Japanese language, a phrase command and an accent command typically occur at the beginning of each breath group and over the range of accent nucleus in each accentual phrase, respectively.

2.2. Probabilistic formulation of F_0 contour model (SPACE)

Here, we briefly review the conventional probabilistic pitch contour model based on the discrete-time version of the Fujisaki model [16, 17, 15].

In the original Fujisaki model, phrase commands and accent commands are assumed to respectively consist of Dirac deltas and rectangular pulses. In addition, they are not allowed to overlap each other. To incorporate these requirements, we find it convenient to model the $u_p[k]$ and $u_a[k]$ pair, i.e., $o[k] = (u_p[k], u_a[k])^\top$, using a path-restricted HMM. In previous studies [16, 17, 15], the authors assumed that $\{\boldsymbol{o}[k]\}_{k=1}^K$ is a sequence of outputs generated from an HMM with the specific topology illustrated in Fig. 1.

In state r_0, $\mu_p[k]$ and $\mu_a[k]$ are both constrained to be zero. In state p_1, referred to as the "phrase state", $\mu_p[k]$ can take a non-zero value, $C_p[k]$, whereas $\mu_a[k]$ is still restricted to zero. At the phrase state, no self transitions are allowed. In state r_1, $\mu_p[k]$ and $\mu_a[k]$ become zero again. This path constraint restricts $\mu_p[k]$ to consisting of isolated deltas. State r_1 leads to states $a_1, \cdots, a_N$, referred to as "accent states". At each accent state, $\mu_a[k]$ can take a different non-zero value $C_n^{(a)}$, whereas $\mu_p[k]$ is forced to be zero. A direct state transition from an accent state to a different state without passing through state r_1 is not allowed. This path constraint restricts $\mu_a[k]$ to consisting of rectangular pulses. The output distribution of each state is assumed to be the Gaussian distribution

$$\boldsymbol{o}[k] \sim \mathcal{N}(\boldsymbol{c}_{s_k}, \Gamma_{s_k}), \quad (6)$$

where s_k indicates the state variable. Namely, the mean vector $\boldsymbol{\mu}[k] = (\mu_p[k], \mu_a[k])^\top = \boldsymbol{c}_{s_k}$ and the covariance matrix $\boldsymbol{\Sigma}[k] = \Gamma_{s_k}$ is considered to evolve in time as a result of the state transition $s_1, \cdots, s_K$. The definition of the above HMM can be summarized as follows:

output sequence : $\boldsymbol{o}[k] = (u_p[k], u_a[k])^{\mathrm{T}}$

state set $\boldsymbol{S} = \{r_0, p_0, r_1, a_0, \cdots, a_{N-1}\}$

state sequence : $\boldsymbol{s} = \{s_k\}_k$

output distribution : $p(\boldsymbol{o}[k] | s_k = i) = \mathcal{N}(\boldsymbol{c}_i[k], \boldsymbol{\Sigma}_i)$

$$\boldsymbol{c}_i[k] = \begin{cases} (0, 0)^{\mathrm{T}} & (i \in r_0, r_1) \\ (C^{(p)}[k], 0)^{\mathrm{T}} & (i \in p_0) \\ (0, C_n^{(a)})^{\mathrm{T}} & (i \in a_n) \end{cases},$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} v_{p,i}^2 & 0 \\ 0 & v_{a,i}^2 \end{bmatrix}.$$

state transition probability : $\phi_{\hat{\imath},i} = p(s_k = i | s_{k-1} = \hat{\imath})$

Given the state sequence $\boldsymbol{s} = \{s_k\}_{k=1}^K$, the above HMM generates the $u_p[k]$ and $u_a[k]$ pair. Taking into consideration (2) and (4), we then fed $u_p[k]$ and $u_a[k]$ through the different critically damped filters $G_p[k]$ and $G_a[k]$ to generate the phrase and accent components $y_p[k]$ and $y_a[k]$:

$$x_p[k] = G_p[k] * u_p[k], \quad (7)$$

$$x_a[k] = G_a[k] * u_a[k] \quad (8)$$

where $*$ denotes convolution over k. An F_0 contour is then given by

$$y[k] = x_p[k] + x_a[k] + x_b, \quad (9)$$

where x_b denotes the baseline value.

3. Proposed model (DNN-SPACE)

As discussed in sec. 2.1, phrase and accent commands are closely associated with linguistic information such as breath

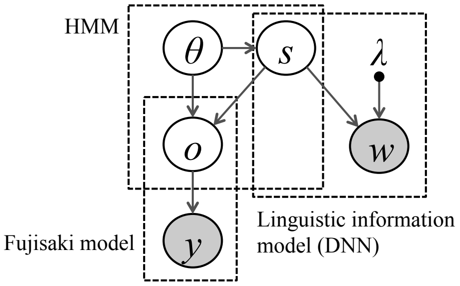

Figure 2: *Graphical representation of proposed model.*

groups or accent nucleus. Thus, we may expect to improve the command estimation accuracy by leveraging auxiliary linguistic information. To incorporate the linguistic information model to the conventional SPACE method, we constructed a DNN that maps the linguistic feature vectors to the state posterior probabilities of the HMM. We formulated the generative model of the F_0 contours by combining the DNN with the SPACE model in such a way that a parameter estimation algorithm could be derived on the basis of the DNN-HMM [24] framework and an auxiliary function method. The key to our formulation is incorporating a linguistic information model to the conventional SPACE method without changing its basic parameter optimization algorithms.

Given the linguistic information sequence $\boldsymbol{w} = \{\boldsymbol{w}[k]\}$, we constructed a DNN that maps $\boldsymbol{w}[k]$ to the posterior probability of the HMM state s_k. For simplicity, the DNN was trained to discriminate the three sets of HMM states shown below given the linguistic feature $\boldsymbol{w}[k]$.

- Phrase state ($\{p_0\}$)
- Accent states ($\{a_0, \cdots, a_N\}$)
- The other states ($\{r_0, r_1\}$)

All the accent states share the same posterior probability given the linguistic information. In the work we report in this paper, we assumed that the DNN can be trained with speech corpora that include linguistic annotations and manually annotated phrase/accent command functions. Accordingly, we formulated a generative model of the F_0 contours by combining the DNN with the SPACE method on the basis of the DNN-HMM [24] framework. More specifically, we assumed the generative probability $p(\boldsymbol{w}|\boldsymbol{s}, \lambda)$ was time-independent and transformed by applying Bayes' theorem:

$$p(\boldsymbol{w}|\boldsymbol{s}, \lambda) = \prod_k p(\boldsymbol{w}[k]|s_k, \lambda) \tag{10}$$

$$= \prod_k \frac{p(s_k|\boldsymbol{w}[k], \lambda)p(\boldsymbol{w}[k])}{p(s_k)}. \tag{11}$$

Here, λ denotes the parameters of the DNN. We used the relative frequency for $p(s_k)$ and assumed $p(\boldsymbol{w}[k])$ as a constant. The formulation for the generative process of the other parameters ($\{\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{s}, \theta\}$) is the same as that described in sec. 2.2. Figure 2 shows the graphical representation of the proposed model.

There are two benefits for this formulation. First, by utilizing the linguistic feature $\boldsymbol{w}[k]$ as an input vector of the DNN, we can assume there was no specific distribution for p($\boldsymbol{w}[k]|s_k, \lambda$). Therefore, we can use the proposed DNN-SPACE method framework to leverage auxiliary information other than linguistic information, such as spectral features or image features. Second, the basic parameter estimation algorithms are not changed from those for the conventional SPACE method. The parameter estimation algorithm for the proposed method is derived to maximize the posterior probability $p(\boldsymbol{o}, \boldsymbol{s}, \theta|\boldsymbol{y}, \boldsymbol{w}, \lambda)$ with respect to $\boldsymbol{o}, \boldsymbol{s}$ and θ. In a similar way that Sato et al. did for their proposed parameter optimization algorithm [25], for our method we derived an iterative algorithm to search the maximum posterior probability:

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o}, \hat{\boldsymbol{s}}, \hat{\theta}|\boldsymbol{y}, \boldsymbol{w}, \lambda), \tag{12}$$

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p(\hat{\boldsymbol{o}}, \boldsymbol{s}, \hat{\theta}|\boldsymbol{y}, \boldsymbol{w}, \lambda), \tag{13}$$

$$\hat{\theta} = \arg\max_{\theta} p(\hat{\boldsymbol{o}}, \hat{\boldsymbol{s}}, \theta|\boldsymbol{y}, \boldsymbol{w}, \lambda). \tag{14}$$

It should be noted from the decomposition

$$p(\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{w}, \boldsymbol{s}, \theta|\lambda) = p(\boldsymbol{y}|\boldsymbol{o})p(\boldsymbol{o}|\boldsymbol{s}, \theta)p(\boldsymbol{w}|\boldsymbol{s}, \lambda) p(\boldsymbol{s}|\theta)p(\theta) \tag{15}$$

that the objective function of (12) and (14) is independent from $\boldsymbol{w}$ and λ. This means the updates of (12) and (14) can be derived in the same way as Sato et al. described [25] by using the auxiliary function method. As for the update of (13), the objective function is in the same form as that of DNN-HMM [24]. Therefore, a Viterbi algorithm-based update algorithm can be derived.

4. Experiments

4.1. Experimental conditions

To evaluate the parameter estimation accuracy of the proposed method, we conducted an experiment using the ATR Japanese speech database B-set [26]. This database consists of 503 phonetically balanced sentences. We selected speech samples of one male speaker (MHT). The ground truth data of the Fujisaki model parameters had been manually annotated by an expert in the speech prosody field. In these ground truth data, the baseline values were all set at log 60 Hz. We compared the performance of the proposed method with that of two conventional Fujisaki model parameter extractors that do not utilize linguistic information; Narusawa's method [11] and the SPACE method [17]. We used 450 sentences for training the DNN and then tested the parameter estimation accuracy on the remaining 53 sentences.

We used the DNN with two hidden layers with 256 units, a fully-connected network and a sigmoid activation function. As linguistic feature $\boldsymbol{w}$, we used 495 dimensional vectors that are used as the linguistic feature vectors for DNN-based speech synthesis. The vectors include information such as phonemes and accent types [27]. The weights of the DNN were initialized randomly, then optimized to minimize the cross entropy between the target and predicted value, using the Adam [28]-based back-propagation algorithm. The parameters for the Adam algorithm were set as $\alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 1e-8$. We used 22 sentences as the development set.

We obtained V/UV segments by simple energy thresholding. The constant parameters were fixed at $t_0 = 8$ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $v_p^2[k] = 0.2^2$, $v_a^2[k] = 0.1^2$, $v_b^2 = 0.001^2$ and $v_n^2[k] = 10^{15}$ for unvoiced regions and $v_n^2[k] = 0.2^2$ for voiced regions. The parameter x_b was set at the minimum log F_0 value in the voiced regions. The initial values for the proposed method were set at those obtained with a conventional method [17] that does not utilize linguistic information. The parameter estimation algorithm was run for 20 iterations.

We evaluated the accuracy of the parameter estimation on the basis of two criteria: log F_0 RMSE (root mean sequared error) and detection rates. Our aim was to confirm whether the proposed model can achieve higher parameter estimation accu-

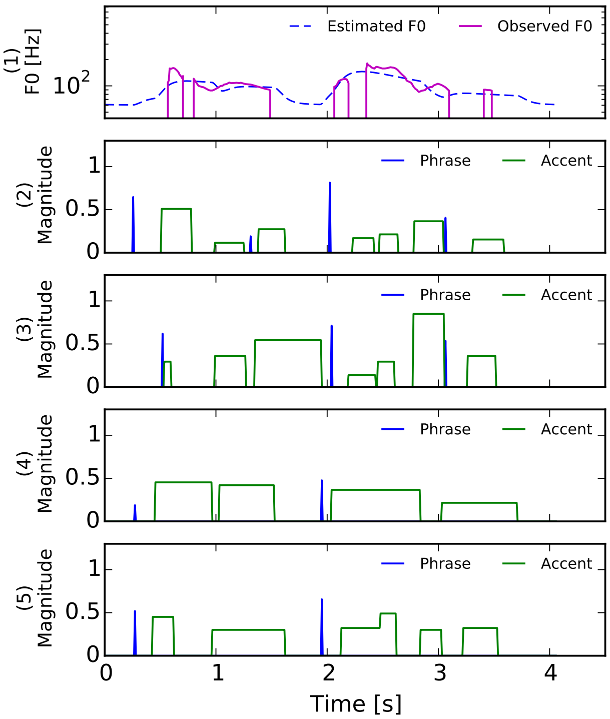

Figure 3: *Example of command detection. (1) An observed F_0 contour in voiced regions (solid line) and the estimated F_0 contours by the proposed method (dotted line). The estimated phrase and accent commands. (2) Narusawa's method [11]. (3) SPACE [17]. (4) DNN-SPACE (the proposed method). (5) Grand truth (manually labeled by an expert).*

racy. The detection rate was used to evaluate the accuracy of the parameter estimates, which was calculated in the following way: we performed matching between the estimated and ground truth command sequences on a command-by-command basis by using dynamic programming algorithm. If the time difference between the estimated and ground truth phrase commands was shorter than S seconds, the estimated phrase command was considered "matched" and the local distance was set at zero. Otherwise the local distance was set at 1. As for the accent commands, we took the average of the time difference between the onsets of the estimated and ground truth accent commands and the time difference between the offsets of the estimated and ground truth accent commands. In the same way, when the average time difference was shorter than S seconds, the estimated accent command was considered matched. The magnitudes of the phrase and accent commands were not taken into account in our evaluation. This is because the magnitude estimation was very sensitive to the baseline F_0 value, which was set differently in the proposed method and in manual annotation. Let N_{E} and N_{A} be the total number of commands in the estimated and ground truth command sequences, N_{M} be the number of the matched commands between the two sequences, and N_{Esum}, N_{Asum} and N_{Msum} be the sum of $N_{\mathrm{E}}, N_{\mathrm{A}}, N_{\mathrm{M}}$ for all 53 sentences. We defined the insertion error rate E_{I} as $(N_{\mathrm{Esum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$, the deletion error rate E_{D} as $(N_{\mathrm{Asum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$, and the detection rate D as $1 - E_{\mathrm{I}} - E_{\mathrm{D}}$.

Table 1: *Detection rates and log F_0 RMSE (S=0.3s)*

	Detection rates	log F_0 RMSE
Narusawa [11]	0.666	0.1160
SPACE [17]	0.700	0.0540
DNN-SPACE (proposed)	**0.717**	0.0636

4.2. Experimental results

Table 1 shows the results obtained with our quantitative evaluation with $S = 0.3$ s. The "Narusawa" and "SPACE" rows shows the detection rate and log F_0 RMSE of the command sequences obtained with the conventional method which does not utilize linguistic information. The "DNN-SPACE" row shows the command sequence obtained with the proposed method using linguistic information. We can see that the detection rate of the proposed method is improved compared to the conventional methods. We can conclude from the result that leveraging auxiliary linguistic information for parameter estimation improves the command estimation accuracy. While the log F_0 RMSE of the proposed method was improved compared to Narusawa's method, it was slightly worse than the SPACE method. These results show that improvement in command detection accuracy by utilizing the linguistic information does not always lead to more accurate F_0 contour estimation. This is because the Fujisaki model has the nature that even though a set of command functions that ignores the relationship with the linguistic information, it can approximate a given observed F_0 contour.

Figure 3 shows the parameter estimation results along with the ground truth data. We can also confirm that the proposed model is able to estimate parameters more accurately than the previous models. For example, while the conventional Narusawa and SPACE methods had insertion errors of phrase commands around $t = 3.0$, our proposed method eliminated them. This is because our method can utilize linguistic information in which there is no phrase boundary around $t = 3.0$.

5. Conclusion

In this paper, we described a method we propose to extract prosodic features from a speech signal by leveraging auxiliary linguistic information. The DNN-SPACE method we propose was extended from the conventional SPACE method to model the relationship between phrase/accent commands and linguistic information. We evaluated the method's parameter estimation accuracy and revealed that using linguistic information was effective in improving the command estimation accuracy. Our future work will include incorporating sequence model such as RNN instead of DNN to model the time dependence of command functions and further improve the parameter estimation accuracy.

6. References

[1] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal physiology: Voice production, mechanisms and functions*, pp. 347–355, 1988.

[2] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[3] H. Fujisaki and S. Ohno, "Analysis and modeling of fundamental frequency contours of English utterances," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[4] H. Mixdorff and H. Fujisaki, "Analysis of voice fundamental frequency contours of german utterances using a quantitative model," in *Third International Conference on Spoken Language Processing*, 1994.

[5] H. Fujisaki, S. Ohno, K.-i. Nakamura, M. Guirao, and J. Gurlekian, "Analysis of accent and intonation in spanish based on a quantitative model," in *Third International Conference on Spoken Language Processing*, 1994.

[6] H. Fujisaki, M. Ljungqvist, and H. Murata, "Analysis and modeling of word accent and sentence intonation in swedish," in *Proc. ICASSP 1993*, vol. 2. IEEE, 1993, pp. 211–214.

[7] H. Fujisaki, S. Narusawa, S. Ohno, and D. Freitas, "Analysis and modeling of f_0 contours of portuguese utterances based on the command-response model." in *Proc. INTERSPEECH 2003*, 2003.

[8] C. Wang, H. Fujisaki, R. Tomana, and S. Ohno, "Analysis of fundamental frequency contours of standard chinese in terms of the command-response model and its application to synthesis by rule of intonation." in *Proc. INTERSPEECH 2000*, 2000, pp. 326–329.

[9] H. Fujisaki, W. Gu, and K. Hirose, "The command-response model for the generation of f_0 contours of cantonese utterances," in *Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on*, vol. 1. IEEE, 2004, pp. 655–658.

[10] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proc. ICASSP 2000*, vol. 3. IEEE, 2000, pp. 1281–1284.

[11] 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. ICASSP 2002*, vol. 1. IEEE, 2002, pp. I–509.

[12] P. S. Rossi, F. Palmieri, and F. Cutugno, "A method for automatic extraction of Fujisaki-model parameters," in *Speech Prosody 2002, International Conference*, 2002.

[13] P. S. Rossi, F. Palmieri, and F. Cutugno, "Inversion of F0 model for natural-sounding speech synthesis." in *Proc. ICASSP 2003*, 2003, pp. 520–523.

[14] H. R. Pfitzinger, H. Mixdorff, and J. Schwarz, "Comparison of Fujisaki-model extractors and F0 stylizers." in *Proc. INTERSPEECH 2009*, 2009, pp. 2455–2458.

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

[16] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech F0 contours." in *Statistical And Perceptual Audition*, 2010, pp. 43–48.

[17] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Hidden markov convolutive mixture model for pitch contour analysis of speech." in *Proc. INTERSPEECH 2012*, 2012, pp. 390–393.

[18] K. Kadowaki, T. Ishihara, N. Hojo, and H. Kameoka, "Speech prosody generation for text-to-speech synthesis based on generative model of F0 contours." in *Proc. INTERSPEECH 2014*, 2014, pp. 2322–2326.

[19] K. Tanaka, H. Kameoka, T. Toda, and S. Nakamura, "Statistical F0 prediction for electrolaryngeal speech enhancement considering generative process of F0 contours within product of experts framework," in *Proc. ICASSP 2016*. IEEE, 2016, pp. 5665–5669.

[20] K. Hirose, Y. Furuyama, S. Narusawa, N. Minematsu, and H. Fujisaki, "Use of linguistic information for automatic extraction of f_0 contour generation process model parameters." in *Proc. INTERSPEECH 2003*, 2003.

[21] K. Hirose, Y. Furuyama, and N. Minematsu, "Corpus-based extraction of F0 contour generation process model parameters." in *Proc. INTERSPEECH 2005*, 2005, pp. 3257–3260.

[22] H. Torres and J. Gurlekian, "Parameter estimation and prediction from text for a superpositional intonation model," in *Proc. 20 Konferenz Elektronische Sprachsignalverarbeitung*, 2009, pp. 238–247.

[23] H. M. Torres, J. A. Gurlekian, H. Mixdorff, and H. Pfitzinger, "Linguistically motivated parameter estimation methods for a superpositional intonation model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 28, 2014.

[24] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks." in *Proc. INTERSPEECH 2011*, 2011, pp. 437–440.

[25] R. Sato, H. Kameoka, and Y. Kashino, "Fast algorighm for statistical phrase/accent command estimation based on generative model incorporating spectral features (in Japanese)," *IEICE Technical Report, Spoken Language Processing*, vol. 2016, no. 11, pp. 1–6, 2016.

[26] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[27] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP 2013*. IEEE, 2013, pp. 7962–7966.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.