Evaluation of Fundamental Validity in Applying AR-HMM with Automatic Topology Generation to Pathology Voice Analysis

Akira Sasou [1]

[1] National Institute of Advanced Industrial Science and Technology, AIST, Japan

a-sasou@aist.go.jp

Abstract

Voice-pathology detection from a subject's voice is a promising technology for the pre-diagnosis of larynx diseases. Glottal source estimation in particular plays a very important role in voice-pathology analysis. To more accurately estimate the spectral envelope and glottal source of the pathology voice, we propose a method that can automatically generate the topology of the Glottal Source Hidden Markov Model (GS-HMM), as well as estimate the Auto-Regressive (AR)-HMM parameter by combining the AR-HMM parameter estimation method and the Minimum Description Length-based Successive State Splitting (MDL-SSS) algorithm. This paper evaluates the fundamental validity of pathology-voice analysis based on the proposed method. The experiment results confirmed the feasibility and fundamental validity of the proposed method.

Index Terms: pathology voice, glottal source, AR-HMM

10.21437/Interspeech.2013-28

1. Introduction

Voice-pathology detection from a subject's voice is a promising technology for the pre-diagnosis of larynx diseases and is also expected to enable organic growth located on the posterior part of the vocal folds to be identified [1]. So far, a number of researchers have devoted themselves to the automatic detection and classification of voice pathologies by means of acoustic analysis, parametric and non-parametric feature extraction, and so on. Glottal source estimation in particular plays a very important role in voice-pathology analysis. Many of the previous works seem to have adopted a frequency-domain analysis of the glottal source rather than a time-domain analysis like the parameterization-based method in the L-F model. This may due to the inaccurate glottal source estimation in the time-domain, and the frequency-domain-based methods may tend to be more robust to estimation errors than the time-domain-based methods [2].

To improve the estimation accuracy of the spectral envelope and glottal source from a voiced sound by the linear prediction method, we proposed the Glottal Source Hidden Markov Model (GS-HMM) and accompanying parameter-estimation method, in which the HMM was introduced as a non-stationary glottal source model [3,4]. Figure 1 depicts an example of the proposed source-filter model, which combines an Auto-Regressive (AR) filter and the GS-HMM. We refer to this model as AR-HMM. The glottal source can be obtained by inverse-filtering the speech signal. Fig.1 depicts the glottal source filtered by a high pass filter. The GS-HMM in the example has four states that are concatenated in a ring state so that the state transition occurs in order. Therefore, this GS-HMM can be adopted to analyze periodic voiced sounds. The state transitions of the four states were evaluated by the Viterbi algorithm from the glottal source signal. Each state has a single Gaussian distribution as an output Probability Distribution Function (PDF). The red and green lines in the graph were obtained by aligning the expectation and variance of each state's output PDF according to the state transition. Thus far, we have analyzed periodic voiced sounds based on this ring GS-HMM, where we adjust only the number of states in the GS-HMM and the prediction order of the AR filter with regard to the analyzed speech signal, while the topology remains concatenated in a ring state. In addition to the probability and statistical parameters of the GS-HMM, the topology automatically generated by the bottom-up approach according to a speech signal might also reflect much useful information related to speech production, voice pathology, the individual speaker's characteristics, and emotion/voice quality.

This paper proposes a method that can automatically generate the topology of the GS-HMM and estimate the AR-HMM parameter by combining the AR-HMM parameter estimation method [3,4] and the Minimum Description Length-based Successive State Splitting (MDL-SSS) algorithm [5]. We also evaluate the fundamental validity of the pathology-voice analysis based on the proposed method.

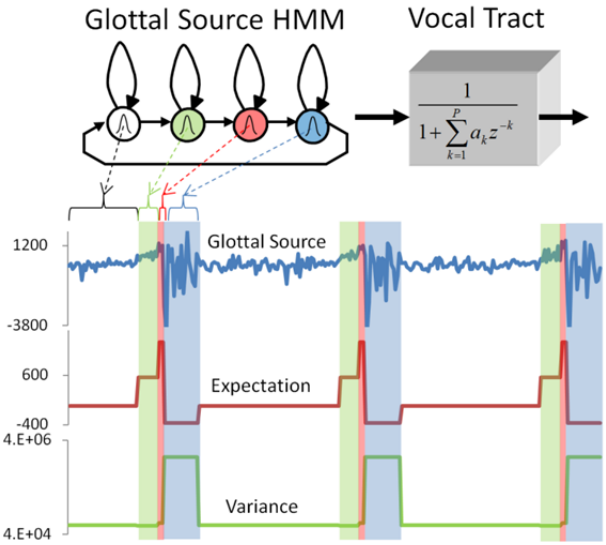

Figure 1: *Source filter model using glottal source HMM*

2. AR-HMM Parameter Estimation

The AR-HMM parameters consist of the AR coefficients and the parameters of the GS-HMM. Previously, we presented an algorithm [3] that iteratively estimates these parameters from a signal $x(t), t = 0, \cdots, T-1$. Let $\mathbf{a}^{(i)} = [a^{(i)}(1), \cdots, a^{(i)}(P)]^T$ represent the i-th estimate of the AR coefficients, where P denotes the prediction order of the AR filter. The *i*-th estimate of the glottal source signal $e^{(i)}(t), t = P, \cdots, T-1$ is given by

$$\mathbf{e}_P^{(i)} = \mathbf{x}_P - \mathbf{\Omega a}^{(i)} \quad (1)$$

where

$$\mathbf{e}_{\mathrm{P}}^{(\mathrm{i})} = \left[\mathrm{e}^{(\mathrm{i})}(\mathrm{P}), \mathrm{e}^{(\mathrm{i})}(\mathrm{P}+1), \cdots, \mathrm{e}^{(\mathrm{i})}(\mathrm{T}-1)\right]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{T}-\mathrm{P}}$$

$$\mathbf{x}_{\mathrm{t}} = [\mathrm{x}(\mathrm{t}), \mathrm{x}(\mathrm{t}+1), \cdots, \mathrm{x}(\mathrm{t}+\mathrm{T}-\mathrm{P}-1)]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{T}-\mathrm{P}}$$

$$\boldsymbol{\Omega} = \left[\mathbf{x}_{\mathrm{P}-1}, \mathbf{x}_{\mathrm{P}-1}, \cdots, \mathbf{x}_{0}\right] \in \mathbb{R}^{(\mathrm{T}-\mathrm{P})\times\mathrm{P}}.$$

We allocate a unique number from $\mathrm{S} = \{1, \cdots, \mathrm{N_s}\}$ to each state of the GS-HMM to distinguish it from other states, where $\mathrm{N_s}$ is the number of states. Here $\mu_{\mathrm{s}}^{(\mathrm{i})}, \sigma_{\mathrm{s}}^{2,(\mathrm{i})}, \mathrm{s} \in \mathrm{S}$ represent the i-th estimates of the population parameters of each state's output single Gaussian distribution. Given a state-transition sequence $\mathrm{s}(\mathrm{t}) \in \mathrm{S}, \mathrm{t} = \mathrm{P}, \cdots, \mathrm{T}-1$, the population parameters of a glottal source signal at time t are given by $\mathrm{m}^{(\mathrm{i})}(\mathrm{t}) = \mu_{\mathrm{s(t)}}^{(\mathrm{i})}, \mathrm{v}^{(\mathrm{i})} = \sigma_{\mathrm{s(t)}}^{2,(\mathrm{i})}$. Hence, the expectation vector of the glottal source signal is represented by

$$\mathbf{m}_{\mathrm{P}}^{(\mathrm{i})} = \left[\mathrm{m}^{(\mathrm{i})}(\mathrm{P}), \mathrm{m}^{(\mathrm{i})}(\mathrm{P}+1), \cdots, \mathrm{m}^{(\mathrm{i})}(\mathrm{T}-1)\right]^{\mathrm{T}}. \quad (2)$$

Based on the assumption that the samples of the glottal source signal at different instants are mutually independent, the covariance matrix of the glottal source signal is defined as a diagonal matrix given by:

$$\boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i})} = \mathrm{diag}\left\{\mathrm{v}^{(\mathrm{i})}(\mathrm{P}), \mathrm{v}^{(\mathrm{i})}(\mathrm{P}+1), \cdots, \mathrm{v}^{(\mathrm{i})}(\mathrm{T}-1)\right\}. \quad (3)$$

The AR-HMM parameter estimation algorithm consists of the following processes.

1. The initial population parameters of the glottal source signal vector are prepared as $\mathbf{m}_{\mathrm{P}}^{(0)} = \mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{P}}^{(0)} = \mathbf{I}$. The following processes are repeated from $\mathrm{i} = 0$.
2. The AR coefficients $\mathbf{a}^{(\mathrm{i}+1)}$ and the glottal source signal vector $\mathbf{e}_{\mathrm{P}}^{(\mathrm{i}+1)}$ are estimated by maximizing the occurrence probability given by $\mathrm{L}\left(\mathbf{e}_{\mathrm{P}}^{(\mathrm{i}+1)}; \mathbf{m}_{\mathrm{P}}^{(\mathrm{i})}, \boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i})}\right)$.
3. The population parameters $\mathbf{m}_{\mathrm{P}}^{(\mathrm{i}+1)}, \boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i}+1)}$ of the glottal source signal are estimated by maximizing the likelihood given by $\mathrm{L}\left(\mathbf{e}_{\mathrm{P}}^{(\mathrm{i}+1)}; \mathbf{m}_{\mathrm{P}}^{(\mathrm{i}+1)}, \boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i}+1)}\right)$.
4. If the likelihood has converged, the algorithm stops. Otherwise, the above processes are repeated for $\mathrm{i} \leftarrow \mathrm{i}+1$ from step 2.

By iterating the above processes, the likelihood increases almost monotonically in practical situations and converges to optimum or local optimum values.

The details of each step are as follows. In step 2, the AR coefficients vector can be obtained by

$$\mathbf{a}^{(\mathrm{i}+1)} = \left[\boldsymbol{\Omega}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i})}\right)^{-1}\boldsymbol{\Omega}\right]^{-1}\boldsymbol{\Omega}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i})}\right)^{-1}\left(\mathbf{x}_{\mathrm{P}} - \mathbf{m}_{\mathrm{P}}^{(\mathrm{i})}\right). \quad (4)$$

The glottal source signal vector $\mathbf{e}_{\mathrm{P}}^{(\mathrm{i}+1)}$ is derived from (1).

In Step 3, the population parameters of the glottal source signal vector are estimated according to the following processes.

3.1 The Baum-Welch algorithm estimates the population parameters $\mu_{\mathrm{s}}^{(\mathrm{i}+1)}, \sigma_{\mathrm{s}}^{2,(\mathrm{i}+1)}, \mathrm{s} \in \mathrm{S}$ of each state's output PDF from the glottal source signal vector $\mathbf{e}_{\mathrm{P}}^{(\mathrm{i}+1)}$.

3.2 The Viterbi algorithm estimates the state transition sequence $\mathrm{s}(\mathrm{t}) \in \mathrm{S}, \mathrm{t} = \mathrm{P}, \cdots, \mathrm{T}-1$.

3.3 The expectation vector $\mathbf{m}_{\mathrm{P}}^{(\mathrm{i}+1)}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathrm{P}}^{(\mathrm{i}+1)}$ of the glottal source signal vector are estimated using (2) and (3).

3. Successive State Splitting of AR-HMM

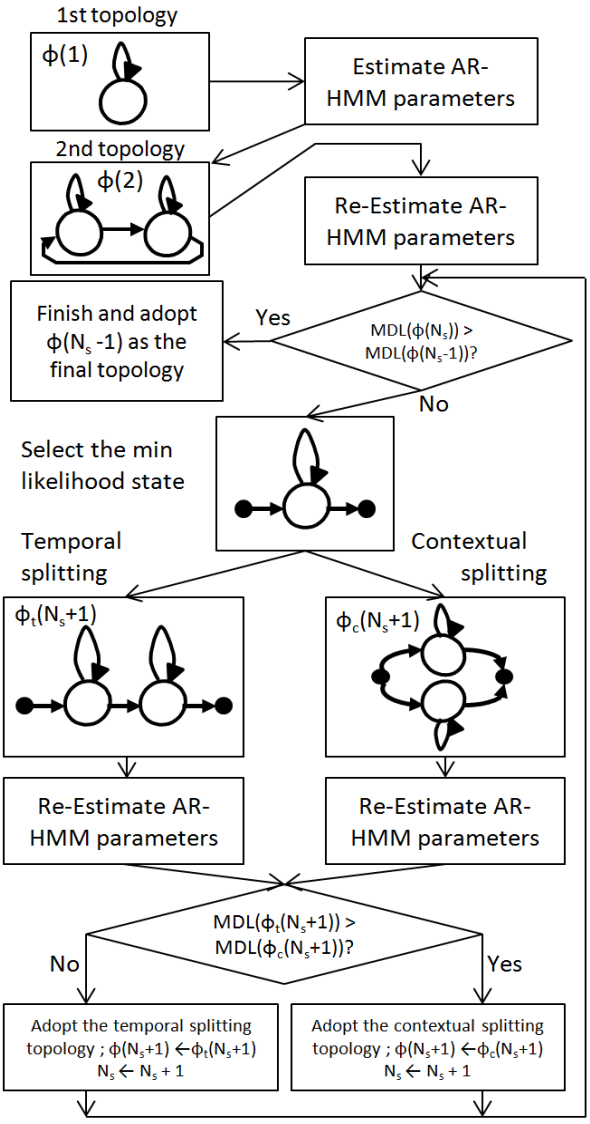

Figure 2: *Flow chart of MDL-SSS for AR-HMM*

The Successive State Splitting (SSS) algorithm was originally proposed to make a network of HMM states optimized to an individual speaker [6] and was expanded to the MDL-SSS algorithm, which conducts both contextual and temporal splitting with the MDL criterion as the splitting and stop criteria [5]. We apply the MDL-SSS to the topology generation of the GS-HMM. Figure 2 represents the flow chart of the proposed MDL-SSS for AR-HMM. The first topology has only one state. In the following, $\phi(\mathrm{N_s})$ represents the AR-HMM parameter set that consists of the AR coefficients and the parameters of the GS-HMM with $\mathrm{N_s}$ states. After the AR-HMM parameters $\phi(1)$ are estimated, the $\phi(2)$ parameters are estimated where the two states are concatenated in a ring state. These two models are then compared using the MDL criterion, which is given by the following equation.

$$\mathrm{MDL}\left(\phi(\mathrm{N_S})\right) = -\log\left(\mathrm{L}\left(\mathbf{x}_{\mathrm{p}}; \phi(\mathrm{N_s})\right)\right) + \frac{1}{2}\kappa\left(\phi(\mathrm{N_s})\right)\log(\mathrm{T}). \quad (5)$$

Here, $L\left(\mathbf{x}_p; \phi(N_s)\right)$ and $\kappa(\phi)$ represent the likelihood and the number of the AR-HMM parameters, respectively. If the MDL of $\phi(N_s)$ increases in comparison with the MDL of $\phi(N_s - 1)$, $\phi(N_s - 1)$ is selected as the final model, and the algorithm finishes. Otherwise, the likelihood of each state is evaluated according to the following equation.

$$l(s) = \prod_{t \in \left\{ t \middle| \substack{t=P,\cdots,T \\ s(t)=s} \right\}} \frac{q_{s(t),s(t+1)}}{\sqrt{2\pi\sigma_s^{2,(I)}}} \exp\left\{ -\frac{\left(e^{(I)}(t) - \mu_s^{(I)}\right)^2}{2\sigma_s^{2,(I)}} \right\} \quad (6)$$

where $q_{s1,s2}$ represents the transition probability from a state s1 to a state s2. The minimum likelihood state that is selected by $s^* = \arg\min_{s \in S} l(s)$ is then split in both temporal and contextual directions as depicted in Fig.2, where ϕ_t represents a temporally split AR-HMM parameter set and ϕ_c represents a contextually split AR-HMM parameter sets. The MDLs of both AR-HMMs are evaluated, and the parameter set of the AR-HMM with the least MDL is adopted as $\phi(N_s + 1)$. These processes are iterated until the stop condition is satisfied.

4. Experiments

4.1. Synthesis of Pseudo-Pathology Voices

In order to evaluate the fundamental validity of the pathology-voice analysis based on the proposed method, we synthesized pseudo-pathology voices according to the processes in Fig. 3. The AR coefficients to be used as vocal tract filters in the synthesis are estimated by applying an auto-correlation Linear Prediction (LP) method to a Japanese vowel /a/ of a normal speaker. We then estimate the AR coefficients by applying the Sample Average and Variance-based Iterative MAximum Likelihood (SAVIMAL) method to the pathology voices of five Japanese vowels uttered by each of twelve speakers with disorders [7]. In this experiment, we adopt prediction residuals obtained by inverse filtering of SAVIMAL's AR coefficients as the glottal sources of the pseudo-pathology voices. The prediction residuals obtained by SAVIMAL are expected to retain the time-varying characteristics of the pathology voices' glottal source. The pseudo-pathology voices are synthesized by exciting the vocal tract filter with the normal voice's AR coefficients by the prediction residuals of the pathology voices' SAVIMAL analyses. The sampling frequency was 44.1kHz. The pre-emphasis filter is given by $H(z) = 1 - 0.99z^{-1}$. The prediction order of the auto-correlation LP and SAVIMAL was set to 44. The number of samples used for evaluating the sample average and variance in SAVIMAL was set to T_s =10, which is described in the following section. The total number of synthesized pseudo-pathology voices was 60.

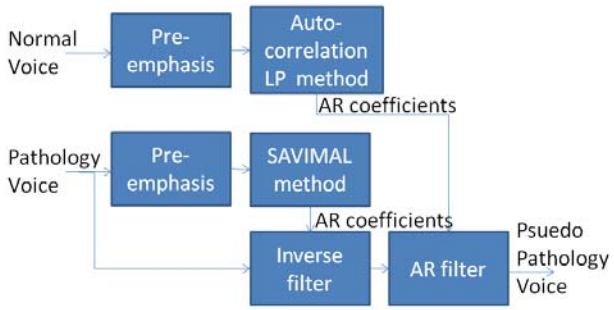

Figure 3: *Block diagram of pathology voice synthesis.*

4.2. Sample Average and Variance-based Iterative MAximum Likelihood (SAVIMAL) method

The AR-HMM parameter estimation method described in Section 2 has an HMM in order to estimate the population parameters $\mathbf{m}_P^{(i+1)}, \boldsymbol{\Sigma}_P^{(i+1)}$ of the glottal source in step 3. The SAVIMAL method has the same processes as the AR-HMM parameter estimation method except for adopting the sample averages and variances of the glottal source signal for the population parameter estimation in Step 3.

$$m^{(i+1)}(t) = \frac{1}{T_s} \sum_{k=-T_s/2}^{(T_s/2)-1} e^{(i+1)}(t+k) \quad (7)$$

$$v^{(i+1)}(t) = \frac{1}{T_s - 1} \sum_{k=-T_s/2}^{(T_s/2)-1} \left\{ e^{(i+1)}(t+k) - m^{(i+1)}(t) \right\}^2 \quad (8)$$

Here, T_s is the number of samples. The population parameters $\mathbf{m}_P^{(i+1)}, \boldsymbol{\Sigma}_P^{(i+1)}$ of the SAVIMAL method are then obtained by substituting the above sample averages and variances into (2) and (3). These population parameters estimated from the pathology voices do not always represent the exact time-varying characteristics of the glottal source. The prediction residuals of the SAVIMAL method, however, are expected to represent the time-varying characteristics better than those estimated by the conventional LP method, assuming that the glottal source conforms to an Independent and Identically-Distributed (IID) Gaussian.

4.3. Analysis Results of Pseudo-Pathology Voices

In the following experiments, we applied the proposed AR-HMM analysis with automatic topology generation to 60 synthesized pseudo-pathology voices. We transformed the estimated AR coefficients and the ideal AR coefficients used for the pseudo-pathology voice synthesis to LPC Mel-CEPstrum (LPC-MCEP) and then evaluated the error distances between the estimated LPC-MCEPs and the ideal one. In order to compare the results, we also conducted the AR-HMM analysis with the GS-HMM fixed in a ring topology of *N* states and conducted the conventional LP analysis of a covariance method. Figure 4 presents the average error distance of each method. The experiment results confirmed that the proposed AR-HMM analysis can outperform the conventional LP method and AR-HMM analysis with a fixed-ring topology.

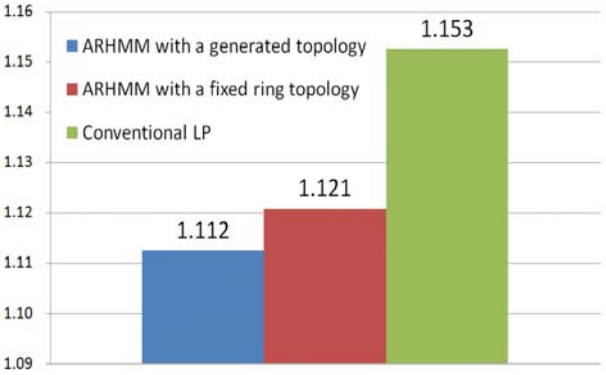

Figure 4: *Average Error Distances.*

4.4. Analysis Results of Real Pathology Voices

In the following experiments, we apply the proposed AR-HMM analysis with automatic topology generation to the 60 real pathology voices consisting of the five Japanese vowels uttered by the twelve speakers with disorders [7]. We also analyze 37 normal voices selected from the five Japanese vowels uttered by the 16 normal speakers from the JVPD database[8], so that all the vowel sounds have a stable duration exceeding 250ms. All the vowel sounds were digitized at a sampling frequency of 44.1kHz and quantized at 16 bits. The pre-emphasis filter ($H(z) = 1 - 0.99z^{-1}$) is applied to the vowel sounds prior to the analysis. The proposed method analyzed the data of the analysis frames shifting periodically during each vowel's stable duration. The analysis frame was set to 200ms. The prediction order of the AR filter was set to 44.

Figure 5 presents the average frequency in percent of adopting the GS-HMMs consisting of *N* states by the MDL criteria. GS-HMM of three states (four states) was adopted most frequently in the pathology voices (normal voices). The state-splitting processes in the normal voice analyses tend to proceed further than in the pathology voice analyses because the pathology voice's glottal sources have greater amplitude perturbations than the normal voices. In order to demonstrate this, we evaluated the normalized variances that can be obtained by dividing all the variances by the maximum variance in each GS-HMM. Figure 6 presents the average occurrence frequency of the normalized variances in percent. The X axis indicates the normalized variance, and the value of 1.0 represents the maximum variance. For the normal voices, other normalized variances are distributed around the lower area of the X axis compared with the maximum variance. As depicted in Fig. 1, this is because the blue state following the vocal fold closure tends to have a maximum variance far greater than other variances. For pathology voices, the maximum variance and other variances are more closely distributed than for normal voices. This might be caused by air leaking through the vocal folds in the pathology voices. We also counted the number of GS-HMMs with ring topology of all the GS-HMMs. The average frequency in percent of adopting the GS-HMMs with the ring topology in the pathology voices was 61.54%, which was somewhat greater than the normal voices' average frequency of 55.73%.

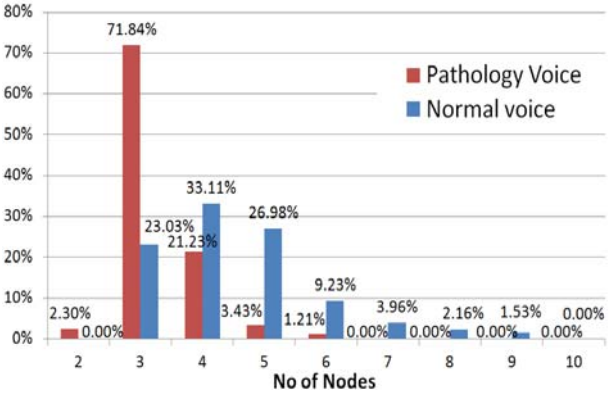

Figure 5: *Average Frequency of Adopting the GS-HMM with N states.*

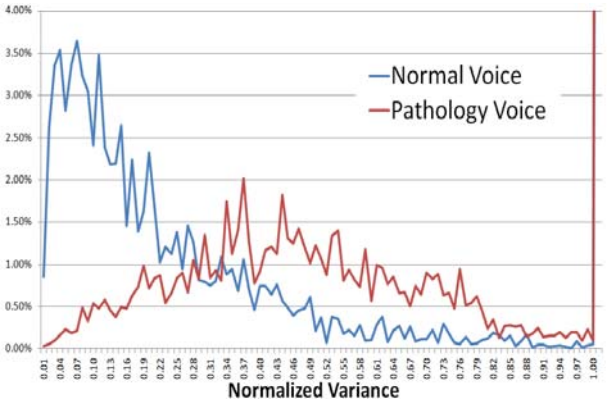

Figure 6: *Average Occurrence Frequency of Normalized Variance.*

5. Conclusions

To more accurately estimate the spectral envelope and glottal source of the pathology voice, we proposed a method that can automatically generate the topology of the Glottal Source Hidden Markov Model (GS-HMM), as well as estimate the Auto-Regressive (AR)-HMM parameter by combining the AR-HMM parameter estimation method and the Minimum Description Length-based Successive State Splitting (MDL-SSS) algorithm. This paper evaluated the fundamental validity of pathology-voice analysis based on the proposed method. The experiment results confirmed the feasibility and fundamental validity of the proposed method.

6. References

[1] Marcelo de Oliveira Rosa et al, "Adaptive Estimation of Residual Signal for Voice Pathology Diagnosis," IEEE trans. on Biomedical Engieering, Vol.47, No.1, Jan. 2000

[2] Pedro Gomez-Vilda et al, "Glottal Source biometrical signature for voice pathology detection," Speech Communication 51 (2009) pp.759-781.

[3] A.Sasou, K.Tanaka, "Glottal Excitation Modeling Using HMM with Application to Robust Analysis of Speech Signal," in Proc of ICSLP2000, Vol.IV, pp.704-707, Oct. 2000

[4] A.Sasou, M.Goto, S.Hayamizu, K.Tanaka, "Comparison of Auto-Regressive, Non-Stationary Excited Signal Parameter Estimation Methods," in Proc. of IEEE MLSP, pp.295-304, Sep. 2004.

[5] T.Jitsuhiro, T.Matsui, S.Nakamura, "Automatic Generation of Non-Uniform Context-Dependent HMM Topologies Based on The MDL Criterion," in Proc. of EUROSPEECH2003, pp.2721-2724,2003.

[6] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," in Proc. ICASSP'92, vol. 1, pp.573.576, 1992.

[7] http://www.intern.co.jp/books/bookdetails17.html

[8] http://research.nii.ac.jp/src/en/JVPD.html