A Model for Infant Acquisition of Spoken Words Using Genetic Algorithm and Fujisaki Model

Tomio Takara
Faculty of Engineering
University of the Ryukyus / Okinawa Polytechnic College
Okinawa, Japan
e-mail: takara@ie.u-ryukyu.ac.jp

Ryoichi Eto
International Systems Development Co. Ltd.
Okinawa, Japan
e-mail: ryoichi.eto@isd.co.jp

***Abstract*—We propose a new model of speech imitation and acquisition process of infants. We regard the vowel space parameters as the articulatory gesture, such as the tongue hump position and the degree of constriction of vowels. We represent the coarticulation effect using Fujisaki's generative model of speech. We model a trial and error process of the infant's speech imitation using the Genetic Algorithm (GA). In our model, we regard "command" in the Fujisaki model as the articulatory gesture and detect it from the spectral sequence using the GA. In other words, the original phonemic target is inversely estimated as the Fujisaki's command from the phonemically ambiguous speech spectrum caused by the coarticulation. Our model simulates the hypothesis that human infants acquire the normalization (inverse estimation) skill of the coarticulation through the process of imitating spoken words. We evaluated the model in listening tests using synthesized speech. We also show that the model can represent the phenomenon of "predicted sound", which is unconsciously heard as the effect of the normalization of the coarticulation, by comparing this predicted sound with the inversely estimated sound.**

Keywords- model for acquisition of spoken word; coarticulation; Fujisaki model; genetic algorithm; vowel space parameter

I. INTRODUCTION

Human infants learn articulatory behaviors, such as the tongue hump position and the degree of constriction, which are human internal function, by controlling their speech organs and imitating other people's speech. In this manner, infants acquire mapping skills between perception and articulation of speech [1].

We think that this mapping is the most important factor of human speech communication because it means that not only the other people's internal information but their intention is known. We think also that the skill of normalization of coarticulation, which is to grasp clearly ambiguous speech caused by the coarticulation, is acquired in this imitation training process.

However, there has been no simulation research on infant's acquisition process of spoken words using perception and production models combined whereas only the motor theory has been existed [4].

In this paper, we propose a new model of this speech imitation and acquisition process of infants. We show result of listening tests on the coarticulation effect, and they can be understood by our proposed model.

We regard the vowel space parameters [2][3] as the articulatory gesture [4], such as the tongue hump position and the degree of constriction of vowels. We represent the coarticulation effect [5][6] using Fujisaki's generative model of speech (Fujisaki model) [7]. We model a trial and error process of the infant's speech imitation using the Genetic Algorithm (GA) [8]. The papers [5] and [6] used the Fujisaki model, however, were not the acquisition models.

In our model, we regard "command" in the Fujisaki model as the articulatory gesture and detect it from the spectral sequence using the GA. In other words, the original phonemic target is inversely estimated as the Fujisaki's command from the phonemically ambiguous speech spectrum caused by the coarticulation.

Our model also simulates the hypothesis that human infants acquire the normalization (inverse estimation) skill of coarticulation through the process of imitating spoken words. We evaluate the model in listening tests using synthesized speech. We also show that the model can represent the phenomenon of "predicted sound", which is unconsciously heard as the effect of the normalization of the coarticulation, by comparing this predicted sound to the inversely estimated sound.

In section II, we describe our model of speech production and perception. In section III, we describe our model to imitate the speech production process. In section IV, we describe results of analysis of the coarticulation by our model and comparison of them to listening tests. Section V is the conclusion of the paper.

II. THE MODEL OF SPEECH PRODUCTION AND PERCEPTION

In this section, we describe parts of the proposed model of speech production and perception: speech analysis synthesis system, the vowel space parameter, Fujisaki model, and the genetic algorithm.

 ISBN: 978-1-61208-768-9

A. Speech analysis synthesis system

Spectral envelopes are extracted by the improved cepstral method [9] with sampling frequency of 10 kHz, frame length of 25.6ms and frame shift of 10ms. Speech is synthesized using the Log Magnitude Approximation (LMA) filter [10].

B. The vowel space parameter

Principal vectors of the principal component analysis [11] are calculated from many log amplitude spectra, which we call simply "spectra", of isolated vowels. Spectrum at each frame of a spoken word is transformed to components on the axis of principal vectors. We call this space constructed by the principal vectors as the vowel space and the components on the principal vectors as the vowel space parameters [2][3]. The vowel space can represent effectively the space where the vowel spectra vary widely because a principal axis with lager in the spectral space. Therefore, we can reconstruct a spectrum using only the components at the principal vectors with larger eigen values. The vowel space parameters were shown to have much more compressed information than that of the cepstra and can sufficiently express the phonemic characteristics of consonants [2][3].

Figure 1 shows the distribution of the vowel space parameters of isolated vowels spoken by a Japanese male speaker. The central frames of the isolated vowels were analyzed to be log magnitude spectra. Using these spectra, the principal component analysis was performed. These spectra were analyzed using the resulting principal components. This figure is very similar to the distribution of isolated vowels at two dimensional space of the first and the second Formant frequencies. This figure also corresponds very much to the place of vowels expressed in the tongue hump position and the degree of constriction. Therefore, we can use the vowel space parameters in place of the Formant frequencies or the articulatory gesture. Formant frequency cannot be detected always correctly whereas the vowel space parameter is not and can have higher order parameters than the Formant.

C. Fujisaki model

The Fujisaki model [7] is an effective model for approximating the contour of the fundamental frequency precisely for the source model of speech synthesis.

The phrase command and one accent command of the Fujisaki model were used in our proposed model to imitate information of the source. In the model of imitating information of the articulatory position, only accent commands were used and connected to be a command pattern [5]. Fujisaki model was originally applied to the fundamental frequency and then to the Formant frequency [5]. The vowel space parameter is similar to the Formant frequency. Therefore, we can apply Fujisaki model to the vowel space parameter.

D. Genetic algorithm

The (GA) [8] is a searching algorithm which simulates biological evolution. *N* individuals with chromosomes made by random numbers are generated first and they become the first generation. The fitness is decoded from the chromosomes using expression for each model. We adopted the ranking and the elite strategies in the selection mode. The cross over and mutation are carried out and they become a new generation. This algorithm was repeated to be 10th generation in this study. We adopted the Real number Coded Genetic Algorithm (RCGA) using the chromosomes with the real number.

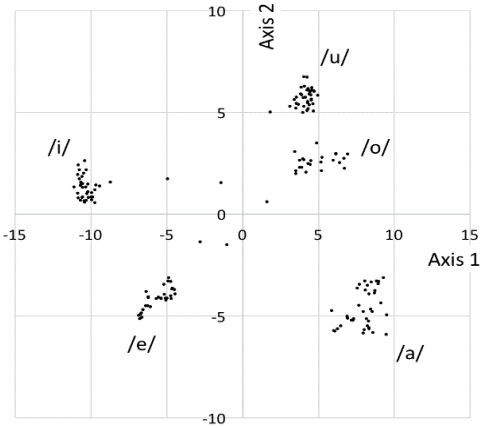

Figure 1. Analytic result of isolated vowels by the vowel space parameter

III. MODEL TO IMITATE THE SPEECH PRODUCTION PROCESS

Speech production and perception processes of this model involve four steps for each: linguistic, psychological, physiological and physical. These processes are similar to the usual speech chain, but a difference exists in the psychological process of this model. The psychological process represents one of the function of speech processing done in the human brain and is performed unconsciously. We use the vowel space parameter as the speech parameter at the psychological process. The processing of the coarticulation is performed at this process for both production and perception. We adopt the Fujisaki model as the production model for the coarticulation and the GA as the training algorithm to eliminate (or normalize) the coarticulation at the perception.

A. Acquisition of source control

1) Imitation of voiced/unvoiced decision using the GA

Figure 2 shows the coding method for voiced/unvoiced decision. The voiced or unvoiced decision is coded 1, 0, respectively. The starting time t_i is coded in the real number. First, word length is divided by number of phonemes to get equal length parts of a word. Each point with the same interval is set as an initial start point of the algorithm, then the times measured from the initial staring points are set as a starting time [×10ms] for each phoneme.

GA operation of the voiced/unvoiced decision and the fitness are as follows.

Crossover: For voiced/unvoiced decision, the code values (1 or 0) are exchanged at the crossover point. For the starting times, pairs of corresponding genes of two

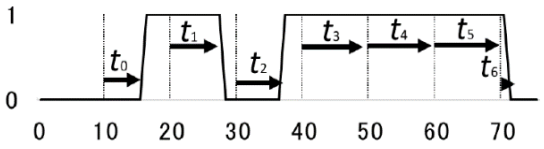

Figure 2.Coding method of voiced/unvoiced decision.

individuals are selected randomly. A new gene is generated as a random number between values of two genes. A new individual is generated by changing the gene of one of the individuals to the new gene. The other individual is similarly treated with other individual. Nine children are generated and the elite individual is added. The size of a generation was set at ten.

Mutation: For the voiced/unvoiced decision, a gene is selected with probability 0.6 for each individual and the value 1 or 0 is inversed. For the starting time, a gene is selected with a probability 0.6 for each individual and changed to be a number generated randomly between "value of the gene -20% of interval among the initial staring points" and "value of the gene +20% of interval among the initial staring points". The probability 0.6 was selected by preliminary tests.

Fitness is the percentage of the agreement of voiced/unvoiced decision for each frame.

2) Imitation of fundamental frequency using the GA

Figure 3 shows the coding method of the fundamental frequency. We adopted one phrase and one accent for the Fujisaki model. Expression of the Fujisaki model becomes as follows [7].

$$\ln F_0 = \ln F_{\min} + A_p G_p(t - T_0) + A_a\{G_a(t - T_1) - G_a(t - T_2)\} \quad (1)$$

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

$$G_a = \begin{cases} 1 - (1 + \beta t)\exp(-\beta t) & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (3)$$

Where F_0, $F_{\min}$, A_p, G_p, A_a, G_a, T_0, T_1, T_2, α, and β are fundamental frequency, minimal fundamental frequency, amplitude of phrase command, phrase component function, amplitude of accent command, accent component function, time point of beginning of phrase, time point of beginning of accent, time point of end of accent, time constant of phrase command function, time constant of accent command,

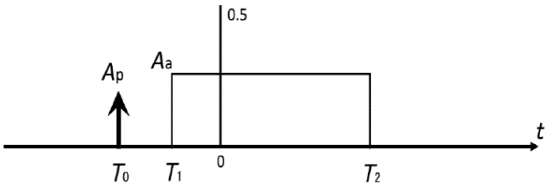

Figure 3. Coding method of fundamental frequency.

respectively. According to the preliminary test, we set $\alpha = 10$, $\beta = 25$, $A_p = 0.08$. Therefore, in order to imitate the fundamental frequency, the unknown parameters to be detected were time at phrase command T_0 and amplitude A_a, as well as beginning point T_1 and ending point T_2 of the accent command, which were coded in real numbers and were different for each word in their acquisition. The other parameters were decided using preliminary tests and set constant. The times are restricted in the conditions:

$$-20 < T_0 < 0 \quad (4)$$

$$T_0 < T_1 < T_2 < T_E \quad (5)$$

Where T_E is word length. The amplitude of accent command is

$$0 < A_a < 1. \quad (6)$$

The GA operations and the fitness of the fundamental frequency are as follows.

Crossover is the same as the above crossover of the staring time in III.*A.1 Imitation of voiced/unvoiced decision using the GA.*

Mutation is the same as the above mutation of the staring time in III.*A.1) Imitation of voiced/unvoiced decision using the GA.*

Fitness is Euclidean distance between a fundamental frequency pattern of an original word and the generated contour of the Fujisaki model, which is evaluated at the voiced interval.

3) Evaluation of the imitation model of source control

The information for source is obtained by combining information of the fundamental frequency and the voiced/unvoiced decision obtained by the GA.

In order to evaluate the obtained information of the source, we performed listening tests to check the quality of the synthesized speech made by the obtained parameters [12]. As a result of the tests, it was shown that the quality (3.2) attained was near to that of the analysis synthesis speech (3.6) at the third generation. It is very similar to human infant's linguistic performance that the model can acquire the source information in a short time with a few trials.

B. Imitation of articulatory position control [13]

The articulatory position is acquired by imitating time pattern of each dimension of the vowel space parameter extracted from speech spectrum. According to the former method applying the Fujisaki model to Formant frequencies [5], a few accent command components of the Fujisaki model are prepared to be the same number as phonemes of a word. In this expression, the ending times of phonemes are not used but the beginning times only. Therefore, the expression of the Fujisaki model for each dimension of the vowel space parameter is as follows.

$$A(t) = A_{\min} + \sum_{j=0} (A_{j+1} - A_j) G_{aj}(t - T_j) \quad (7)$$

$$G_{aj}(t) = \begin{cases} 1 - (1 + \beta t)\exp(-\beta t) & (t \geq 0) \quad (8) \\ 0 & (t < 0) \quad (9) \end{cases}$$

The time constant β was set to 20 from the preliminary test, referring to that of fundamental frequency. The unknown parameters to be detected to imitate the vowel space

parameters are A_j and T_j of the Fujisaki model, which represent characteristics of each phoneme of a word. The parameters are prepared for each phoneme and coded in a real number. The other parameters were decided using preliminary tests and set constant. The initial staring points are set similarly to the above staring time in III.*A.1) Imitation of voiced/unvoiced decision using the GA.*

The GA operations and the fitness of the vowel space parameters, which represent information of the coarticulation, are as follows.

Crossover: We used the all point crossover. Pair of corresponding genes of two individuals are selected. For each gene, a new gene is generated as a random number between values of the two genes. A new individual is generated by changing all genes of one of the individuals to the new genes. The other individual is similarly treated with other individual. Nine children are generated and the elite individual is added. The size of a generation becomes ten.

Mutation: One of the genes is selected with mutation probability 0.6 for each individual. The value 0.6 was decided by preliminary tests. The amplitude A_i is changed to be a number generated randomly between "the gene value - its 50%" and "the gene value + its 50%". The staring time T_j is changed to be the number generated randomly between "the gene value - 20% of interval among the initial staring points" and "the gene value + 20% of interval among the initial staring points".

Fitness is the Euclidian distance between the vowel space parameter obtained from the original spectrum of a word and the Fujisaki contour generated by the GA.

In order to evaluate the acquired vowel space parameters, we synthesized speech using the spectrum obtained from the imitated and acquired vowel space parameters and performed listening tests to check the quality of the synthesized speech [13]. As a result of the tests, it was shown that at the third generation, words were acquired with the quality of correct rate 90% in the listening test for vowels. It is very similar to human infant's linguistic performance that the model can acquire the articulatory position control with a few trials.

IV. ANALYSIS OF THE COARTICULATION

For spectra of a word, it was shown that the phonemic feature can be extracted as the command of the Fujisaki model from the vowel space parameter with a very ambiguous phonemic feature because of the coarticulation [13]. In this section, we discuss in more detail using listening tests and analysis by the proposed model, in which the command of the Fujisaki model represents the effect of inversely estimating (normalizing) the process of coarticulation.

Typical objects of study regarding the coarticulation are three Symmetrically Connected Consecutive Vowels [14][6][15] (3SCCV) and two Consecutive Vowels (2CV). The 3SCCV is the connected three vowels in which the beginning and the ending vowels are the same and has been used very much to analyze and model the coarticulation.

In the 2CV, there exists the phenomenon [6] that the phoneme at transitional part is not heard. For example, the transitional part of the 2CV /ai/ is acoustically [e]. Where /.../ shows that "..." are the phonemic representation. But usually humans cannot hear this transitional sound. This is also the effect of the normalization of the coarticulation.

We newly think that the following phenomena are also the effect of the normalization of the coarticulation. First, we cut the 2CV at the center of the transitional part and delete the later part. We call this sound as the Cut two Consecutive Vowels (C2CV). When we listen to the C2CV, not only the first vowel but also the second vowel is shortly heard. We think that this phenomenon shows the predicted sound which is heard because information of the second vowel exists at the former and the transitional part as the effect of the coarticulation.

A. *Three symmetrically connected consecutive vowels*

We can clearly discriminate phonemic feature of a center vowel of the 3SCCV whereas acoustic characteristics at the center do not reach the target of an isolated vowel. We think this phenomenon to be the effect of the normalization of the coarticulation in the process of human speech information processing. For example, speech /i/ in /aia/ is sometimes acoustically [e], but human hear it like [i].

We prepared the 3SCCV of /aia/, /aua/, /iui/. Figure 4 shows an example of analyzed result of the /aia/. This is the value at the third generation where the GA is performed 100 times and an individual with the best fitness at the tenth generation is selected. We set the number of phonemes to be five including the preceding and the following silence. The vertical axis shows the first dimension of the vowel space parameter. The curved solid line shows the observed vowel space parameter and the dashed line is the contour of the Fujisaki model. The straight line is the command of the Fujisaki model. The value of the command at the first /a/ is about 9, that of /i/ is about -10 and that of the last /a/ is about 8. Comparing these values to the horizontal axis of Figure 1, we can see that the value of the /i/ is grasped as the Fujisaki's command at the place acoustically [e] where the observed value of the solid line is about -7. That is to say, the model grasps the target of the phoneme while the acoustical value does not reach the target. For other dimensions of the vowel space parameter and speech data, we got almost the same results.

Figure 4 can be understood as follows. The proposed coarticulation model is already trained coefficients of the

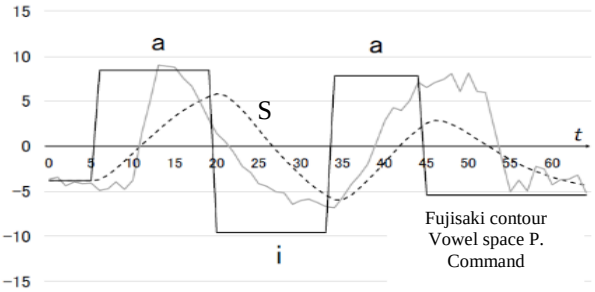

Figure 4. An example of analyzed result of three symmetrical connected consecutive vowels /aia/ by the proposed method.

expression of the Fujisaki model including the time constants. As the result, the model predicts (inversely estimates) the value of the target as the command of the Fujisaki model using the slope from the preceding phoneme to the target (S in Figure 4) under the condition of the set number of phonemes.

B. Cut two consecutive vowels

The second vowel of the C2CV doesn't exist physically but is heard auditorily. For example, when we delete the following part of /ai/ and listen, it sounds like [aj], which is usually perceived as [ai]. This phenomenon is thought to show the effect of unconscious normalization processing of the coarticulation. We study whether the proposed model of acquisition of spectral information can be a model for such a phenomenon of normalization of the coarticulation. For investigation of the C2CV, we adopted /ai/, /ia/, /au/, /ua/, /iu/, /ui/. They are the transitional phonemes among /a, i, u/ which exist in almost all languages in the world.

First, we performed listening tests to check what the inexistent vowel of the C2CVs sounds like. We used the speech data uttered three times by a Japanese male. We cut the data at 3/4 point in the transitional interval by observing the transitional contour of Formants and processed the tail by the linear fadeout with 5ms length. The beginning of the fadeout is the center of the two vowels.

All the C2CVs were listened randomly and the listeners were asked to answer in two second. The instruction to the listeners was "listen to a consecutive two vowels and answer the last vowel among /a, i, u, e, o/". The listeners were five Japanese university students with normal hearing ability. One speech specimen was used seven times and the last four results were used for the statistical analysis. Therefore, there were 60 answers for each C2CV.

Table I shows a confusion matrix of the result of the listening test. The numbers represent percentages of the perceived phonemic value, shown in the shaded block, of the inexistent vowels, shown in italic at IN for each C2CV. When we see the shaded part in the table, we find that the target vowels of C2CVs are answered at a high rate.

Figure 5 shows an example of the result of the analyzed C2CV /ai/. The vertical axis represents the first dimension of the vowel space parameter. We hypothesize to be four phonemes: silence, C2CV, silence. This is a result at the third generation which is selected as the individual with the best fitness at the tenth generation after the GA operations were repeated 100 times. We can see that the values of the command at [a] is about 9 and at [i] is about -9. Comparing to the horizontal axis of Figure 1, we can see that not only /a/ but also the value of the physically inexistent sound [i] is grasped as the Fujisaki's command. Our understanding for the mechanism of this prediction is the same as the description at the above section.

V. CONCLUSION

We proposed an infant model, which imitates and acquires spoken words. For the imitation of the source information, the fundamental frequency pattern was expressed in the Fujisaki model. For the imitation of the articulatory gesture,

TABLE I. A CONFUSION MATRIX OF THE LISTENING TEST [%].

OUT / IN	i	e	a	o	u
a_*i*	63	38	0	0	0
u_*i*	51	4	3	8	33
i_*a*	0	18	82	0	0
u_*a*	0	0	99	1	0
i_*u*	0	0	0	0	100
a_*u*	0	0	0	3	97

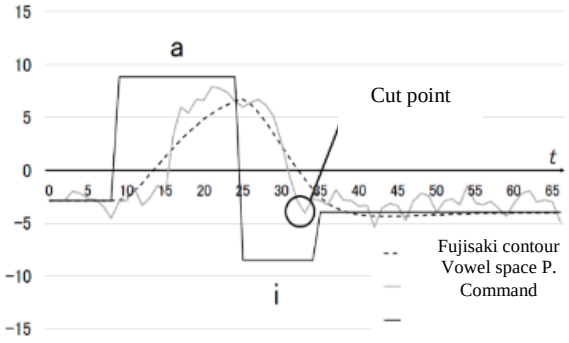

Figure 5. An example of analyzed result of the cut two consecutive vowels /ai/.

spectral information was represented in the vowel space parameter. Trial and error evident in infant's imitation process was modeled in the genetic algorithm.

We confirmed similar performance to human infants in the model of imitating spoken words, which could acquire spoken words with reasonable quality in a few trials. We showed in the analysis of the coarticulation that this model expresses suitably the normalization (inverse estimation) of the coarticulation.

This model shows that a new speech analysis method which predicts a spectrum or inversely estimates the coarticulation can be possible, especially predicts a following phoneme from the proceeding phoneme. The model also shows that a new method is possible to make a very ambiguous phonemic characteristic of spectrum in continuous speech clearer.

REFERENCES

[1] GP. K. Kuhle, B. T. Conboy, S. C. D., Padden, M. Rivera Gaxiola, and T. Nelson: "Phonetic learning as a pathway to language: new data and native language Magnet theory expanded (NLM-e)," Philos. Trans. R. Soc. B, 363, pp.979 – 1000, 2008.

[2] T. Takara, S. Yamashiro, and T. Ooishi, "Study on speech synthesis using principal components on spectral space of vowels," 2015 Annual Meeting of Acoustical Society of Japan, 2-Q-34, pp. 379-380, 2015.

[3] Tomio Takara, Akichika Higa, Syouki Kaneshiro, and Yuuya Oshiro: "Speech analysis-synthesis system using principal components of vowel spectra," The Journal of the Acoustical Society of America 140, p. 2962, 2016

[4] S. Hiroya, "Brain science of 'speaking and listening to speech': The relationship between speech production and

perception in brain," J. Acoust. Soc. Jpn, vol. 73, no.8, pp. 509-516, 2017.

[5] Y. Saito and H. Fujisaki, "Formulation of the process of coaticulation in terms of formant frequencies and its application to automatic speech recognition," J. Acoust. Soc. Jpn., vol. 34, pp. 177-185, 1978.

[6] M. Akagi and S. Furui, "Modeling of vowel target pre-diction mechanism in speech perception," Proc. of IECE, vol. 69A, pp. 1277-1285, 1986.

[7] H. Fujisaki and K. Hirose: "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn. (E) vol. 5, no. 4, pp.233 – 242, 1984.

[8] Z. Michalewicz, Genetic Algorithm + Data Structures = Evolution, Programs, 3rd ed., Springer-Verlag, Berlin, Heidelberg, pp. 97-106, 1996.

[9] S. Imai and Y. abe, "Extraction of spectral envelope using improved cepstral method," Proc. IEICE A, J62-A, 4, pp. 217-223, 1979.

[10] S. Imai, "Log magnitude approximation (LMA) filter," Proc. IEICE A, J63-A, 12, pp. 886-893, 1980.

[11] R. Schalkoff, Pattern recognition, pp. 306-307, John Wiley & Sons, Inc., 1992.

[12] T. Takara and R. Eto, "Source information for the acquisition model of spoken word using genetic algorithm and Fujisaki model," Annual meeting of ASJ autumn, 1-8-5, pp. 183 - 184, 2017.

[13] T. Takara, A. Higa, R. Eto, and Go Ishikawa, "Generative model of spectra for a word using Fujisaki model and genetic algorithm," J. Acoust. Soc. Jpn., vol. 39, no. 2, pp.147 – 149, 2018.

[14] N. Kuwahara and H. Sakai, "Normalization of coarticulation effect for a sequence of vowels in connected speech," J. Acoust. Soc. Jpn., vol. 29, no. 2, pp. 91-99, 1973.

[15] Tomio Takara and Motonori Tamaki: "A normalization of coarticulation of connected vowels using neural network," ICSLP 90 pp.1369-1372, 1990.

 ISBN: 978-1-61208-768-9