Order estimation of speech production model
based on the eigenvalue ratios of
quasi-covariance matrix

Introduction

Recently, speech analysis based on the autoregressive moving average (ARMA) model has been employed extensively to nasals and nasalized sounds. In such cases, it is necessary to estimate the orders of both AR and MA processes of the speech production system. This paper presents a method for estimating the order of the AR part of speech production. It is emphasised that the order of the AR part in speech can be estimated by this method independently of the MA part even if the speech is generated by an ARMA model. This method utilizes the eigenvalues of the quasi-covariance matrix of the speech waveform within the glottal closure interval of voiced speech.

Theoretical Basis

The speech production system here is assumed to be an ARMA model of order \( p_0 \) for AR and \( q_0 \) for MA defined as follows:
\[
s_n = \sum_{k=1}^{p_0} a_k s_{n-k} + \sum_{m=0}^{q_0} b_m u_{n-m}
\]
where \( s_n \) and \( u_n \) are the \( n \)-th samples of speech and the excitation waveforms, respectively, \( a_k \) is the \( k \)-th AR parameter and \( b_m \) is the \( m \)-th MA parameter. Let \( \{s_0, s_1, s_2, \ldots, s_{N-1}\} \) be a set of \( N \) speech samples obtained from the speech production system, and a new quasi-covariance matrix \( \Phi^{(i_0)} \) is defined as
\[
\phi^{(i_0)}_{ij} = \phi^{(i_0)}
\]
where \( i_0 \) is the initial point parameter for the matrix. If \( p \), the size of \( \Phi^{(i_0)} \), is greater than \( p_0 \), the following equation is obtained from Eqs. (1) and (2).
\[
\phi^{(p)}_{ij} = \sum_{k=1}^{p_0} a_k \phi^{(i_0)}_{i+j-k} + \sum_{m=0}^{q_0} b_m \sum_{n=p+i_0}^{N-1} s_{n-i_0-i} u_{n-j-m},
\]
for \( i = 1, 2, \ldots, p \), \( j = 1, 2, \ldots, p-p_0 \).

If the analysis frame is located within the interval in which \( u_n \) remains zero, rank \( \Phi^{(i_0)} \) coincides with \( p_0 \) according to Eq. (3). Glottal closure intervals in voiced speech can be candidates for the interval in which \( u_n \) remains zero. For natural voiced speech, however, \( u_n \) would not remain zero but it might be assumed to be small magnitude white noise even in the closure intervals. If \( i_0 < q_0 \) and the whole analysis frame is located within such an interval, the following approximation holds according to the causal relation between input \( u_n \) and output \( s_n \).
\[
\sum_{n=p+i_0}^{N-1} s_{n-i_0-i} u_{n-j-m} \approx 0,
\]
for \( i \geq j, j = 1, 2, \ldots, p-p_0 \), \( m = 0, 1, 2, \ldots, q_0 \).

From Eqs. (3) and (4), \( \phi^{(p)}_{ij} \) is approximated by the linear combination of \( \phi^{(i_0)}_{ij} \). Therefore rank \( \Phi^{(i_0)} \) is nearly equal to \( p_0 \) and we can expect that the ratio \( |\lambda_i/\lambda_{i+1}| \) yields a comparatively large value at \( i = p_0 \) provided that the eigenvalues \( \lambda_i \) are numbered in descending order of the absolute value.

It is difficult to investigate the validity of Eq. (4) and its effects on the eigenvalue ratios of the quasi-covariance matrix for short analysis frames such as the glottal closure intervals. In the following section, computer simulation of the proposed order estimation method is made using synthetic and natural speech.

Experimental Results

(a) Synthetic speech

Figure 1 shows the block diagram for speech synthesis. The excitation wave \( u_n \) for the ARMA system is the sum of glottal wave \( u_{n0} \) and white noise \( w_n \). The signal to noise ratio \( S/N \) of \( u_n \) is defined as follows:
\[
S/N = 10 \log_{10} ( \sum_{n=0}^{N_0-1} u_{n0}^2 / N_0 \sigma^2 ),
\]
where \( N_0 = T_0 / T_p \), \( T_0 \) is the pitch period of glottal wave, \( T_p \) is the sampling interval and \( \sigma^2 \) is the variance of the noise.

In order to investigate the performance of the proposed method, several synthetic sounds were prepared. The synthetic sounds used in the simulation have five poles and one zero for vocal tract characteristics and another two zeros on the real axis; one for radiation characteristics and the other for spectral slope compensation. Accordingly, \( p_0 = 10 \) and \( q_0 = 4 \). A set of
Fig. 1 Block diagram of speech synthesis.

\[ u_n = a_n + \sum_{m=1}^{a_k} a_m u_{n-k} + \sum_{m=1}^{a_k} b_m u_{n-m} \]

(a) When \( i_0 = 0 \) and (b) when \( i_0 = 7 \).

Fig. 2 Eigenvalue ratio \( |\lambda_i/\lambda_{i+1}| \) of the matrix \( \phi^{(i_0)} \), where the signal to noise ratio of the excitation wave is 40 dB.

Table 1 Poles and a zero used as the vocal tract characteristics.

<table>
<thead>
<tr>
<th>Frequency (Hz)</th>
<th>Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>700.0</td>
<td>54.1</td>
</tr>
<tr>
<td>1300.0</td>
<td>64.1</td>
</tr>
<tr>
<td>2500.0</td>
<td>102.1</td>
</tr>
<tr>
<td>3500.0</td>
<td>152.1</td>
</tr>
<tr>
<td>4500.0</td>
<td>218.7</td>
</tr>
<tr>
<td>Zero</td>
<td>1750.0</td>
</tr>
</tbody>
</table>

Fig. 2 illustrates the eigenvalue ratio \( |\lambda_i/\lambda_{i+1}| \) of the matrix \( \phi^{(i_0)} \) when \( S/N = 40 \) dB and the analysis frame is located at the glottal closure interval. If we set \( i_0 = 0 \) equal to zero, the eigenvalue ratio has a maximum value at \( i = 12 \), but if we select \( i_0 = 7 \) \((>q_0)\), the maximum is at \( i = 10 \) \((=p_0)\).

From these results, it can be stated that rank \( \phi^{(i_0)} \) is nearly equal to \( p_0 \) even in such a short analysis frame as the glottal closure interval if the parameter \( i_0 \) is selected properly.

(b) Natural speech

In order to confirm the performance of the proposed method on natural speech sounds, the method was applied to the CV syllables composed of nasals and five vowels uttered by two male adults. Those syllables can be considered to be produced by an ARMA model.

To apply the proposed method successfully, the analysis frame must be located at the glottal closure interval. However, it is very difficult to detect the glottal closure interval in a natural speech waveform. As a simple and convenient method to locate the analysis frame properly, the front end of the analysis frame is located near the zero-cross point just before the maximum amplitude point in each period. As for the length of the analysis frame, about a half pitch period is used.

The method was applied to the above-mentioned speech sounds and the results were successful for all the data. One of the analysis results is depicted in Fig. 3, where the power spectrum and the eigenvalue ratio of nasal consonant /m/ in a syllable /ma/ are shown. The order of the AR part estimated by the proposed method is six. As the number of dominant peaks of the spectrum envelope of Fig. 3(b) appears to be three, the order of the AR part of this nasal consonant should be estimated to be six. The order of the AR part estimated by the proposed method coincides with the value presumed from the spectrum.

Conclusion

A method for estimating the order of AR part of the speech production system has been presented by using the eigenvalue ratios of the quasi-covariance matrix of
speech introduced here. Although the amount of speech data is not so large, the experimental results affirm that the performance of the proposed method is not affected by the MA part even if the speech is generated by an ARMA model. Considering the successful results on speech sounds with zeros, it is obvious that this method is applicable to voiced speech sounds without zero.

Reference