Common Acoustical Pole Estimation from Multi-Channel Musical Audio Signals

Takuya YOSHIOKA†, Student Member, Takafumi HIKICHI††, Masato MIYOSHI††, Members, and Hiroshi G. OKUNO††, Nonmember

SUMMARY This paper describes a method for estimating the amplitude characteristics of poles common to multiple room transfer functions from musical audio signals received by multiple microphones. Knowledge of these pole characteristics would make it easier to manipulate audio equalizers, since they correspond to the room resonance. It has been proven that an estimate of the poles can be calculated precisely when a source signal is white. However, if a source signal is colored as in the case of a musical audio signal, the estimate is degraded by the frequency characteristics of its spectral envelope and fine structure. We assume that musical pieces can be classified into several categories according to their average amplitude spectral envelopes. Based on this assumption, the amplitude spectral envelope of the musical audio signal can be obtained from prior knowledge of the average amplitude spectral envelope of a musical piece category into which the target piece is classified. On the other hand, the fine structure is identified based on its time variance. By removing both the spectral envelope and the fine structure from the amplitude spectrum estimated with the conventional method, the amplitude characteristics of the acoustical poles can be extracted. Simulation results for 20 popular songs revealed that our method, the amplitude characteristics of the acoustical poles, can be estimated precisely when a source signal is a musical audio signal. We consider that the prior knowledge of such average amplitude spectral envelopes can be used to represent a room transfer function, which describes sound transmission characteristics from a source to a microphone in a room [1]. The AR part represents the poles common to transfer functions between multiple sources and microphones in a room and describes the room resonance characteristics. The MA part varies depending on the source and microphone positions.

If such common acoustical poles were estimated from signals received by microphones, even technically untrained musical performers would be able to control the effects of resonances easily. Such pole estimation is possible prior to performances with the conventional method by using a white source signal [2]. However, pole estimation during the performances seems much more convenient for the performers because it requires no additional task for the pole estimation. In addition, compared to the a priori pole estimation, it might be robust to room soundfield fluctuation caused by various acoustic properties such as room temperature or presence of audiences because it does not require any a priori measure of the soundfield. However, when a source signal is colored as in the case of a musical audio signal, the poles estimated with the conventional method are smeared by the characteristics of this source signal [3].

In this paper, the conventional common-pole estimation method mentioned above is extended so that it can estimate the amplitude characteristics of such poles even when a source signal is a musical audio signal. We consider that the amplitude spectrum estimated with the conventional method from a musical audio signal windowed in a short-time frame may include the time-invariant amplitude spectrum of the common acoustical poles and the amplitude spectral envelope and fine structure of the musical signal. We assume here musical pieces can be classified into several categories according to their average amplitude spectral envelopes; the average of the amplitude spectral envelopes over all the time frames covering the musical signal is similar to that of a musical piece category to which the target piece belongs. We use the prior knowledge of such average amplitude spectral envelopes of musical piece categories as the average amplitude spectral envelope of the musical signal. On the other hand, since the amplitude spectral fine structure of the musical signal seems to change frame by frame, the fine structure in each time frame may be extracted by comparing the spectrum in that frame with those in several adjoining frames. By eliminating the spectral envelope and the fine structure from the spectrum obtained with the conventional method, we can estimate the amplitude spectrum of...
the common acoustical poles.

2. Estimation of Common Acoustical Poles from Musical Audio Signals

In this paper, we deal with a single-source and two-microphone system as shown in Fig. 1. The two room transfer functions $G_1(z)$ and $G_2(z)$ can be expressed as

$$G_i(z) = \frac{H_i(z)}{A(z)} = \sum_{n=0}^{\infty} h_i(n) z^{-n} / \sum_{n=1}^{\infty} a(n) z^{-n},$$  \hspace{1cm} (1)$$

where $a(n)$ and $h_i(n)$ denote the coefficients of the $K$-th order polynomial $A(z)$ (the AR part of $G_i(z)$) and those of the $J$-th order polynomial $H_i(z)$ (the MA part of $G_i(z)$), respectively [1]. Source signals $s(k)$ are received by two microphones after being transmitted through $G_1(z)$ and $G_2(z)$. The signals observed by the microphones are denoted as $x_1(k)$ and $x_2(k)$.

Here, we assume the following conditions.

- **Assumptions for room soundfields**
  - **(A1)** Room transfer functions $G_1(z)$ and $G_2(z)$ are time invariant.
  - **(A2)** The MA polynomials of $G_1(z)$ and $G_2(z)$, $H_1(z)$ and $H_2(z)$, have no common zeros [4].

- **Assumptions for musical audio signals**
  - **(A3)** A source signal observed in the $i$-th time frame is produced through an AR process:

$$\frac{1}{B_i(z)} = \frac{1}{B_i'(z)B_i''(z)}.$$  \hspace{1cm} (2)$$

The amplitude spectrum of $1/B_i'(z)$, $1/B_i''(e^{j\omega})$, corresponds to the amplitude spectral envelope of the source signals in the $i$-th time frame that is obtained by adapting the one-step forward linear prediction analysis [5] of a small order $O$ to the source signals in that frame. On the other hand, $1/B_i''(e^{j\omega})$ corresponds to the amplitude spectral fine structure of the signals. Figure 2 shows examples of these two spectra.

By adopting assumption (A3), the system in Fig. 1 can be expressed as the schematic diagram shown in Fig. 3.

(A4) Musical pieces may be classified into several categories according to their average amplitude spectral envelopes. In other words, we assume that for each musical piece belonging to a category $S$, the average of $\log \left| 1/B_i'(e^{j\omega}) \right|$ converges to the average logarithmic amplitude spectral envelope $\log \left| 1/B_S'(e^{j\omega}) \right|$ of category $S$:

$$E\{\log \left| \frac{1}{B_i'(e^{j\omega})} \right| \} = \log \left| \frac{1}{B_S'(e^{j\omega})} \right|.$$  \hspace{1cm} (3)$$

where $E\{\cdot\}$ is an average function.

The average logarithmic amplitude spectral envelope for category $S$, $\log \left| 1/B_S'(e^{j\omega}) \right|$, is calculated as follows. We consider $M$ musical pieces that belong to category $S$. First, the logarithmic amplitude spectral envelope of the $i$-th time frame of the $m$-th piece, $\log \left| 1/B_{i,m}'(e^{j\omega}) \right|$, is calculated using $W$-point Hamming window shifted by $I$ points to obtain $N'$ signal excerpts ($1 \leq i \leq N'$ and $1 \leq m \leq M$). Then, these spectra are averaged over time frames and pieces to acquire the average logarithmic amplitude spectral envelope of the category $S$:

$$\log \left| \frac{1}{B_S'(e^{j\omega})} \right| = \frac{1}{MN'} \sum_{m=1}^{M} \sum_{i=1}^{N'} \log \left| \frac{1}{B_{i,m}'(e^{j\omega})} \right|.$$  \hspace{1cm} (4)$$

Figure 4 shows all the procedures for estimating the logarithmic amplitude characteristics of the AR process contained in room transfer functions $G_1(z)$ and $G_2(z)$. First,
observed signals $x_1(k)$ and $x_2(k)$ are windowed by a $W$-point Hamming window with a shift of $I$ points to obtain $N$ excerpts of the signals. Next, each excerpt is processed by using the conventional common-pole estimation method [2] (see Appendix) to estimate the amplitude spectrum, $|1/C_i(e^{j\omega})|$, of the AR process common to the two signal-transmission paths in the $i$-th time frame. Looking into Fig. 3, it is clear that $|1/C_i(e^{j\omega})|$, namely the estimate of $|1/C_i(e^{j\omega})|$ calculated with the conventional method, can be expressed as

$$
\log \left| \frac{1}{C_i(e^{j\omega})} \right| = \log \left| \frac{1}{B_i(e^{j\omega})B_i(e^{j\omega}) + \delta_{B_i}(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega}) + \delta_{A_i}(e^{j\omega})} \right|,
$$

(5)

where $\delta_{A_i}(e^{j\omega})$ and $\delta_{B_i}(e^{j\omega})$ are computational errors. Then, the additional procedures depicted in Fig. 4 and proposed in this paper are used to improve the estimate of $\log |1/A(e^{j\omega})|$. These procedures are described below.

2.1 Spectral Smoothing

From Eq. (5), we have

$$
\log \left| \frac{1}{C_i(e^{j\omega})} \right| = \log \left| \frac{1}{B_i(e^{j\omega})B_i(e^{j\omega}) + \delta_{B_i}(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega}) + \delta_{A_i}(e^{j\omega})} \right|.
$$

(6)

We can rewrite this equation as

$$
\log \left| \frac{1}{C_i(e^{j\omega})} \right| = \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega})} \right| + \log \left| \frac{1}{1 + \delta_{B_i}(e^{j\omega})B_i(e^{j\omega})} \right|.
$$

(7)

When we assume that the computational errors are sufficiently small:

$$
|\delta_{B_i}(e^{j\omega})| \ll |B_i(e^{j\omega})B_i(e^{j\omega})|,
$$

(8)

$$
|\delta_{A_i}(e^{j\omega})| \ll |A(e^{j\omega})|.
$$

(9)

We then consider the last two terms of the right-hand side of Eq. (7) as the logarithmic amplitude of spectral error $\varepsilon_i(e^{j\omega})$:

$$
\log \left| \frac{1}{C_i(e^{j\omega})} \right| = \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega})} \right| + \log |\varepsilon_i(e^{j\omega})|.
$$

(10)

Applying an operator, $S \{ \cdot \}$, which smoothes temporally local spectral peaks to Eq. (10) gives

$$
S \left\{ \log \left| \frac{1}{C_i(e^{j\omega})} \right| \right\} \approx \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega})} \right| + S \left\{ \log \left| \frac{1}{B_i(e^{j\omega})} \right| \right\} + \log |\varepsilon_i(e^{j\omega})|.
$$

(11)

$$
\approx \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega})} \right| + \log |\varepsilon_i(e^{j\omega})|.
$$

(12)

This expansion is based on an assumption that peaks in $\log |1/C_i(e^{j\omega})|$ are caused by those in $\log |1/B_i(e^{j\omega})|$, which usually holds in a practical case.

Operator $S \{ \cdot \}$ was implemented using the following algorithm. Figure 5 is a view showing the frame format of this algorithm.

1. Detect the spectral peaks in $\log |1/C_i(e^{j\omega})|$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{system_overview.png}
\caption{System overview.}
\end{figure}
2. From the detected spectral peaks in each time frame, select the temporally local ones. Each spectral peak is judged to be temporally local if its frequency varies by more than 50 cents (half of the semitone difference) from the any of the peak frequencies in the other time frames. The conversion of frequency \( f_{\text{Hz}} \) in hertz to frequency \( f_{\text{cent}} \) in cents is given by

\[
f_{\text{cent}} = 1200 \log_2 \frac{f_{\text{Hz}}}{440 \times 2^{\frac{1}{12}}}. \tag{13}\]

where 440 Hz is converted to 5700 cents [6].

3. Calculate a third-order spline function using spectral points except those in the spectral “hills” containing each temporally local peak so as to eliminate these peaks. We defined a spectral point as being in a spectral hill if the frequency of the point is between the two local-minimum frequencies adjacent to the peak of the hill (Fig. 6).

By employing the expectation of Eq. (12), we have

\[
E \left( S \left( \log \left| \frac{1}{C_i(e^{j\omega})} \right| \right) \right) \\
\approx E \left( \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega})} \right| \right) \\
+ E \left\{ \log |e_i(e^{j\omega})| \right\}. \tag{14}\]

Assuming that a spectral error \( e_i(e^{j\omega}) \) is a white noise spectrum, we obtain

\[
E \left\{ \log |e_i(e^{j\omega})| \right\} = 0. \tag{15}\]

Substituting Eq. (15) into Eq. (14) leads to

\[
E \left( S \left( \log \left| \frac{1}{C_i(e^{j\omega})} \right| \right) \right) \approx E \log \left| \frac{1}{B_i(e^{j\omega})} \right| + \log \left| \frac{1}{A(e^{j\omega})} \right| \tag{16}\]

Thus, the spectral smoothing procedure will remove source-signal spectral fine structures from the estimates obtained with the conventional method.

2.2 Spectral Subtraction

According to assumption (A4), the first term of the right-hand side of Eq. (16) coincides with the average spectrum \( \log |1/B_i(e^{j\omega})| \). Therefore, it can be eliminated as

\[
E \left( S \left( \log \left| \frac{1}{C_i(e^{j\omega})} \right| \right) \right) - \log \left| \frac{1}{B_i(e^{j\omega})} \right| \approx -\log \left| \frac{1}{A(e^{j\omega})} \right|. \tag{17}\]

Thus, the spectral subtraction procedure will remove source-signal spectral envelopes.

3. Simulations

Simulations were conducted to evaluate the proposed method.

3.1 Conditions

Figure 7 outlines the simulation setup, where a simple soundfield is surrounded by three reflectors. Room transfer functions \( G_1(z) \) and \( G_2(z) \) were simulated as follows. MA coefficients \( h_i(n) \) were simulated using the image method [7]. The AR coefficients, \( a(n) \), of the room were obtained based on the theoretical resonant frequencies calculated with \( f_i = ic/2L \), where L is the distance between two opposing walls, \( c \) is sound velocity, and \( i \) is an integer. The parameters used in this simulation are listed in Table 1. Figure 8 shows the impulse responses of room transfer functions \( G_1(z) \) and \( G_2(z) \).

The source signals were 20 popular Japanese musical pieces taken from the RWC Music Database: Popular Music (RWC-MDB-P-2001: No.1–20) [8]. In the simulations, we assumed that these 20 songs are classified into a single musical piece category. Prior to the simulations, we verified this assumption. When we calculated the average of the spectral distortions (SDs) [9] between the average amplitude spectral envelope of each piece and that of the category obtained under the assumption, the average of the SDs was 1.29 dB. The
Table 1  Simulation parameters.

<table>
<thead>
<tr>
<th>Simulation condition parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( J ) Order of zeros</td>
<td>100</td>
</tr>
<tr>
<td>( K ) Order of poles</td>
<td>17</td>
</tr>
<tr>
<td>( f_1 ) Lowest resonance frequency</td>
<td>243 Hz</td>
</tr>
<tr>
<td>( f_i ) Resonance frequencies</td>
<td>( i = 1, \cdots, 16 )</td>
</tr>
<tr>
<td>( f_s ) Sampling frequency</td>
<td>8 kHz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimation parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( W, W' ) Frame length</td>
<td>6000</td>
</tr>
<tr>
<td>( I, I' ) Frame interval</td>
<td>3000</td>
</tr>
<tr>
<td>( 2L - 1 ) Order of estimated poles</td>
<td>239 (see Appendix)</td>
</tr>
<tr>
<td>( N, N' ) Number of time frames</td>
<td>30</td>
</tr>
<tr>
<td>( O ) Order of linear prediction</td>
<td>5</td>
</tr>
</tbody>
</table>

Fig. 8  Room impulse responses for each microphone.

SD represents the mean squared error on a dB scale between two spectra \( P(f) \) and \( P'(f) \), which is defined by

\[
SD = \frac{1}{F} \sum_{f=0}^{F-1} (20 \log |P(f)| - 20 \log |P'(f)|)^2, \tag{18}
\]

where \( F \) denotes the number of frequency bins. Because the average SD was not so large, we conducted the simulation using the single musical piece category.

We performed a 10-fold cross validation. That is, we first divided all of the songs into 10 sets, and then repeated the test 10 times. Each time, we extracted one set for evaluation and used the rest to obtain the average logarithmic amplitude spectral envelope of the category.

The estimation results were evaluated using the SD between the given and estimated all-pole spectra. The given all-pole spectrum was calculated as \( |1/A(e^{j\pi})| \).

3.2 Results

We obtained an average SD of 3.11 dB for the 20 songs. Figure 9 summarizes the results. An example of the estimated and given amplitude characteristics is shown in Fig. 10(a). The estimated spectrum in Fig. 10(a) is very similar to that of the room AR process. Indeed, almost all the spectral peak frequencies were precisely detected, and the peak amplitudes were also estimated well. When only the spectral subtraction procedure was disabled, the average SD was 6.35 dB. This performance degradation can be seen in detail in Fig. 10(b). Although many of the peak frequencies of the room AR process were correctly detected, the differences in amplitudes were significant. These results show that the spectral subtraction procedure plays a major role in reducing SD values. In contrast, when only the spectral smoothing procedure was disabled, the average SD was 3.34 dB. The SD value reduction obtained with the spectral smoothing is much smaller than that obtained with the spectral subtraction. However, when we compare the estimated spectrum shown in Fig. 10(c) with that in Fig. 10(a), we find that many spectral peaks originally contained in the source signals were attenuated by the spectral smoothing.

3.3 Discussion

Although the spectral smoothing procedure was able to attenuate spectral peaks originally contained in the source signals, several sharp spectral peaks of the source signals still remained as shown in Fig. 10(a). In order to improve the effectiveness of this procedure, the following problems should be addressed.

- The current spectral smoothing procedure cannot distinguish spectral peaks caused by source signals from those of a room AR process when source-signal spectral peaks happen to appear at the same frequencies in all time frames. This problem can be found especially in the low frequency range since the harmonic components of various tones are likely to overlap in this frequency region.
- The current spectral smoothing procedure cannot distinguish between the spectral peaks of source signals and those of a room AR process when they overlap.
Fig. 10 Examples of the given and estimated all-pole spectra when (a) estimated with the proposed method, (b) spectral subtraction was disabled, (c) spectral smoothing was disabled, and (d) estimated in the closed condition.

When we used the average logarithmic amplitude spectral envelope of target source signals for the spectral subtraction (closed condition), the average SD was 2.95 dB (see Fig. 9). The amplitude characteristics of the common acoustical poles are shown in Fig. 10(d). Here, we used the same source signals as those used in Fig. 10(a). Comparing these two figures, we can see that the spectral peak frequencies were detected very well in both open and closed conditions. Figure 10(a) shows more degradation in the frequency range above 3 kHz. These observations indicate that our proposed method, which uses the average spectrum of a musical piece category into which a target piece is classified, is effective especially in estimating spectral peak frequencies.

All the songs used in the simulation are typical popular musical pieces, which have melody vocals and bass lines that form the predominant sounds of the pieces. Hence, we speculate that the good match between the average amplitude spectral envelope of a source signal and that of its piece category were resulted from such similarity in predominant sounds. A finding in the field of musical instrument identification may partly support our thinking. Brown [10] reports that musical instruments can be identified using amplitude spectral envelopes calculated from long-term signals. It should be noted, however, that his result was obtained by using solo tones. Although we have confirmed that classification of musical pieces based on their logarithmic amplitude spectral envelopes is possible for a part of popular, orchestral, and jazz music in our subsequent experiments [11], we still need to investigate whether more variety of musical pieces can be classified.

4. Conclusion

This paper has described a method for estimating the amplitude characteristics of common poles of room transfer functions from signals received by microphones and a musical piece category into which a target piece is classified. The advantage of the proposed method is that such characteristics can be estimated even when a musical audio signal is used as a source signal, while the conventional method needs a stationary random signal. This improvement is mainly the result of having prior knowledge of an average logarithmic amplitude spectral envelope of a musical piece category into which a target piece is classified. The simulation results for 20 popular songs showed that the proposed method could estimate the amplitude characteristics of the poles with a spectral distortion of 3.11 dB and that most of the spectral peaks, corresponding to the poles, were successfully detected.
The simulation results are promising in that the proposed method provides both an expert audio engineer and a non-expert performer with the simple but useful information for controlling room resonances. Currently, we assume that a performer knows the musical piece category into which a target musical piece is classified. If a number of musical pieces could be classified according to their logarithmic amplitude spectral envelopes, information on the musical pieces could be classified according to their logarithmic amplitude spectral envelopes, information on the musical piece may be then available, for instance, from a table of those musical piece categories.

References


Appendix: Calculation of 1/\( \hat{C}(z) \) with the Conventional Method

The conventional common-pole estimation method [2] for calculating \( 1/\hat{C}(z) \), an estimate of \( 1/C(z) \) in Fig. A-1, is summarized below.

1. First, a two-channel linear prediction matrix \( Q \) defined by the following equations is calculated.

\[
Q = E\{x_{n-1}x_n^T\} \quad (A \cdot 1)
\]

\[
x_n = [x_1(n), \cdots, x_1(n-L), x_2(n), \cdots, x_2(n-L)]^T,
\quad (A \cdot 2)
\]

where \( A^+ \) denotes the Moore-Penrose inverse matrix [12] of \( A \), \( A^T \) denotes the transposed matrix of \( A \), and \( L \) is a positive integer satisfying

\[
L \geq J - 1
\quad (A \cdot 3)
\]

and

\[
L \geq K - J.
\quad (A \cdot 4)
\]

2. Polynomial \( \hat{C}(z) \) is then calculated as the characteristic polynomial of matrix \( Q \), which theoretically corresponds to the AR polynomial, \( C(z) \), common to two signal transmission paths [2]. In practice, since computational errors are inevitable, \( \hat{C}(z) \) can be expressed as Eq. (5).

Takuya Yoshioka was born in Osaka, Japan, on December 21, 1981. He received the B.E. degree from Kyoto University, Kyoto, Japan, in 2004. He is currently pursuing the M.E. degree in the same university. He is a student member of ASJ and IPSJ.

Takafumi Hikichi was born in Nagoya, in 1970. He received his Bachelor and Master of Electrical Engineering degrees from Nagoya University in 1993 and 1995, respectively. In 1995, he joined the Basic Research Laboratories of NTT. He is now working at the Signal Processing Research Group of the Communication Science Laboratories, NTT. His research interests include physical modeling of musical instruments, room acoustic modeling, and signal processing for speech enhancement and dereverberation. He was honored to receive the Kiyoshi-Awaya Incentive Awards from the ASJ in 2000. He is currently a visiting associate professor of the Graduate School of Information Science, Nagoya University. He is a member of IEEE, ASA, ASJ, and IPSJ.
Masato Miyoshi received M.E. from Doshisha University in Kyoto in 1983. Since joining NTT as a research staff member this year, he has been engaging on the research and development of acoustic signal processing technologies. Currently, he is in charge of a group leader of the Media Information Lab. of NTT Communication Science Labs. He was honored to receive 1988 IEEE ASSP Senior Awards, 1989 ASJ Kiyoshi-Awaya Incentive Awards, and 1990 ASJ Satoh Paper Awards. He also received Ph.D. from Doshisha Univ. in 1991. He is a member of IEEE, AES, and ASJ.

Hiroshi G. Okuno is a Professor of Computer Science at the Graduate School of Informatics, Kyoto University, in Kyoto, Japan. He received his B.A. in Pure and Applied Sciences from the University of Tokyo, and his Ph.D. in Computer Science from the same university. Before joining Kyoto University, he worked for NTT, Kitano Symbiotic Systems Project, and Science University of Tokyo. His research interests include Computational Auditory Scene Analysis and its subfields Robot Audition, Speech Media Processing, Automatic Onomatopoeia Recognition, and Music Scene Analysis, with emphasis on development of enabling technologies for intelligent systems. He received the Best Paper Award from JSAI in 1991, the Best Paper Award of IEA/AIE-2001 in 2001, IEEE/RSJ IROS 2001 Best Paper Nomination Finalist, and Funai Information Technology Achievement Award in 2003. His books include “Internet Utilization Methods” (Iwanami Publishers, 1996), “Computational Auditory Scene Analysis” (Eds., Lawrence Erlbaum Associates, 1998) and “Advanced Lisp Technology” (eds., Taylor and Francis, 2002). He is a member of IPSJ, JSAI, JSSST, RSJ, JCSS, ACM, AAAI, IEEE, and ASA.