Automatic Topology Generation of Glottal Source HMM

Akira Sasou

1National Institute of Advanced Industrial Science and Technology (AIST)

a-sasou@aist.go.jp

Abstract

We previously proposed the Auto-Regressive Hidden Markov Model (AR-HMM) for speech signal analysis, where the HMM was introduced as a non-stationary glottal source model. In this paper, we propose a novel method that can automatically generate the topology of the Glottal Source Hidden Markov Model (GS-HMM), as well as estimate the AR-HMM parameter obtained by combining the AR-HMM parameter estimation method and the Minimum Description Length-based Successive State Splitting (MDL-SSS) algorithm. In the experiments, we apply the proposed method to analyze the laryngeal and esophageal voices. The topology generated from the laryngeal voices tended to form a ring state, compared with the topology of the esophageal voices; this result indicates that the glottal sources of the laryngeal voices exhibit clearer periodicity than the sound sources of the esophageal voices. We also compared the vocal tract characteristics estimated by the proposed method and a conventional LP method. From these results, we were able to confirm the feasibility and the validity of the proposed method.

Index Terms: glottal source, AR-HMM, MDL-SSS

1. Introduction

The Linear Prediction (LP) method is widely used to analyze speech signals [1,2]. However, with the LP method, local peaks of LP spectral estimates are strongly biased toward harmonics, especially for high-pitched speech [3], because the glottal source signal of high-pitched speech tends to exhibit stronger non-stationary properties. In order to overcome this difficulty, the author proposed the Glottal Source Hidden Markov Model (GS-HMM) and an accompanying parameter estimation method [4,5], in which the HMM was introduced as a non-stationary glottal source model. 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 GS-HMM in the example has four states that are concatenated in a ring state so that the state transition occurs in order. This type of GS-HMM can be adopted to analyze periodic voiced sounds. The glottal source can be obtained by inverse-filtering the speech signal. The state transitions of the four states were evaluated by the Viterbi algorithm from the glottal source signal. In the figure, white represents the closed-glottis phase, green and red represent the open glottis phase, and blue includes the instant of glottal closure. 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 transitions.

Figure 1: Source filter model using glottal source HMM

2. AR-HMM Parameter Estimation with Successive State Splitting

The AR-HMM parameters consist of the AR coefficients and the parameters of the GS-HMM. Previously, we presented an algorithm [4] that iteratively estimates these parameters from a signal $x(t)$, $t = 0, \ldots, T - 1$. Let $\mathbf{a}^{(i)} = [a^{(i)}(1), \ldots, a^{(i)}(P)]^T$ 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 $e^{(i)}(t)$, $t = P, \ldots, T - 1$ is given by

$$e_p^{(i)} = x_p - \Omega a^{(i)}$$

(1)

where

$$e_p^{(i)} = [e^{(i)}(P), e^{(i)}(P + 1), \ldots, e^{(i)}(T - 1)]^T \in \mathbb{R}^{T - P}$$

$$x_t = [x(t), x(t + 1), \ldots, x(t + T - P)]^T \in \mathbb{R}^{T - P}$$

$$\Omega = [x_{T-1}, x_{T-2}, \ldots, x_0] \in \mathbb{R}^{(T - P) \times P}.$$

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

$$m^{(i)} = [m^{(i)}(P), m^{(i)}(P + 1), \ldots, m^{(i)}(T - 1)]^T.$$

(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:

$$\Sigma^{(i)} = \text{diag}(v^{(i)}(P), v^{(i)}(P + 1), \ldots, v^{(i)}(T - 1)).$$

(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 $m^{(0)} = 0$, $\Sigma^{(0)} = I$. The following processes are repeated from $i = 0$.
2. The AR coefficients $a^{(i+1)}$ and the glottal source signal vector $e_p^{(i+1)}$ are estimated by maximizing the occurrence probability given by $L(e_p^{(i+1)}, m^{(i)}_p, \Sigma^{(i)}_p)$.
3. The population parameters $m^{(i+1)}_p$, $\Sigma^{(i+1)}_p$ of the glottal source signal are estimated by maximizing the likelihood given by $L(e_p^{(i+1)}, m^{(i+1)}_p, \Sigma^{(i+1)}_p)$.
4. If the likelihood has converged, the algorithm stops. Otherwise, the above processes are repeated for $i \leftarrow i + 1$ 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

$$a^{(i+1)} = \left[ \begin{array}{c} \Omega^T (m^{(i)}_p)^{-1} \end{array} \right]^{-1} \left[ \Omega^T (m^{(i)}_p)^{-1} (x_p - m^{(i)}_p) \right].$$

(4)

The glottal source signal vector $e_p^{(i+1)}$ is derived from (1).

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

1.1 The Baum-Welch algorithm estimates the population parameters $\mu^{(i+1)}$, $\sigma^{(i+1)}$, $s \in S$ of each state's output PDF from the glottal source signal vector $e_p^{(i+1)}$.
1.2 The Viterbi algorithm estimates the state transition sequence $s(t) \in S, t = P, \ldots, T - 1$.
1.3 The expectation vector $m^{(i+1)}_p$ and the diagonal covariance matrix $\Sigma^{(i+1)}_p$ of the glottal source signal vector are estimated using (2) and (3).

### 2.2. Successive State Splitting of AR-HMM

The Successive State Splitting (SSS) algorithm was originally proposed to make a network of HMM states optimized to an individual speaker [7] 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 [6]. 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.

![Flow chart of MDL-SSS for AR-HMM](image)

The first topology has only one state. In the following, $\phi(N_s)$ represents the AR-HMM parameter set that consists of the AR coefficients and the parameters of the GS-HMM with $N_s$ 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.
Other Topology. Finally, the ASD is 40.71%.

These processes are iterated until the stop condition is evaluated according to the following equation.

\[ I(s) = \prod_{t \in \{1:T \}} \frac{1}{2\pi\sigma_{s,t}^2} \exp \left( -\frac{(e^{(t)} - \mu_{s,t}^{(t)})^2}{2\sigma_{s,t}^2(t)} \right) \]  

(6)

The minimum likelihood state that is selected by \( s^* \) is then split in both temporal and contextual directions as depicted in Fig.2, where \( \Phi_1 \) represents a temporally split AR-HMM parameter set and \( \Phi_2 \) represents a contextually split AR-HMM parameter set. 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.

3. Experiments

In order to confirm the feasibility of the proposed method, we apply the proposed method in the following experiments to analyze voiced sounds uttered by different speech production methods, using laryngeal and esophageal voices. The laryngeal voices were uttered by the usual speech production method involving oscillation of vocal folds. The esophageal voices were uttered with oscillation of the esophagus by injecting air into the upper esophagus and then releasing in a controlled manner.

3.1. Topology Generated by AR-HMM Analysis

In the experiments, five laryngeal speakers and five esophageal speakers uttered five Japanese vowel sounds (/a/, /e/, /i/, /o/, and /u/). A total of fifty vowel sounds were used for the experiments. The vowel sounds were digitally recorded at a sampling frequency of 16kHz and quantization of 16 bits. Each vowel sound was analyzed frame by frame. The topology of the GS-HMM was generated from each frame. The frame size was 20ms, and the shift period was 10ms. The prediction order \( P \) was fixed to 16 for all analyses.

Figure 3 indicates the average frequency in percent of adopting the GS-HMM with \( N \) states.

<table>
<thead>
<tr>
<th>Topology Generated by AR-HMM Analysis</th>
<th>Ring Topology</th>
<th>Other Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laryngeal Voice</td>
<td>59.29 %</td>
<td>40.71 %</td>
</tr>
<tr>
<td>Esophageal Voice</td>
<td>45.30 %</td>
<td>54.70 %</td>
</tr>
</tbody>
</table>

Table 1: Average Frequency of Adopting the GS-HMM with Ring Topology.

3.2. Comparison of Vocal Tract Characteristics

In the following experiments, the conventional LP method with the Hanning window was also used for analysis of the same vowel sounds. We compared the vocal tract characteristics estimated by both the AR-HMM and the LP methods. Figures 4 and 5 illustrate examples of the vocal tract characteristics estimated by the LP method vary rapidly along the time axis. The vocal tract does not vary as rapidly. This example demonstrates that the LP method cannot correctly separate the characteristics of the vocal tract and the esophageal sound source, and the rapid variations caused by the sound source of the esophageal voice remain in the estimated vocal tract characteristics. This is because the LP method's assumption that the sound source conforms to an Independent Identically Distributed (IID) normal distribution is not appropriate for such a non-stationary sound source of the esophageal voice. In order to compare variations of the estimated vocal tract characteristics, we evaluate the Average Standard Deviation (ASD) as follows. Let \( A_1(1), \ldots, A_1(P) \) denote the AR coefficients extracted from the t-th frame. The logarithmic amplitude response of the AR filter is given by

\[ A_t(w) = -20\log_{10}[1 + \sum_{n=1}^{P} a_t(n)e^{-j\omega_n}] \]  

At the discrete frequencies, \( \omega_n = \frac{n\pi}{100}, n = 1, \ldots, 100 \), the mean is evaluated according to

\[ \overline{A}(\omega_n) = \sum_{t=1}^{T} A_t(\omega_n) / T \]  

and the standard deviation is evaluated according to

\[ \sigma(\omega_n) = \sqrt{\sum_{t=1}^{T} (A_t(\omega_n) - \overline{A}(\omega_n))^2 / T} \]  

Finally, the ASD is obtained by \( \overline{\sigma} = \sum_{\omega_n}^{N_{\omega_n}} \sigma(\omega_n) / 100 \). Figure 6 indicates the averages of the five ASDs evaluated using the AR coefficients estimated from each vowel's voices uttered by the five esophageal speakers by the LP and the AR-HMM methods. The periodicity of the esophageal voices’ sound sources are not always as clear, and the characteristics of the esophageal voices’ sound sources tend to vary in each pitch.
results confirm that the vocal tract characteristics estimated by the AR-HMM were smoother than those estimated by the LP method; thus, the AR-HMM approach can separate the vocal tract characteristics and the sound source more accurately.

Figure 7 indicates the averages of the five ASDs evaluated from the five laryngeal voices for each vowel. The vocal tract characteristics estimated by the LP method from the laryngeal voices were smoother than those of the esophageal voices. Compared with the LP method, the AR-HMM-based approach can more accurately separate the vocal tract and the glottal source not only in esophageal voice analyses but also in laryngeal voice analyses.

4. Conclusions

In this paper, we proposed a novel method that can automatically generate the topology of the GS-HMM, in addition to AR-HMM parameter estimation obtained by combining the AR-HMM parameter estimation method and the Minimum Description Length-based Successive State Splitting (MDL-SSS) algorithm. In the experiments, we applied the proposed method to analyze the laryngeal and esophageal voices. The topology generated from the laryngeal voices tended to form a ring state, compared with that of the esophageal voices; this result indicates that the glottal sources of the laryngeal voices exhibit clearer periodicity than the sound sources of the esophageal voices. We are now planning to develop an AR-HMM-based approach that achieves detailed analysis by referring to the topology and integrating the probability and statistical parameters of the GS-HMM.

5. References