INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Discriminative Method to Extract Coarse Prosodic Structure and Its Application for Statistical Phrase/Accent Command Estimation

Yuma Shirahata, Daisuke Saito, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo, Japan
{shirahata, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

This paper introduces a method of extracting coarse prosodic structure from fundamental frequency (F_0) contours by using a discriminative approach such as deep neural networks (DNN), and applies the method for the parameter estimation of the Fujisaki model. In the conventional methods for the parameter estimation of the Fujisaki model, generative approaches, in which the estimation is treated as an inverse problem of the generation process, have been adopted. On the other hand, recent development of the discriminative approaches would enable us to treat the problem in a direct manner. To introduce a discriminative approach to the parameter estimation of the Fujisaki model in which the precise labels for the parameter are expensive, this study focuses on the similarities between the acoustic realization of the prosodic structure in F_0 contours and the sentence structure of the read text. In the proposed method, the sentence structure obtained from the text is utilized as the labels for the discriminative model, and the model estimates the *coarse* prosodic structure. Finally this structure is refined by using a conventional method for the parameter estimation. Experimental results demonstrate that the proposed method improves the estimation accuracy by 18% in terms of detection rate without using any auxiliary features at inference.

Index Terms: the Fujisaki model, SPACE, voice F_0 contours

1. Introduction

In the field of text-to-speech (TTS), the quality of synthesized speech has improved dramatically since the advent of WaveNet [1], and more recently, the end-to-end model has achieved naturalness almost equivalent to natural speech [2]. However, the control of para/non-linguistic information such as emotion and speaking style in TTS has not yet reached that level. Prosodic information, especially the fundamental frequency (F_0), has a large impact on these para/non-linguistic information. Therefore, modeling F_0 is still very important for TTS, dialogue systems, and voice conversion.

The Fujisaki model is a well-founded mathematical model that formulates an F_0 contour as the superposition of phrase and accent components, which correspond to the pitch variation of phrase units and those of accent units, respectively [3]. These components are controlled by the parameters of Fujisaki model, i.e., the positions and the magnitudes of impulse-like phrase commands and stepwise accent commands. Since this model incorporates the process of human speech with an explicit formula and can approximate F_0 contours of real utterances with a small number of parameters, it has been applied to various languages and shown its validity [4, 5, 6, 7, 8, 9].

The estimation of the Fujisaki model commands from raw F_0 contours is not an easy problem, because parameters should approximate the observed F_0 contour while meeting the constraint imposed in the Fujisaki model. Recently, however, a powerful method for the parameter estimation, called SPACE, which translates the Fujisaki model into a probabilistic generative model, has been proposed [10]. Since the Fujisaki model is a generative model of F_0 contours, this method treats the estimation problem of the parameters as an inverse problem of the generation process. In order to further improve this method, researchers have focused on the close relationship between the command of the Fujisaki model and linguistic information, and proposed methods combining SPACE with auxiliary linguistic or spectral features [11, 12, 13, 14].

On the other hand, the remarkable development of discriminative methods in recent years would make it possible to solve the problem of the parameter estimation in a direct manner. Generally speaking, a large amount of labeled data which directly correspond to a target problem is required for adopting a discriminative method. However the precise labels for the parameters of the Fujisaki model is quite expensive. These labels have been manually annotated by professionals of speech prosody field. Preparing a large amount of the target information enables us to obtain the benefits of discriminative models.

In this paper, we investigate the integration of a discrimnative approach with the SPACE method. To address the data-hungry issue in discriminative approaches, we focus on the similarities between the prosodic structure of a speech and the syntactic structure of the corresponding text, and used the syntactic structure obtained from text as the target labels of the discriminative model. As this model ignores the aspect of the Fujisaki model as a generative model and the target of the model is obtained only from text, the output is coarse prosodic structure. To obtain the finer prosodic structure, the output of the discriminative model is integrated with SPACE. For the discriminative model, SPACE method is regarded as a refiner of the coarse prosodic structure. On the other hand, for SPACE, the output of the discriminative models is viewed as the auxiliary linguistic information. In the proposed approach, the generative and the discriminative approach are complementary to each other.

The rest of this paper is organized as follows: Section 2 briefly introduces the original Fujisaki model and the parameter estimation framework, SPACE. Section 3 presents the proposed method using a discriminative approach for the parameter estimation of the Fujisaki model. Section 4 describes the experimental evaluations. Section 5 concludes this paper.

2. Generative Model of Speech F_0 Contours

2.1. Fujisaki Model

The Fujisaki model formulates an F_0 contour in logarithmic scale, $y[k]$, where k is time, as the superposition of a phrase component $x_{\mathrm{p}}[k]$, an accent component $x_{\mathrm{a}}[k]$ and a base component μ_{b}:

$$y[k] = x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + \mu_{\mathrm{b}}. \tag{1}$$

The phrase component $x_{\mathrm{p}}[k]$ represents long-scale pitch variations over the duration of prosodic units, and the accent com-

http://dx.doi.org/10.21437/Interspeech.2020-2566

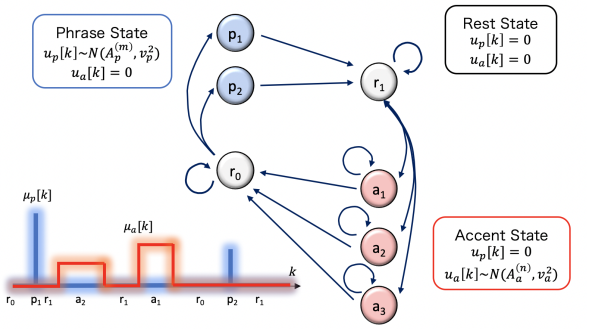

Figure 1: *The state transition topology of the command sequence.* p, a, r *denotes phrase states, accent states and rest states, respectively.*

ponent $x_{\mathrm{a}}[k]$ represents relatively short-scale pitch variations in accent units. μ_{b} is a constant value which represents the lower bound of the speaker's $\log F_0$. The phrase and accent components are generated by second-order, critically-damped linear filters $G_{\mathrm{p}}[k]$ and $G_{\mathrm{a}}[k]$ in response to an impulse-like phrase command $u_{\mathrm{p}}[k]$ and a stepwise accent command $u_{\mathrm{a}}[k]$, respectively:

$$x_{\mathrm{p}}[k] = G_{\mathrm{p}}[k] * u_{\mathrm{p}}[k], \tag{2}$$

$$x_{\mathrm{a}}[k] = G_{\mathrm{a}}[k] * u_{\mathrm{a}}[k]. \tag{3}$$

In these equations, $*$ denotes convolution over time.

2.2. Stochastic formulation of F_0 Contours model (SPACE)

In this section, we briefly review a conventional powerful framework for the command estimation of the Fujisaki model, called SPACE [10]. In SPACE, $\boldsymbol{u}[k] = (u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^\top$ is treated as a model parameter, which is emitted from the path-restricted hidden Markov model (HMM) illustrated in Figure 1. To model the duration of rest states and accent states, each state is split into frame-level sub states. The output distribution of each HMM state is a Gaussian distribution:

$$\boldsymbol{u}[k] \sim \mathcal{N}\left(\boldsymbol{u}[k]; \boldsymbol{\mu}[k], \boldsymbol{\Sigma}\right), \tag{4}$$

where $\boldsymbol{\mu}[k] = (\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k])^\top, \boldsymbol{\Sigma} = \mathrm{diag}(v_{\mathrm{p}}^2, v_{\mathrm{a}}^2)$ are the mean vector and the covariance matrix of the output distribution of state in time k, respectively. The output sequence of the above HMM $u_{\mathrm{p}}[k], u_{\mathrm{a}}[k]$ is then convoluted with different second-order filters $G_{\mathrm{p}}[k]$ and $G_{\mathrm{a}}[k]$, to generate the phrase and accent component $x_{\mathrm{p}}[k], x_{\mathrm{a}}[k]$ as described in (2) and (3), respectively. The logarithmic F_0 contour $x[k]$ is then derived from (1).

In order to incorporate the uncertainty of observed F_0 contours, an observed F_0 contour $y[k]$ is modeled as the superposition of the above $x[k]$ and a noise component $x_{\mathrm{n}}[k] \sim \mathcal{N}\left(0, v_{\mathrm{n}}[k]^2\right)$:

$$y[k] = x[k] + x_{\mathrm{n}}[k]. \tag{5}$$

By marginalizing $x_{\mathrm{n}}[k]$ out, the probability density function of $\boldsymbol{y} = \{y[k]\}_{k=1}^{K}$, given $\boldsymbol{u} = \{u[k]\}_{k=1}^{K}$, can be written as follows:

$$P\left(\boldsymbol{y} \mid \boldsymbol{u}\right) = \prod_{k=1}^{K} \mathcal{N}\left(y[k]; x[k], v_{\mathrm{n}}^2[k]\right). \tag{6}$$

The parameters $\boldsymbol{u}$ and $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ denotes the parameters of the HMM, are optimized by locally maximizing $P\left(\boldsymbol{u}, \boldsymbol{\theta} \mid \boldsymbol{y}\right)$ using EM algorithm. The details of this process are described in [10].

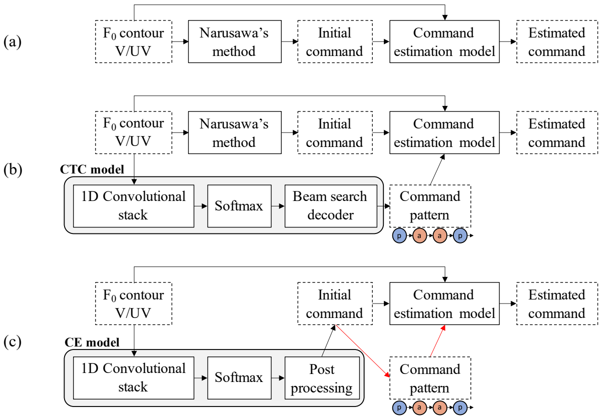

Figure 2: *Overview of the proposed model. (a) SPACE [10]. (b) FIX model. (c) INIT model and INIT_FIX model. Only INIT_FIX model uses the red lines.*

3. Discriminative Approaches for Command Estimation of the Fujisaki Model

3.1. Overview

The command estimation of the Fujisaki model is equivalent to the maximization problem of $P\left(\boldsymbol{u} \mid \boldsymbol{y}\right)$ with respect to $\boldsymbol{u}$. This section introduces methods for the parameter estimation of the Fujisaki model using discriminative approaches, which solve the problem in a direct manner. However, building a model that receives $\boldsymbol{y}$ as input and emits the parameters $\boldsymbol{u}$ involves two problems. The first problem is that the precise labels for the parameters of the Fujisaki model are quite expensive. To avoid this data-hungry problem, this study focuses on the similarity between the acoustic realization of prosodic structure and the sentence structure of the read text, and uses the sentence structure obtained from text analysis as the target of the model. The second problem is that since a discriminative model does not consider the generation process, there would be mismatch between the observed F_0 contours and the reconstructed ones from the obtained parameters. To address it, this study adopts a discriminative model that outputs coarse prosodic structure and then refines this structure using SPACE. As the target labels obtainable from text, this study investigated two types of structure: the order of phrase/accent commands and the probability of command type (phrase, accent, rest) at each frame. These types of coarse structure are utilized as the state transition topology and the initial commands of SPACE, respectively. The discriminative models that correspond to the former is referred to as connectionist temporal classification (CTC) model and the latter as cross entropy (CE) model. The overview of the proposed model is illustrated in Figure 2.

3.2. CTC model

The command occurrence in the naive SPACE algorithm is very sensitive to the hyperparameter v_{n} and the initial commands, since there is no constraint that restricts the number and the order of commands. As a proper constraint, the order of the command occurrence (command pattern) is expected to help SPACE to obtain more accurate command sequences. If the occurrence of commands is restricted by a proper command pattern, the estimated command sequence keeps the appropriate number and the order, which leads to stable command estimation.

Command patterns correspond to the state transition topology in SPACE, which is described in Figure 1. Therefore, in order to integrate the constraint of the command pattern into SPACE, it is necessary to transform the state transition topology to a left-to-right one. In this case, SPACE estimates the timing and the duration of each command in the pattern. Since it is reported that the hard E-step accelerates the estimation process and slightly increases the estimation accuracy, it is implemented by adopting the hard E-step [11], which replaces the E-step of EM algorithm with a point estimation procedure. To obtain command patterns from F_0 contours, it is a reasonable approach to define the loss of the output sequence by using CTC in the same way as the speech recognition [15], since F_0 contours are continuous and command patterns are discrete. Hereinafter, this pattern estimation model is referred to as "CTC model" and the command estimation model whose transition topology is fixed by the output of CTC model is referred to as "FIX model".

3.3. CE model

Since the parameter optimization of SPACE adopts EM algorithm, its results of command estimation converge to a local optimal solution and strongly depend on the initial commands. In most of methods utilizing SPACE, the initial commands are calculated by Narusawa's method [16]. This method reasonably extracts commands from F_0 contours, but its performance is limited due to its strong hypothesis that F_0 contours can be approximated by third-order polynomials.

On the other hand, by using the rough positions of commands that are emitted from a discriminative model are expected to perform well as the initial commands of SPACE, since they are generated considering overall shape of an F_0 contour. Hence, we train a discriminative model that outputs the probability of command type at each frame. Since the raw output of this model is just probability and ignores the constraint of the Fujisaki model, the following process is executed for the output: 1) The maximum phrase command probability is detected for each unvoiced regions, and if it is larger than threshold T_p, there exists a phrase command of magnitude A_p. 2) Moving average filter with window size L_a is applied to the accent command probability, and if the value is larger than T_a, there exists an accent command of magnitude A_a. Also, accent commands which are completely included in unvoiced regions or with duration shorter than D_a are deleted. 3) The commands obtained from the above processing are used as the initial commands.

Hereinafter, we denote this model as "CE model", and SPACE with these initial commands is referred to as "INIT model". In addition, it is also possible to automatically obtain the command pattern from the initial commands. As the obtained command pattern has the same number and order as itself, fixing the transition topology of INIT model using the command pattern would further enhance the performance of command estimation. This model is referred to as "INIT_FIX model". Although estimating command magnitude by a regression model is also a possible approach, we investigated only the discriminative model since magnitude is very sensitive to the base component and hard to estimate directly.

3.4. Related works

There have been some approaches that expand SPACE using auxiliary linguistic features. Sato *et al.* incorporated the relationship between the onset of the accent commands and the phoneme boundary into SPACE by using time variation of the spectral features and the phoneme alignment [11, 12]. Hojo *et al.* constructed a DNN that maps frame-level linguistic feature vectors to the state posterior probabilities of the HMM on the basis of DNN-HMM framework [17], to model the relationship between the commands and linguistic information [13]. Our previous study explicitly linked the occurrence of phrase commands to the boundaries of phrase structure to treat phrase structure as the minimal unit of focus control [14]. The proposed method differs from them in that it treats linguistic information as flexible coarse structure, i.e., the transition topology and the initial commands of SPACE, and SPACE itself is only utilized to refine the structure.

4. Experiments

4.1. Experimental conditions

To evaluate the performance of the proposed methods, experiments of command estimation were conducted. To investigate the effects of the training data, two different datasets were prepared: 450 utterances of the ATR Japanese sentence database B-set spoken by a male speaker (0.5 hours), *ATR* henceforth [18], and the JVS corpus spoken by 49 male speakers (12.4 hours), *JVS* henceforth [19]. The other 53 utterances of the ATR database were used for the evaluation. F_0 contours were extracted by Kameoka's method [20].

To prepare the target labels of the discriminative models, two types of procedures were adopted. The first procedure is *TEXT*, where the labels are automatically prepared from pairs of speech and text in the following way: 1) phone alignment and text analysis are executed using Julius and Open JTalk [1], respectively [21]. 2) Phrase and accent commands are allocated to the pause positions and the accented morae, respectively. The second procedure is *MANUAL*, where a professional of speech prosody field manually annotates the ground-truth commands. The purpose of training models with *MANUAL* labels is to know the upper bound of the performance of the proposed models. From these commands obtained from the above procedures, one-hot labels for 3 states in frame level for CE model and command pattern labels for CTC model were generated.

For both CE and CTC models, the input consisted of two dimensions, which were $\log F_0$ and V/UV information. To remove the influence of the difference of base component μ_b, the lowest value in voiced region was subtracted from $\log F_0$ for each utterance. The model architecture was CNN of 4 layers with RELU as activation function. The window size was set at 101, 81, 15, 5, respectively, which was also common between the two models. As for CE model, the loss of phrase states was weighted by 20 times because the occurrence of them is less frequent than the other states. The parameters were set at $T_\mathrm{p} = 0.3, T_\mathrm{a} = 0.5, A_\mathrm{p} = 0.4, A_\mathrm{a} = 0.5, L_\mathrm{a} = 40$ ms, $D_\mathrm{a} = 16$ ms, respectively. The hyperparameters of SPACE were fixed at the same value as [10]. For FIX and INIT_FIX model, $v_\mathrm{n} = 0.03$ was adopted, instead of 0.2.

The accuracy of command pattern was calculated as the mean of Levenshtein distance between the estimated pattern and the ground-truth one. The performance of command estimation was measured on the basis of two criteria: $\log F_0$ RMSE (root mean squared error) and detection rates. The detection rate is a measure of how accurate the positions of the estimated commands are, and calculated in the following way: First, matching between the estimated and ground-truth commands is performed using dynamic programming algorithm. If the time difference between an estimated command and a ground truth one is shorter than predefined tolerance S, the estimated command is considered "matched" and the local distance is set at 0. Otherwise the local distance is 1. The time difference of two accent

[1] http://open-jtalk.sp.nitech.ac.jp

Table 1: *Levenshtein distance between the estimated and the ground-truth command pattern with 95% confidence intervals.*

	JVS_TEXT	ATR_TEXT	ATR_MANUAL
CTC model	2.00 ± 0.27	1.33 ± 0.25	0.96 ± 0.21
CE model	1.27 ± 0.23	1.12 ± 0.24	0.94 ± 0.26
Narusawa		2.47 ± 0.43	

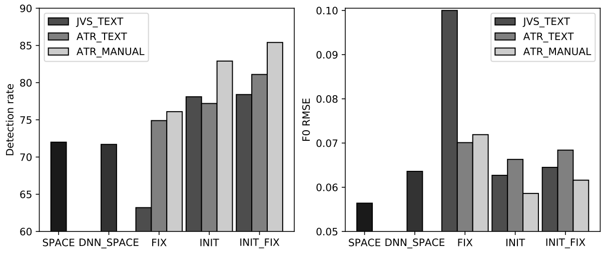

Figure 3: *Result of command estimation. The left and right graph shows the detection rate and* $\log F_0$ *RMSE, respectively. Note that the performance of DNN-SPACE is taken from [13].*

commands is calculated as the average of time difference between the two onsets and two offsets of them. Let N_{E} and N_{A} be the number of commands in the estimated and ground truth sequences, N_{M} be the number of the matched commands between the two sequences. The insertion error E_{I} is defined as $(N_{\mathrm{E}} - N_{\mathrm{M}})/N_{\mathrm{A}}$, the deletion error rate E_{D} as $(N_{\mathrm{A}} - N_{\mathrm{M}})/N_{\mathrm{A}}$, and the detection rate is calculated as $1 - E_{\mathrm{I}} - E_{\mathrm{D}}$. Note that the magnitudes of commands were not evaluated, because estimation of magnitudes is very sensitive to the base component μ_{b}, which is set differently in SPACE and manual annotation.

4.2. Experimental results

Table 1 shows the result of command pattern estimation. We can see that regardless of the target label and the model, the proposed methods estimate closer command patterns to the ground-truth ones than the conventional Narusawa's method. This result shows that command patterns can be estimated from the overall shape of F_0 contours reasonably well by training the discriminative model. Interestingly, it is also shown that CE model can estimate the command pattern more accurately than CTC model, while the latter directly minimizes the loss of the command pattern sequence. The reason for this result may be that CE model is trained with frame-level target labels, which have rich information, while CTC model knows only the command pattern and do not know the exact position of each command.

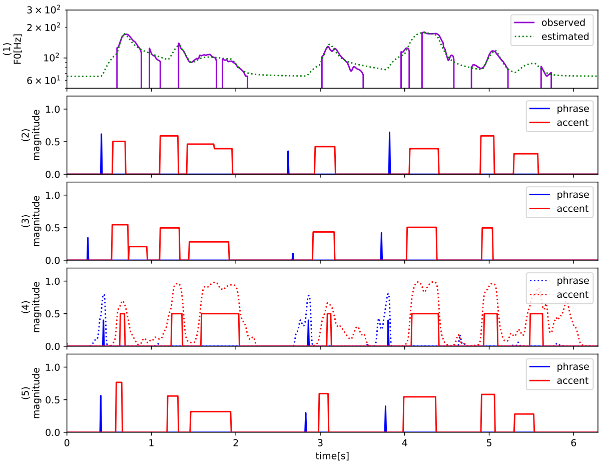

Figure 4: *Example of command estimation by INIT_FIX model trained with ATR_TEXT. (1) An observed* F_0 *contour in voiced regions (solid line) and the estimated one by the proposed model (dotted line). (2) Ground-truth commands. (3) Commands estimated by SPACE [10]. (4) Output of CE model (dotted line) and initial commands (solid line). (5) Commands estimated by the proposed model.*

The result of command detection rate with $S = 0.3$ s is shown in the left side of Figure 3. This figure shows that almost all of the proposed models outperform the conventional methods in terms of command detection rate. This result demonstrates that using coarse prosodic structure estimated by the discriminative model improves the performance of command estimation model. Specifically, INIT and INIT_FIX improved the detection rate by 7% to 15% and 9% to 18% compared to the original SPACE, respectively. From this result we can conclude that obtaining the initial commands of SPACE using the discriminative model helps the command estimation, and the performance is further improved by fixing the transition topology. In terms of the training data, we can see that the model performance improves even if the target label is automatically generated from text (*_TEXT) or CE model is trained speaker-independently (JVS_TEXT). On the other hand, FIX model only improved the performance when CTC model is trained speaker-dependently (ATR_*). This is probably because the length of the pauses and the strength of the accents vary greatly from person to person, and CTC model, which is trained without using time information, is unable to absorb these differences when it is trained speaker-independently.

The right side of Figure 3 shows the experimental result of $\log F_0$ RMSE. We can see that the performance of F_0 reconstruction in the proposed methods is slightly worse than that of the conventional method. This is because estimating the commands from F_0 contours is a ill-posed problem, and even though the estimated command do not correspond to the ground-truth ones, it can approximate the observed F_0 contour.

Figure 4 shows an example of command estimation. We can see that CE model outputs reasonable values considering the overall shape of the F_0 contour, and the estimated command by the proposed model is very similar to the ground-truth one, while conventional SPACE has an insertion error around $t = 0.8$ and a deletion error around $t = 5.4$.

5. Conclusions

This paper has introduced a method to extract the parameters of Fujisaki model from speech signals using a discriminative approach. To avoid the data-hungry problem, this study focused on the similarities between the prosodic structure of a speech and the sentence structure of the corresponding text, and utilized the sentence structure obtained from text as the target labels of the discriminative model. To refine the obtained coarse prosodic structure, a conventional powerful framework for the parameter estimation was adopted. Experimental results revealed that the proposed method improved estimation accuracy by 18% in terms of the command detection rate without utilizing any auxiliary features at inference, and even if the target labels were generated automatically from text and the model was trained speaker-independently, the improvement was 9% compared to the conventional method. For further works, the application of the proposed approach for TTS should be investigated.

6. References

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] H. Fujisaki, M. Ljungqvist, and H. Murata, "Analysis and modeling of word accent and sentence intonation in Swedish," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1993, pp. 211–214.

[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 *Sixth International Conference on Spoken Language Processing*, 2000.

[9] W. Gu, K. Hirose, and H. Fujisaki, "Analysis of F0 contours of Cantonese utterances based on the command-response model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[10] 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.

[11] R. Sato, H. Kameoka, and K. Kashino, "Fast algorithm for statistical phrase/accent command estimation based on generative model incorporating spectral features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5595–5599.

[12] R. Sato and K. Kashino, "Statistical phrase/accent command estimation algorithm utilizing linguistic information," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5569–5573.

[13] N. Hojo, Y. Ohsugi, Y. Ijima, and H. Kameoka, "DNN-SPACE: DNN-HMM-based generative model of voice F0 contours for statistical phrase/accent command estimation." in *INTERSPEECH*, 2017, pp. 1074–1078.

[14] Y. Shirahata, D. Saito, and N. Minematsu, "Generative modeling of F0 contours leveraged by phrase structure and its application to statistical focus control," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 228–233.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[16] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *proceedings of ICASSP*, vol. 1, 2002, pp. I–509.

[17] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[18] 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.

[19] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[20] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and F0 contour generating process model," *IEICE Technical Report*, vol. 110, pp. 29–34, 2010.

[21] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *EUROSPEECH2001*, 2001, pp. 1691–1694.