2011 International Conference on Communication Engineering and Networks
IPCSIT vol.19 (2011) © (2011) IACSIT Press, Singapore

Study on the Artificial Synthesis of Human Voice using Radial Basis Function Networks.

Yuki Naniwa1, Takaaki Kondo[1], Kyohei Kamiyama[1] and Hiroyuki Kamata[2]

[1]Graduate School of Science and Technology, Meiji University.Kawasaki, Japan
[2]School of Science and Technology,Meiji University.Kawasaki, Japan
[1] ce01067@meiji.ac.jp [2] kamata@isc.meiji.ac.jp

Abstract— In this study, we introduce the method of reconstructing more natural synthetic voice by using radial basis function network (RBF) that is one of neural network that is suitable for function approximation problems and following and synthesizing vocal fluctuations. In the synthetic simulation of RBF, we have set the Gaussian function based on parameters and tried to reconstruct the vocal fluctuations. With respect to parameter estimation, we have adopted to nonlinear least-squares method for making much account of the nonlinearity of human voice. When we have reproduced the synthesized speech, we have tried to reconstruct the nonlinear fluctuations of amplitude by adding normal random number. We have made a comparison the real voice and the synthetic voice obtained from simulation. As a consequence, we have found that it was possible to synthesize the vocal fluctuations for a short time. .

Keywords-component; RBF network, K-means, Nonlinear least square method, Fluctuation of vocal cord.

1. Introduction

In recent year, there are many methods centered upon text−to-speech synthesis because of the performance of computers that handle large amount of speech database being improved. In the study of text−to-speech synthesis, especially in recent, hidden Markov models (HMM) has often been used in some way[1].These studies have human voice as a recording medium. So, when we set the rule of connecting each syllable, phoneme, waveform of 1 cycle and that of controlling prosodic information such as pitch and amplitude precisely, we able to reproduce portable synthetic voice and it is also commercialized.

However, in terms of humanity about generated voice, it is still not quite express. Therefore, the technique of speech production for the machine model has become a hot topic. Sawada has developed a robot that captured the appropriate speech techniques by associating not only leaning of the correspondence between acoustic features and vocal tract shape, but also leaning of the amount of motor control [2]. In regard to speech synthesis on software, Kanda has constructed and validated the vowel acquisition model by segmenting continuous acoustic signal and articulatory movement based on the vocal tract production model by Maeda and the neural model [3, 4]. The following thing has been mentioned in various papers, and we think that it is extremely meaningful in the field of speech recognition and speech synthesis that we make clear language acquisition board as a basic flow of the process of language acquisition from infant[5]. In addition, we believe it is necessary to try to synthesize self-organized speech like with vowel acquisition through voice imitative interaction in speech production.

In this study, we try to create the new artificial voice but in a short time, not voice recording medium by focusing on the human vocal cords. By using the radial basis function neural network (for pattern classification and function approximation problem), we aim to synthesize vocal fluctuations and to produce artificial sounds that are similar to the real voice by keeping track of real voice as much as possible.

Therefore, we will cope with the following three points.

(1) The reconstruction of the fundamental frequency with chaos in short time.
(2) The verification of the synthesized fluctuations improved by adding the vocal fluctuations to the smoothed data in the input.
(3) The reconstruction of the intensity of each period by Box-Muller-method.
In keeping with above points, we attempt to synthesize the nonlinear vocal fluctuations by using RBF network with the nonlinear approximation of parameters.

2. Fluctiation of Vocal Cords

At first sight, the human speech waveform seems to have a periodic structure. However, the shape of the waveform is changing for each cycle and the cycle width subtle changes accordingly[7]. The difference in width of the period is called the fluctuation of the vocal cords.

2.1. Sampling of the vocal fluctuations

With regard to the method of sampling vocal fluctuations, we determine the initial value $x_{max}(k)$(the number of sample $t_0 = 0$) arbitrarily from the maximum (minimum) value of voice waveform, and starting from there, number of sample until the next peak is 1 component of vocal fluctuations. In the case of k fluctuations, the following formula is given.

$$y(k) = x_{max}(k) \tag{1}$$

Visualizing t_k as the total number of samples, we seek in term of the width of each period $x_{max}(k)$.

$$x_{max}(k) = t_{k+1} - t_k \tag{2}$$

We collect the components of vocal fluctuations for the simulation by repeating this procedure. In addition, we set up a threshold to make the exact maximum (minimum) value and are to be used to normalize each of it.

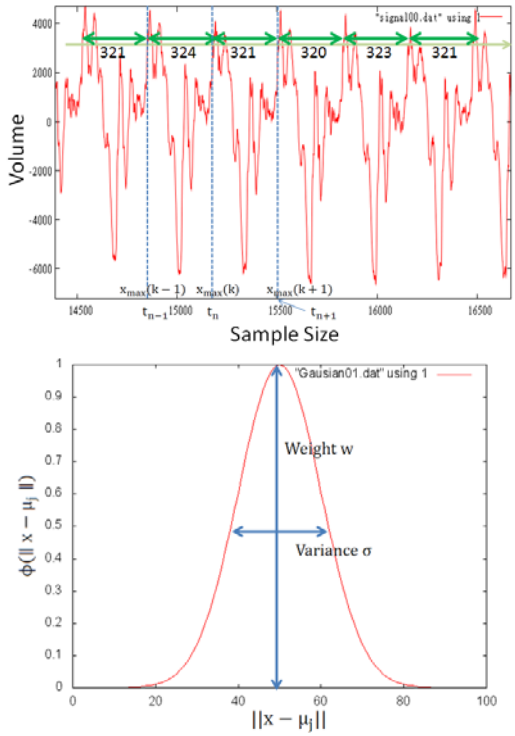

Figure 1. The extraction method of vocal fluctuations and Gaussian function of RBF

3. RBF Network Including the Smoothing Components

RBF(Radial Basis Function Network) is a kind of deformed network across the three-layers, and it was considered as the way to complement desired arbitrary function by overlaying the localized basis functions. With respect to basis functions, there are several functions about Simlate-spline, bell-shaped, but we use Gaussian function (see Fig.1) in this study.

3.1. Structure

The formula that represents the structure of RBF network that obtains 1-dimensional output f(x) to K-dimensional inputs is as follows. In this study, we will try to validate the multi-dimensional output too [8, 13].

$$f(k) = \sum_{i=0}^{M} \omega_i \varphi_i (r) \tag{3}$$

Here, $\varphi_i(r)$ is the basis function, ω_i is the coupling coefficient of it, M is the amount of it. Gaussian function used as RBF $\varphi_i(r)$ is shown below.

$$\varphi_i(r) = \exp\left(-\frac{\|x_k - \mu_i\|^2}{2\sigma_i^2}\right) \tag{4}$$

Here, x_k is the input vector and φ_i is the variance of basis function.

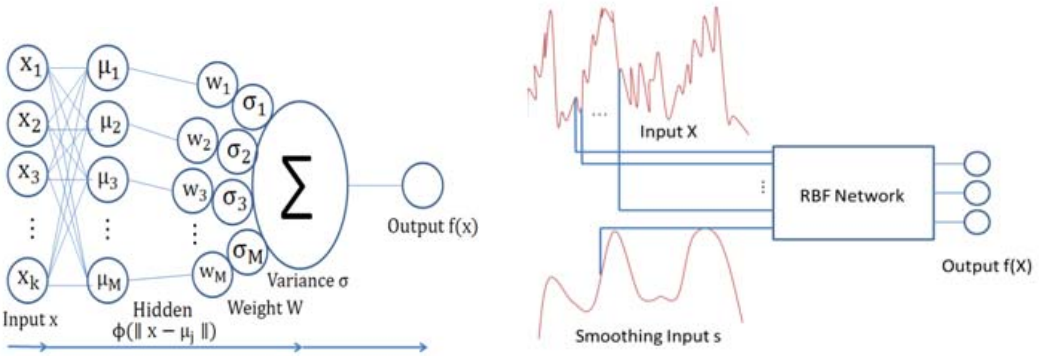

Figure 2. Structure of RBF and Input of RBF

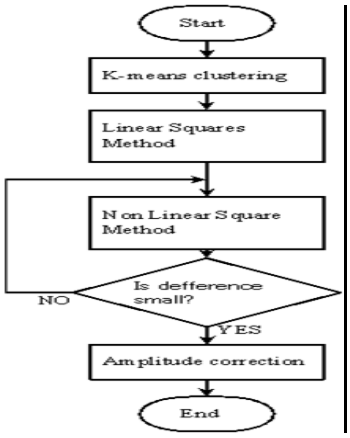

Figure 3. Flowchart of RBF network

μ_i is the i th center vector and determines the number of it according to the number of RBF. The interior of $\|\cdots\|$ also represents the Euclidean norm.

Depending on how to give this vector μ_i and the number of RBF M, the accuracy of predictive value varies significantly. In this study, we have focused on the variance σ_i included in Gaussian function. In most of the studies using RBF, they always have treated the variance σ_i as a constant value, but we treat it as a variable and considered the changing value according to the center vector μ_i to enhance the reproducibility.

The weight w_i is obtained by using least squares from M-input data and Eq.(3) from the general formula of RBF. In addition, utilizing the obtained weight w_i, we seek the variance σ_i by using the Gauss-Newton method from Eq.(4).

We aim to reconstruct the predictive value closer to the actual data from two tasks of this. Fig. 2 shows the structure of RBF that obtains one to several outputs against K-inputs, and Fig. 3 summarizes the flow of RBF.

3.2. K-means method

(1) We give appropriate initial values of M from the actual data of the vocal fluctuations.

(2) We calculate the distance between the center of the cluster and all the data, and classify the cluster most similar to the Euclidean distance.
(3) We seek the center from the formed cluster.
(4) We repeat the step of *(2), (3)* until there is no change in the center of the cluster.

For simplicity, we put the image of the algorithm of the K-means method by using the scatter plot on two-dimensional planes.

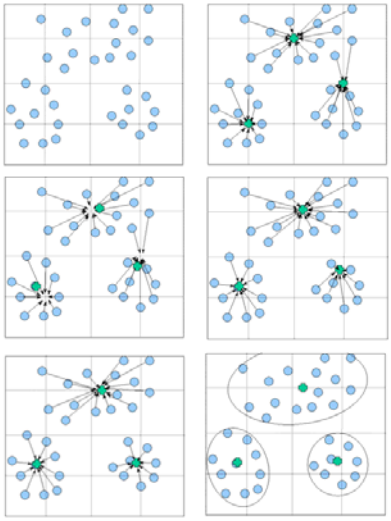

Figure 4. K-means clustering

3.3. Nonlinear least- squares method

In seeking the variance σ_i included in Gaussian function of the nonlinear equation$p(x) = \sum_{i=0}^{M} \omega_i \varphi_i(r)$, we use the Gauss-Newton method of being least squares approximation of Newton's method that we approximate the curve of function p(x) by means of the tangent by as a way to approximate more accurate results by little computation.

1) Gauss-Newton method

We use Gauss-Newton method in M-dimensional input to approximate the variance σ_i included in RBF [12].

In determining the number of M-parameters, given the n data $(x_i, y_i)(i = 1, 2, \cdots, n)$, sum of squared errors of real data y_i and nonlinear equation $p(x_i, a_0, a_1, \cdots, a_M)$is shown as below.

$$f(a_0, a_1, \cdots, a_M) = \sum_{i=1}^{n} (y_i - p(x_i, a_0, a_1 \cdots, a_M))^2 \tag{5}$$

In this equation, we determine the value of the parameter.

The parameters a_M to minimize $f(a_0, a_1, \cdots, a_M)$ is desirable to fulfill the following equation.

$$\frac{\partial f}{\partial a_0}, \frac{\partial f}{\partial a_1}, \cdots, \frac{\partial f}{\partial a_M} = 0 \tag{6}$$

Taylor expansion for a_0 of $q = f(a)$ can be expressed as follow.

$$q_i = f(a_i) + \frac{\partial f(a_i)}{\partial a_0} da_0 + \frac{\partial f(a_i)}{\partial a_1} da_1 + \cdots \tag{7}$$

Upon arranging it as the Taylor expansion of M-dimensions,

$$q_n = \sum_{i=1}^{n} [f(x_i, a_0, \cdots, a_M) + \frac{\partial p(a_0, \cdots, a_M)}{\partial a_0} da_0 + \frac{\partial p(a_0, \cdots, a_M)}{\partial a_1} da_1 + \cdots + \frac{\partial p(a_0, \cdots, a_M)}{\partial a_1} da_M]^2 \tag{8}$$

can be expressed as above. Differentiating this function $a_i(k = 0, 1, \cdots, M)$,

$$\frac{\partial q_n}{\partial(da_k)} = 2\sum_{i=1}^{n}[f(x_i, a_0, \cdots, a_M) + \frac{\partial p(a_0,\cdots,a_M)}{\partial a_0} da_0 + \frac{\partial p(a_0,\cdots,a_M)}{\partial a_1} da_1 + \cdots + \frac{\partial p(a_0,\cdots,a_M)}{\partial a_1} da_M] \frac{\partial p(a_0,\cdots,a_M)}{\partial a_1} = 0 \quad (9)$$

can be expressed as above.This time, q_i takes a minimum value. Expressed as a matrix after transposing the compo-nent, $f(x, a_0, a_1, \cdots, a_M)$,

$$\begin{bmatrix} \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_0}\right)^2 & \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_0}\frac{\partial p_i}{\partial a_0}\right) & \cdots & \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_0}\frac{\partial p_i}{\partial a_M}\right) \\ \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_1}\frac{\partial p_i}{\partial a_0}\right) & \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_1}\right)^2 & \cdots & \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_1}\frac{\partial p_i}{\partial a_M}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_M}\frac{\partial p_i}{\partial a_0}\right) & \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_M}\frac{\partial p_i}{\partial a_1}\right) & \cdots & \sum_{i=0}^{n}\left(\frac{\partial p_i}{\partial a_M}\right)^2 \end{bmatrix}$$

$$\begin{bmatrix} \delta a_1 \\ \delta a_2 \\ \vdots \\ \delta a_M \end{bmatrix} = \begin{bmatrix} -\sum_{i=0}^{n}\frac{\partial p_i}{\partial a_0} R_i \\ -\sum_{i=0}^{n}\frac{\partial p_i}{\partial a_1} R_i \\ \vdots \\ -\sum_{i=0}^{n}\frac{\partial p_i}{\partial a_0} R_i \end{bmatrix} \quad (10)$$

can be represented as above. By solving the matrix equation above, it is necessary to calculate the following equation iteratively to determine the true value a_i obtained from the parameters $\delta a_1, \delta a_2, \cdots, \delta a_1$.

$$a_i = a_i^{(0)} - \delta a_i \quad (11)$$

Here, $a_i^{(0)}$ is the initial value set arbitrarily. When the convergence condition is satisfied, we will terminate the iteration.

$$\sum_{i=0}^{n}(f(a_0, a_1, \cdots, a_M))^2 < \delta^2 \quad (12)$$

At this time, the obtained parameters are the true values.

4. Amplitude Correction of Synthesized Speech Signal

It is obvious that human voice contains a variety of nonlinear elements in addition to vocal fluctuation. Then, now we focus on this nonlinear change in amplitude, and try to vary amplitude for each cycle with random number. For random number generation, we collect the difference between the maximum and minimum of real voice signals (the width of amplitude), and we generate the random number conforming to Gaussian distribution by Box Muller method because distribution of the data relatively close to it [6]. Then, we use the random number as a multiplication of each period of real speech signal.

4.1. Box Muller method

Random numbers with normal distribution (mean μ, variance σ^2) η_1 and η_2 can be expressed as follow using independent uniform random number ξ_1, ξ_2 in [0, 1].

$$\eta_1 = \mu + (\sqrt{-2\log(\xi_1)}\cos 2\pi\xi_2)\sigma \quad (13)$$

$$\eta_2 = \mu + (\sqrt{-2\log(\xi_1)}\sin 2\pi\xi_2)\sigma \quad (14)$$

5. Synthesis of the Vocal Fluctuations

5.1. Simulation conditions

The measurement condition of voice signal used in the simulation is shown in Table 1 bellow.

TABLE I. MEASUREMENT CONDITIONS

The subject	Japanese adult man (22)
Measurement signal	A vowel /a/
Sampling frequency	44.1[kHz]
Bit rete	16[bits]
The number of samples (voice signal)	About 50000[points]
The number of the vocal fluctuations (period)	154[periods]

6. Synthesis Results

In Fig. 4, against the original waveform of vocal fluctuations of /a/, we show a comparison of the synthetic vocal fluctuations in the pattern of the proposal method that we have added the one to several smoothed data in the input. In the Gauss-Newton method used in the approximation of variances σ_i, it is considerably variable whether the initial values converge or not depending on the value. So we compared it to second decimal places strictly.

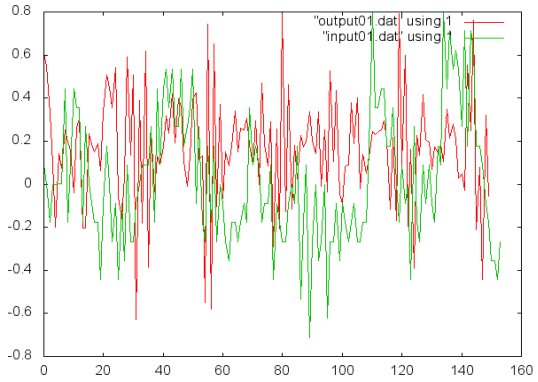

Figure 5. Conparison of the real vocal fluctuations /a/(input00.dat: Real vocal fluctuations),(outpu01.dat: Synthesized vocal fluctuations with smoothed data)

In the conventional method, with respect to the reconstruction of high parts of fluctuations, there was something lacking fidelity [14]. However, the proposal method in Fig. 5 shows that it is relatively better. Also, we succeeded in shortening the error between real vocal fluctuations and synthesized vocal fluctuations than the conventional method. We have been able to get the best results since we adopted the nonlinear approximation. However, there was no significant improvement about the high frecuency components. To workaround this, we think that it is necessary to increase the degree of smoothed fluctuations in input or to apply LPF (low-pass filter) to the output.

After that, we compare between the real speech waveform and the synthesized speech waveform that we have faithfully reconstructed based on the synthesized vocal fluctuations in Fig. 7.

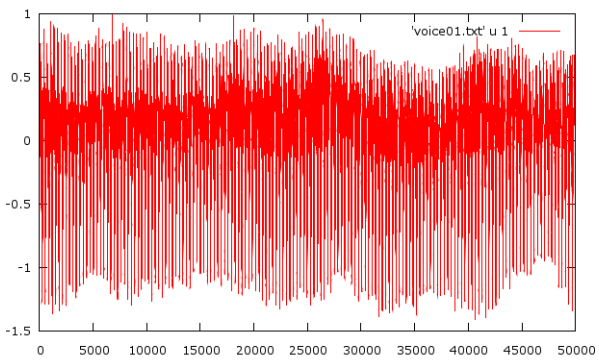

Figure 6. Real voice waveforms /a/

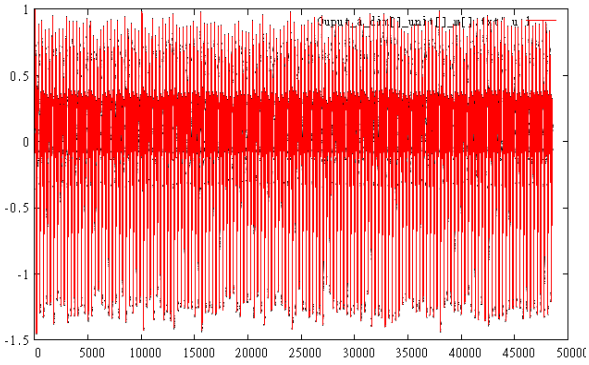

Figure 7. Synthetic voice waveforms /a/

About the wave of details, it is almost identical to the real speech waveform. We have also succeeded in improving the randomness in fluctuations of amplitude by expanding the variance of normal random variable. However, it hasn't become smooth like real speech waveform. So we have tried to reconstruct by adding the ratio of the maximum (minimum) value of the real speech signal and the synthesized speech signal for each period.

Next, we examined the behavior in frequency domain. Relating to the verification, it is considered to be 300Hz~3.4 kHz frequency of human voice, and we made FFT in the range of 0~4096Hz. In comparison with Fig. 8, the frequency spectrum of real speech signal, mainly in the range of 0~500 Hz, generally concentrate at a certain frequency. On the other hand, that of the synthesized speech waveform has some slight dispersion in Fig.9.

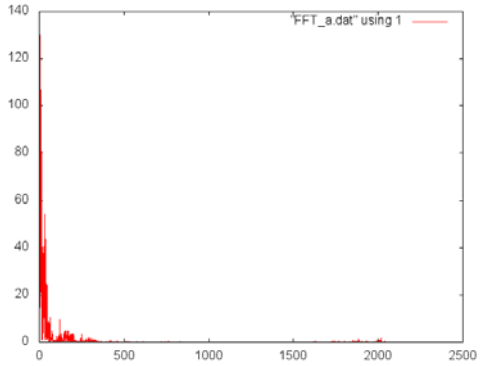

Figure 8. Frequency spectral density of real voice /a/ [2sec]

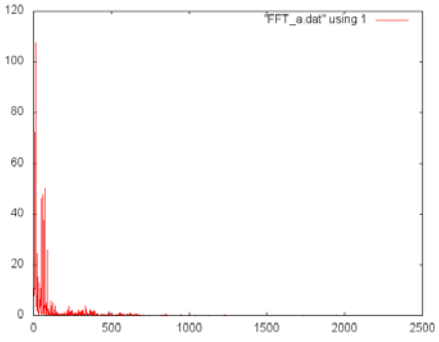

Figure 9. Frequency spectral density of Synthetic voice /a/ [2sec]

Then, in the field of time series analysis, human voice and the fluctuations of vocal cords have been known to be included in the chaotic[7,9,10]. Therefore, we have also decided to estimate the value of the maximum Lyapunov exponents. Regarding estimation, we have carried out the embedding dimension estimation by using the FNN (False Nearest Neighbor) method. This method enabled to reconstruct an attractor with objectivity and accuracy. And then, we have carried out the time-delay estimation based on the mutual information. This will seek Lyapunov exponents for a reconstructed attractor. As a result, in the largest

Lyapunov exponents, there are one or more positive values in both that of the synthesized speech and of the synthesized vocal fluctuations. Therefore, it was found that they have chaos.

TABLE II. LYAPUNOV SPECTRUM OF VOICE SIGNAL

	Vowel of /a/ (Embbeding dimention, Delaytime)	Lyapunov spectrum
Voice /a/	(3 , 5)	$\lambda_1 = -0.333104$ $\lambda_2 = -0.259747$ $\lambda_3 = 0.28725$
Synthetic Voice /a/	(3 , 4)	$\lambda_1 = -0.157536$ $\lambda_2 = -0.050309$ $\lambda_3 = 0.31957$
Synthetic Voice with smoothed data /a/	(5 , 5)	$\lambda_1 = -0.630296$ $\lambda_2 = -0.041686$ $\lambda_3 = 0.001254$ $\lambda_4 = 0.020488$ $\lambda_5 = 0.669256$

TABLE III. LYAPUNOV SPECTRUM OF VOCAL FLUCTUATIONS

	Vowel of /a/ (Embbeding dimention, Delaytime)	Lyapunov spectrum
Voice /a/	(2 , 2)	$\lambda_1 = -0.172612$ $\lambda_2 = 0.28751$
Synthetic Voice /a/	(2 , 2)	$\lambda_1 = -0.657639$ $\lambda_2 = 1.11770$
Synthetic Voice with smoothed data /a/	(2 , 5)	$\lambda_1 = -0.729565$ $\lambda_2 = -1.62867$

7. Conclusion

In this simulation, we have focused on vocal fluctuations of real speech waveform /a/ to reconstruct more natural synthetic voice. To reconstruct these nonlinear vocal fluctuations, by RBF neural network, and by adding the smoothed vocal fluctuations to input, we tried to reproduce more natural synthesized speech. In synthesis of vocal fluctuations, we enable to reconstruct the synthesized vocal fluctuations in fewer dimensions than previous, and the measurement error is improved. In addition, regarding synthesized voice, it close to human voice by complementing voice data according to the synthesized vocal fluctuations and reconstructing fluctuations of amplitude. So we have succeeded in achieving the desired value from sound materials for shorter intervals. However, adding smoothed data of vocal fluctuations to the input didn't seem to directly link to ease the high frequency components.

Finally, throughout, in the field of speech synthesis, we were able to get a limited success by the parameter estimation with nonlinear least square method and improvement of input. But in terms of human voice, we believe that more information is needed to reconstruct the synthesized speech that was aware of the individual components [11]. Therefore, in the future, we put the multidimensional map of the real speech that includes a lot of more information as the materials, and we hope to efforts to improve the reproducibility by fitting it into the nonlinear function.

8. References

[1] Keiichi Tokuda, “FUNDAMENTALS OF SPEECH SYNTHESIS BASED ON HMM,” IEICE, vol.100, no.392, SP2000-74, pp.43--50, Oct. 2000.

[2] Mitsuki Kitani, Tatsuya Hara, Hideyuki Sawada, Autonomous Voice Acquisition of a Talking Robot Based on Topological Structure Learning by Applying Dual-SOM”, *TRANSACTIONS OF THE JAPAN SOCIETY OF MECHANICAL ENGINEERS Series C*, Vol. 77, No. 775 (2011), pp.1062-1070 .

[3] Hisashi Kanda, Tetsuya Ogata, Toru Takahashi, Kazunori Komatani, Hiroshi Okuno, “Simulation of Babbling and Vowel Acquisition based on Vocal Imitation Model using Recurrent Neural Network”, IPS 2009, 2-133"-"2-134", (2009-03-10).

[4] Eri Maeda, Takayuki Arai, Noriko Saika, “Study of mechanical models of the human vocal tract having nasal cavity”, IEICE 2003, 103(219), 1-5, (2003-07-17).

[5] Nobuaki Minematsu, Tazuko Nishimura, Kyoko Sakuraba, “Consideration on infants' speech mimicking and their language acquisition based on the structural representation of speech".
[6] E. R. Golder and J. G. Settle, “The Box-Muller Method for Generating Pseudo-Random Normal Deviates”, Journal of the Royal Statistical Society. Series C 1976, Page 19 of 12-20.
[7] Seiichiro DAI,Tohru HIROHKU,Norihito TOYOTA, “The Lyapunov Spectrum and the chaotic property in speech sounds,” IEICE technical report. Speech 99(576), 37-43, (2000-01-20).
[8] Satoshi OGAWA, Tohru IKEGUCHI, Takeshi MATOZAKI, Kazuyuki AIHARA, “Time Series Analysis using Radial Basis Function Networks,” IEICE technical report. Neurocomputing 95(505), 29-36, (1996-02-02).
[9] Takeshi SUZUKI, Masahiro NAKAGAWA, ``Fluctuashion of the vocal sound and its chaotic and fractal analyses," IEICE technical report. Nonliniea problems, 104(334), (2004-10-7).
[10] Hiroyuki Koga, Kunihiro Nakagawa, ``Chaotic Properties in Vocal Sound and Synthesis Model," IEICE technical report, NLP99-120, (1990-11) .
[11] Wang Xingyuan, Niu Yujin, ``Adaptove synchronization of chaotic systems with nonlinearity inputs,"
[12] H.O.Hartley,"The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares", American Statistical Association and American Society for Quality, Vol.3, No.2 pp.269-280 (1961)
[13] Tatsuo WATANABE,"Consideration of Prediction Accuracy of Chaos Time Series Prediction by RBFN ", The research reports of Oyama Technical College 39, 107-111, (2007-03-10).
[14] Yuki Naniwa, Takaaki Kondo, Kyohei Kamiyama, Hiroyuki Kamata, “The exact reproduction in the voice signal of radial basis function network,” IEICE technical report ,110(387), 199-204, (2011-01-24).