# FRAME-LEVEL ACOUSTIC MODELING BASED ON GAUSSIAN PROCESS REGRESSION FOR STATISTICAL NONPARAMETRIC SPEECH SYNTHESIS

*Tomoki Koriyama, Takashi Nose, Takao Kobayashi*

Tokyo Institute of Technology
Interdisciplinary Graduate School of Science and Engineering
4259-G2-4 Nagatsuta-cho, Midori-ku, Yokohama-City, 226-8502, Japan

## ABSTRACT

This paper proposes a new approach to text-to-speech based on Gaussian processes which are widely used to perform non-parametric Bayesian regression and classification. The Gaussian process regression model is designed for the prediction of frame-level acoustic features from the corresponding frame information. The frame information includes relative position in the phone and preceding and succeeding phoneme information obtained from linguistic information. In this paper, a frame context kernel is proposed as a similarity measure of respective frames. Experimental results using a small data set show the potential of the proposed approach without state-dependent dynamic features or decision-tree clustering used in a conventional HMM-based approach.

***Index Terms—*** acoustic models, statistical speech synthesis, Gaussian process regression, non-parametric Bayesian model, context kernel

## 1. INTRODUCTION

Statistical parametric speech synthesis has been developed for over a decade. Utilization of hidden Markov model (HMM) in speech synthesis, which is one of the most powerful generative models to express time-series information like speech signals, has enabled us to provide not only the naturalness of synthetic speech but also the distinctive techniques for parametric approach such as average voice models and speaker and style adaptation [1, 2].

In the HMM-based speech synthesis, phones are partitioned into some hidden states of HMMs, and the static and dynamic acoustic features are parameterized by their means and variances for each context-dependent state. However HMM is not always an appropriate model for acoustic features to be synthesized. For example, in spite of the continuously changing characteristics of the acoustic features, the hidden-state space is discrete. Although dynamic features enable us to generate smoothly changing feature sequence from the discrete states, the parametric representation of acoustic features has a limitation. In fact, a fixed number of state-dependent dynamic features fail to generate some short-time variations. Moreover, there is a problem in context-dependent decision-tree clustering that aims to make robust models to unseen contexts. Averaging the acoustic features of a leaf node of the tree causes over-smoothing effect, and the synthetic speech samples tend to be muffled.

To overcome the limitations of parametric models, we consider here Gaussian processes (GPs) known as non-parametric Bayesian models for regression and classification [3]. “Non-parametric” implies that the model complexity grows with the increase of data size. This leads to an advantage that GP has a flexibility for the complexity of the model. GP also has a robustness against over-fitting by Bayesian inference. In addition, since GP is a kernel method, various kinds of data can be used as input variables by defining the kernel function of respective samples [4]. In recent years, some approaches using GP have been proposed to speech processing such as speech enhancement [5], voice conversion [6], and speech representation [7]. Henter et al. [7] challenged the problem of the state discreteness. They expanded the discrete states to continuous variables of a latent space and assumed the GP on a frame-level function that transforms the latent space variables into the acoustic features. Specifically, Gaussian process dynamical model (GPDM) was used to express the latent space. However, it is not easy to apply GPDM to text-to-speech directly because of the difficulty in correlating the latent space variables with the linguistic information of a given input sentence to be synthesized.

In this paper, we propose an acoustic modeling technique for speech synthesis based on the GP regression. We use GP on a frame-level function in the same way as [7] except that frame-level information obtained from linguistic information is transformed into acoustic features by the frame-level function. The significant differences from the conventional HMM-based method are that the proposed approach is the frame-level regression and it does not use the states or dynamic features, and it avoids a tree-based clustering because the kernel of contexts can be the alternative to the clustering. We construct kernels for frame information and evaluate the effectiveness through the experiments using a small size of data set.

## 2. GAUSSIAN PROCESS FOR REGRESSION

Suppose that we have $N$ observations as a training data set $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, N\}$, where $\mathbf{x}_i$ is a vector consisting of explanatory variables, and $y_i$ is an output variable. We assume that $y_i$ is given by

$$y_i = f(\mathbf{x}_i) + \varepsilon, \tag{1}$$

where $f(\mathbf{x}_i)$ is noise free observation and $\varepsilon$ represents a Gaussian noise of $\mathcal{N}(0, \sigma_n^2)$. Let $\mathbf{y} = [y_1, \dots, y_N]^\top$ and $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top$ be matrix forms of all training data.

When $f(\mathbf{x})$ is a Gaussian process, the GP prior is given by

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_n^2\mathbf{I}), \tag{2}$$

where $\mathbf{K}$ is a Gram matrix whose element is given by

$$K_{mn} = k(\mathbf{x}_m, \mathbf{x}_n), \tag{3}$$

and $k(\mathbf{x}_m, \mathbf{x}_n)$ is a kernel function, which is also called a covariance function.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

The goal of GP regression is to infer the continuous distribution of $y_*$ given a new input vector $\mathbf{x}_*$. The joint distribution of $\mathbf{y}$ and $y_*$ is given by

$$p\left(\begin{bmatrix}\mathbf{y}\\ y_*\end{bmatrix} \,\middle|\, \mathbf{X}, \mathbf{x}_*\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K}+\sigma_n^2\mathbf{I} & \mathbf{k}_*\\ \mathbf{k}_*^\top & k(\mathbf{x}_*, \mathbf{x}_*)+\sigma_n^2\end{bmatrix}\right), \tag{4}$$

where $\mathbf{k}_*$ is the column Gram vector that has the element $k(\mathbf{x}_n, \mathbf{x}_*)$ $(n = 1, \ldots, N)$. The predictive distribution for the new observation can be obtained by the following conditional distribution:

$$p(y_*|\mathbf{y}, \mathbf{X}, \mathbf{x}_*) = \mathcal{N}(\mu_*, \sigma_*^2) \tag{5}$$

$$\mu_* = \mathbf{k}_*^\top(\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{6}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top(\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}\mathbf{k}_* + \sigma_n^2. \tag{7}$$

The inversion of $(\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}$ requires $O(N^3)$ computations. For the practical implementation, the parameter vector

$$\boldsymbol{\alpha} = (\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{8}$$

that depends only on the training data set is calculated in the training step. When generating the new target mean, we compute the inner product

$$\mu_* = \mathbf{k}_*^\top\boldsymbol{\alpha}, \tag{9}$$

which requires $O(N)$ computational cost.

In order to make the model by GP, we need to construct the kernel function. The requirement for the kernel function is that the Gram matrix should be positive semi-definite and symmetric. In this paper we use typical kernels, a square exponential (SE) kernel and a linear kernel. The SE kernel is the most widely used stationary kernel as the measure of "similarity" of two input vectors. For one-dimensional input, the SE kernel is defined by

$$k(x_m, x_n) = \exp\left(-\frac{(x_m - x_n)^2}{l^2}\right), \tag{10}$$

where $l$ denotes a length-scale hyper-parameter. The predictive mean of new input is obtained by the weighted sum of nearby samples [3]. The linear kernel is given by

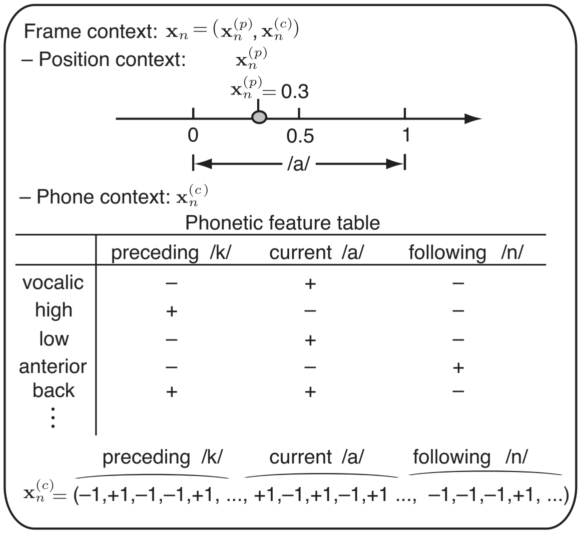
$$k(x_m, x_n) = x_m \cdot x_n. \tag{11}$$

This kernel assumes the linearity between output and input features. In addition, it should be noted that it is possible to construct a new kernel by combining multiple arbitrary kernel functions by means of some operations such as sum, product, and convolution [4].

## 3. ACOUSTIC MODELING USING GP REGRESSION

### 3.1. Frame context

In this paper, we consider a simple set of input features as an initial step of the proposed approach. For the explanatory input variables of the regression model, frame-level features obtained from linguistic information are used. We refer to these features as *frame context*. Figure 1 shows an example of the frame context. The frame context consists of position and phone context. For the position context, the relative frame position in the phone is employed where the beginning and end of the phone are set to 0 and 1, respectively. For the phone context, we use a set of preceding, current, and succeeding phonetic features. We introduce binary variables $(\{\text{positive} = +1, \text{negative} = -1\})$ for each phonetic feature listed in Table 1 based on a balanced distinctive phonetic feature set [8]. Let $M$ be the number of phonetic features, then $3M$-dimensional binary-valued vector is constructed.


Frame context: $\mathbf{x}_n = (\mathbf{x}_n^{(p)}, \mathbf{x}_n^{(c)})$
– Position context: $\mathbf{x}_n^{(p)}$
$\mathbf{x}_n^{(p)}$= 0.3
0
0.5
1
/a/
– Phone context: $\mathbf{x}_n^{(c)}$
Phonetic feature table

| | preceding /k/ | current /a/ | following /n/ |
|---|---|---|---|
| vocalic | – | + | – |
| high | + | – | – |
| low | – | + | – |
| anterior | – | – | + |
| back | + | + | – |
| ⋮ | | | |

preceding /k/
current /a/
following /n/
$\mathbf{x}_n^{(c)}$= (–1,+1,–1,–1,+1, ..., +1,–1,+1,–1,+1 ..., –1,–1,–1,+1, ...)

**Fig. 1**. Example of frame context, i.e., a frame-level input variable set for the GP regression. This example shows the frame context for a frame positioned in phone /a/, which is in between preceding phone /k/ and following phone /n/.

**Table 1**. Binary phonetic features.

| Phonetic features |
|---|
| vocalic, high, low, anterior, back, coronal, plosive, affricative, continuant, voiced, nasal, semi-vowel, silent |

### 3.2. Frame context kernel

A proposed frame context kernel is defined as a product of two kernels.

$$k(\mathbf{x}_m, \mathbf{x}_n) = k_p(\mathbf{x}_m^{(p)}, \mathbf{x}_n^{(p)})k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)}), \tag{12}$$

where $k_p(\mathbf{x}_m^{(p)}, \mathbf{x}_n^{(p)})$ and $k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)})$ denote position kernel and phone context kernel, respectively. The position kernel represents the similarity of the positions in the phones whereas the phone context kernel represents that of the phone context.

#### *3.2.1. Position kernel*

The SE kernel is used for position kernel and is given by

$$k_p(\mathbf{x}_m^{(p)}, \mathbf{x}_n^{(p)}) = \exp\left(-\frac{(\mathbf{x}_m^{(p)} - \mathbf{x}_n^{(p)})^2}{l_p^2}\right), \tag{13}$$

where $p_n$ is the relative position of the $n$-th frame.

#### *3.2.2. Phone context kernel*

We examine two different phone context kernels in this paper. One is sum of SE kernels and the other is a linear kernel. The former one is defined by

$$k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)}) = \sum_{i=1}^{3M}\theta_{ci}^2 \exp\left(-\frac{(x_{m,i}^{(c)} - x_{n,i}^{(c)})^2}{l_{ci}^2}\right), \tag{14}$$

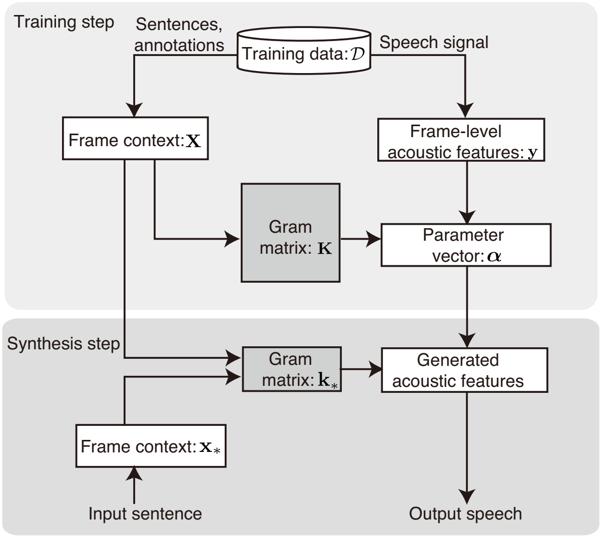

Training step
Sentences, annotations
Training data: $\mathcal{D}$
Speech signal
Frame context: $\mathbf{X}$
Frame-level acoustic features: $\mathbf{y}$
Gram matrix: $\mathbf{K}$
Parameter vector: $\boldsymbol{\alpha}$
Synthesis step
Gram matrix: $\mathbf{k}_*$
Generated acoustic features
Frame context: $\mathbf{x}_*$
Input sentence
Output speech

**Fig. 2**. An outline of speech synthesis process in the proposed approach.

where $l_{ci}$ is a scale hyper-parameter and $\theta_{ci}$ is a hyper-parameter that represents the relevance of the $i$-th phonetic feature. The kernel value becomes maximum when the input phones are the same. By multiplying this by the position kernel, the generated parameter of a new frame results in the weighted sum of the training samples whose values of input variables are close to those of the new frame.

The linear kernel is given by

$$k_c(\mathbf{x}_m^{(c)}, \mathbf{x}_n^{(c)}) = \sum_{i=1}^{3M} \theta_{ci}^2 x_{m,i}^{(c)} x_{n,i}^{(c)}. \tag{15}$$

This kernel assumes that the acoustic features in the same position exist on a hyperplane in the $3M$-dimensional phonetic feature space.

### 3.3. Speech synthesis system

Figure 2 shows an outline of the speech synthesis using the frame context kernel. The framework is based on the general GP regression. Training procedure is as follows:

1. Frame-level acoustic features such as mel-cepstral coefficients and fundamental frequency are extracted from the training data.
2. The frame context is extracted from the transcriptions and annotations including phone boundaries of the training data.
3. Gram matrix between the frames of the training data and the parameter vector $\boldsymbol{\alpha}$ in Eq. (8) is calculated.

Synthesis procedure is as follows:

1. The frame context is extracted from the input sentence.
2. Gram matrix between the frames of the training and new input data is calculated.
3. The means of the predictive distribution are calculated by multiplying the Gram matrix $\mathbf{k}_*$ and $\boldsymbol{\alpha}$, and used as generated acoustic features.
4. The output waveform is synthesized using the generated features.

**Table 2**. Spectral distortions of generated parameter sequence using position context. The values represent mel-cepstrum distances [dB].

| Phoneme | HMM | GPR | Phoneme | HMM | GPR |
|---|---|---|---|---|---|
| a | 6.02 | 6.08 | k | 6.02 | 5.98 |
| i | 7.11 | 7.09 | t | 4.35 | 4.41 |
| u | 7.18 | 7.16 | n | 6.27 | 6.28 |
| e | 6.04 | 6.07 | s | 5.18 | 5.03 |
| o | 6.48 | 6.48 | m | 5.92 | 5.94 |

**Table 3**. Spectral distortion of generated parameter sequence using frame context. The values represent mel-cepstrum distances [dB].

| Phoneme | HMM | GPR-SE | GPR-linear |
|---|---|---|---|
| a | 5.67 | 5.51 | 5.52 |
| i | 6.01 | 5.64 | 5.63 |
| u | 6.10 | 5.94 | 5.94 |
| e | 5.33 | 5.17 | 5.16 |
| o | 5.90 | 5.63 | 5.64 |
| k | 5.09 | 5.05 | 5.05 |
| t | 4.13 | 4.17 | 4.17 |
| n | 5.73 | 5.81 | 5.81 |
| s | 4.74 | 4.57 | 4.57 |
| m | 5.48 | 5.50 | 5.50 |

## 4. EXPERIMENTS

### 4.1. Experimental conditions

Speech database used in experiments consisted of ATR phonetically balanced Japanese sentences recorded by one female speaker. Speech signals were sampled at a rate of 16kHz. Spectral features were extracted by STRAIGHT [9]. The 0-39th mel-cepstral coefficients were used as target variables. Each dimension of the mel-cepstral coefficients was modeled separately.

To examine the potential of GP regression, we chose 5 vowels (/a/, /i/, /u/, /e/, /o/) and 5 consonants (/k/, /s/, /t/, /n/, /m/), which were primary phonemes of Japanese and resulted in an articulatory balanced set. Each phone was segmented using manually annotated phone boundaries. The segments of training set were randomly chosen up to 10,000 frames from 450 sentences for each phoneme. The test 50 segments were randomly chosen from the remaining 53 sentences. When the test segments were synthesized, the manually annotated boundaries of the original utterances were given.

The following experiments were performed separately for each phoneme. The HMM-based speech synthesis was used as the conventional method. Triphones were used for the context set for the HMM training. The model topology was 5-state, left-to-right, no-skip hidden semi-Markov model (HSMM). Each state had a single Gaussian distribution with a diagonal covariance matrix and the feature vector included delta and delta-delta dynamic features.

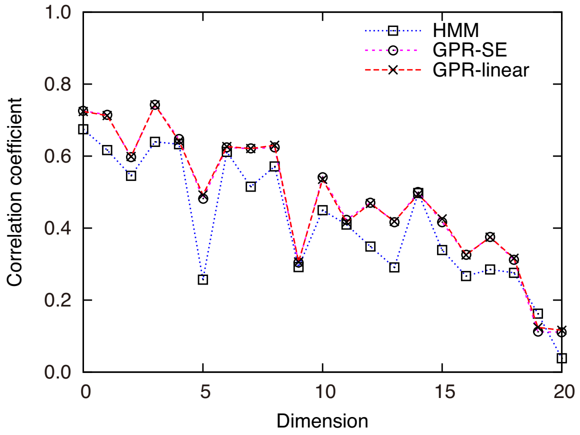


**Fig. 3**. Correlation coefficients between generated and original mel-cepstral coefficients for phoneme /i/.

## 4.2. Results

### 4.2.1. Evaluation on position kernel

To measure the performance of GP regression for generating continuously changing acoustic features, an objective evaluation is performed under a condition where only the position context is given as the input. The kernel for Gaussian process regression corresponded to Eq. (13). All target and explanatory variables are normalized by their means and variances and the hyper-parameters were set to $l_p = 1.0$, $\sigma_n = 1.0$ by preliminary experimental results.

The spectral distortions of the generated parameter sequences from both methods are shown in Table 2. In the table, GPR represents the proposed GP regression. The average mel-cepstral distance was used as the measure of spectral distortion. From the table, it is seen that GP regression has comparable performance with HMM in the generation of continuously changing acoustic features.

### 4.2.2. Evaluation on frame context kernel

The proposed frame context kernels described in Sect. 3.2 were evaluated. We compared the sum of SE kernels and the linear kernel as the phone context kernel. All target and explanatory variables were normalized and the hyper-parameters were given by $l_p = l_{ci} = 1.0$ $(i = 1, \dots, 3M)$, $\sigma_n = 1.0$, and $\theta_{ci} = 1.0/3M$ $(i = 1, \dots, 3M)$ based on preliminary tests. In the case of HMM, triphone HMM was used and decision-tree-based context clustering was performed with a MDL criterion.

Table 3 shows the mel-cepstral distances between the generated and original sequences. GPR-SE and GPR-linear employed the sum of SE kernels and the linear kernel for phone context kernel, respectively. It can be confirmed that phone context reduced the distortions for all methods compared with the case without phone context of Table 2. By comparing GPR with HMM, although there were only small differences for the consonants except /s/, the mel-cepstral distance for the vowels using GPR-SE and GPR-linear decreased significantly without averaging of acoustic features by context-dependent decision-tree clustering used in the HMM-based synthesis. It is also found that the distances of GPR-SE and GPR-linear were comparable. A possible reason is the insensibility of GP to the definition of covariance functions.

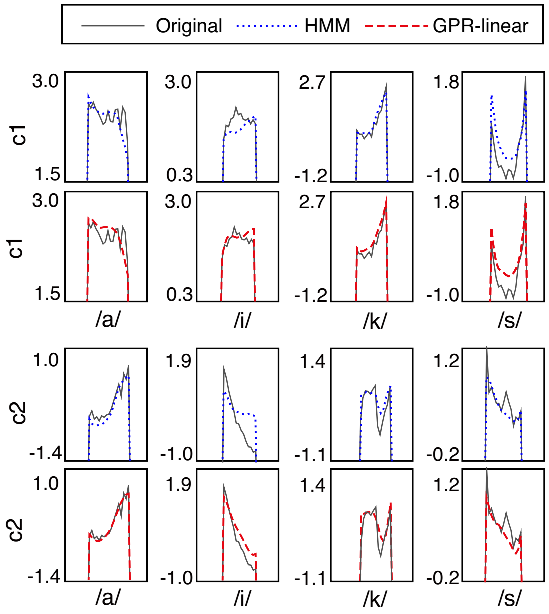


**Fig. 4**. Examples of generated 1st and 2nd mel-cepstral coefficients.

To take a more detailed look, the correlation coefficients between generated and original acoustic features for each mel-cepstral dimension are plotted in Fig. 3. In the figure, the 0-20th dimensions are shown because the correlation coefficients of higher dimensions were too low to discuss. The results of GPR-SE and GPR-linear are very similar and they are almost over-lapped. It is seen that the correlations of GPR-SE and GPR-linear are higher than HMM in most of the dimensions. Figure 4 shows the generated samples of the 1st and 2nd mel-cepstral coefficients for multiple phone context. We can find that GPR-linear could generate the sequences dependently on phone context.

## 5. CONCLUSIONS

We have presented a framework of speech synthesis based on Gaussian process regression and the frame context kernel that represents the similarity of two frame contexts for Gaussian processes. The experiments using a small data set of primary phonemes showed that the proposed method could effectively model the spectral features. For the future work, since the proposed model is currently limited in phone-unit information, we have to expand the model unit to the sentence and incorporate prosodic information for the practical use for text-to-speech. Also since the proposed method needs phone boundary annotations, the effect of the precision of manual or automatic annotation has to be examined. Furthermore, there exists many useful techniques in HMM-based acoustic modeling such as state alignment, therefore, they can be effectively utilized for the regression model.

## 6. ACKNOWLEDGMENTS

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071 and 23700195.

## 7. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2347–2350.

[2] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006.

[4] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*, Cambridge university press, 2004.

[5] S. Park and S. Choi, "Gaussian process regression for voice activity detection and speech enhancement," in *Proc. IJCNN*, 2008, pp. 2879–2882.

[6] N.C.V. Pilkington, H. Zen, and M.J.F. Gales, "Gaussian process experts for voice conversion," in *Proc. INTERSPEECH*, 2011, pp. 2761–2764.

[7] G.E. Henter, M.R. Frean, and W.B. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proc. ICASSP*, 2012, pp. 4505–4508.

[8] Takashi FUKUDA and Tsuneo NITTA, "Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition," *IEICE Trans. Inf. & Syst.*, vol. 87, no. 5, pp. 1110–1118, 2004.

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.