Separation of Individual Instrumental Sounds in Monaural Music Signals Applying a Modified Wiener Filter with a Parameter Identification Technique on the Time-frequency Plane

Yukio Fukayama and Shoya Ogawa
Graduate School of Science and Technology, Hiroshima Institute of Technology
2-1-1 Miyake, Saeki-ku, Hiroshima city 731-5193, Japan
E-mail: y.fukayama.ik@it-hiroshima.ac.jp

Abstract

A signal processor which is suitable for distilling a melody played by a particular instrument from a music ensemble is discussed in this report. The processor separates a monaural music sound into individual tones of every pitch with a particular timbre, combinations of which provide melodies played by an instrument. A mixture of strongly correlated sounds such as the same pitch tones by different instruments can be processed by the proposed technique that is a windowed Wiener filter based on current statistics of the mixture sequentially evaluated in the frequency-domain.

1 Introduction

This research seeks a signal processor for random processes which is suitable for distilling melodies played by a particular instrument from music ensemble. Sound separation techniques that refers multi-inputs from microphones at spatially different positions [e.g. 1] are well-known so far, however, a processor for monaural input has been developed here for broader and more convenient applications. In this field the Wiener filter is the most classical processor [2] to separate a signal from a monaural mixture of the signal with an additive noise. To apply the Wiener filter the second-order statistics of the signal itself and the mixture are required [3], and the requirement is difficult except for simple cases such as the signal and the noise are uncorrelated and time-invariant random processes. While the proposed processor sequentially evaluates statistics of the mixture assuming a model of music sound and employs a windowed Wiener filter so as to separate the mixture of time-variant and correlated random processes such as amplitude changing same pitch tones by different instruments.

The model of music sounds is assumed to be mixtures of standard tones weighted by amplitude factors and corrupted by a colored noise so that the proposed processor at first evaluates the current values of the factors. Here the amplitude factors are coordinates of the signal with respect to the standard tones that are played by different instruments on every pitch and form a basis. Then, according to the evaluated coordinates, the processor adaptively calculates the cross-correlations between the input signal and each standard tone which are essential for statistical evaluations of the mixture. Finally, following the least squares criterion with the evaluated statistics, the processor separate the individual tones with a pitch and timber components of the mixture, combinations of which provide melodies generated by instruments.

2 Wiener filter with window

In this section, a discrete time version of a modified Wiener filter with a window is introduced for processing random processes with time-variant second order statistics.

At first signal \( \{ U_k \} \) and noise \( \{ W_k \} \) at a discrete time \( k \) with an interval \( \Delta t \) are considered. The sampled sequences of them are not obtainable directly and the filter to separate them is assumed to be input \( \{ Y_k \} \) that is mixture of them:

\[
Y_k = U_k + W_k, \quad (1)
\]

where \( \{ U_k \} \), \( \{ W_k \} \) and \( \{ Y_k \} \) are zero-means and real-valued random processes, the autocorrelation \( r_{UU}(m;k) \), \( r_{WW}(m;k) \) and cross-correlation \( r_{UY}(m;k) \) of which are changing slowly with \( k \). In addition, the processes are assumed to be jointly wide sense stationary.
in the neighborhood \( k + m \), \( \{m \} \leq 2M \):

\[
\begin{align*}
    r_{uy}(m; k) &:= \text{E}[U_k Y_{y+m}] = r_{uy}(-m; k), \\
    r_{uu}(m; k) &:= \text{E}[U_k U_{k+m}] = r_{uu}(-m; k), \\
    r_{ru}(m; k) &:= \text{E}[U_k Y_{y+m}] = r_{ru}(-m; k).
\end{align*}
\]

(2)

(3)

(4)

Then, applying a proper window \( \{ d_m \} \) with \( d_m = 0 \), \( \{ m \} > M \), an autocorrelation matrix \( R_{yy}(k) \in \mathbb{R}^{2M+1 \times 2M+1} \) and an cross-correlation vector \( r_{uy}(k) \in \mathbb{R}^{2M+1} \) are defined:

\[
R_{yy}(k) := \text{E}[Y(k) Y^T(k)]
\]

\[
= \begin{pmatrix}
    d_y^2 r_{yy}(0; k) & \cdots & d_y d_m r_{yy}(2M; k) \\
    \vdots & \ddots & \vdots \\
    d_y d_m r_{yy}(2M; k) & \cdots & d_y^2 r_{yy}(2M; k)
\end{pmatrix},
\]

(5)

\[
r_{uy}(k) := \text{E}[Y(k) U_k]
\]

\[
= \begin{pmatrix}
    d_m r_{uy}(M; k) & \cdots & d_m r_{uy}(M-1; k) \\
    \vdots & \ddots & \vdots \\
    d_m r_{uy}(M; k) & \cdots & d_m r_{uy}(M; k)
\end{pmatrix}^T,
\]

(6)

where the input sequence \( \{ Y_k \} \) around \( k \) is multiplied by \( \{ d_m \} \) and forms an input vector \( Y(k) \in \mathbb{R}^{2M+1} \).

\[
Y(k) := (d_y Y_{y+m}) d_{M-1} Y_{y+M-1} \cdots d_m Y_{y+m})^T.
\]

(7)

Here an example of \( \{ d_m \} \) is the following Hanning window:

\[
d_m = \frac{1}{2} \left(1 + \cos \frac{2m \pi}{2M+1}\right), \{m\} \leq M.
\]

(8)

Now a filter that obtains the estimate \( \{ \hat{U}_k \} \) of signal \( \{ U_k \} \) from input \( \{ Y_k \} \) is assumed to be the following FIR form:

\[
\hat{U}_k = h^T(k) Y(k),
\]

(9)

where coefficient vector \( h(k) \in \mathbb{R}^{2M+1} \) is given by the statistical least squares criterion to be minimize the following cost \( C \) :

\[
C(h(k)) := \text{E}[(\hat{U}_k - U_k)^2]
\]

\[
= \text{E}[(\hat{h}(k) Y(k) - U_k)^2]
\]

\[
= \text{E}[h(k) - R_{yy}^{-1}(k) r_{uy}(k)]^2 R_{yy}(k)[h(k) - R_{yy}^{-1} r_{uy}(k)]
\]

\[
+ r_{uu}(0; k) - R_{uy}(0; k) R_{yy}^{-1} R_{uy}(0; k)
\]

\[
\geq r_{uu}(0; k) - R_{uy}(0; k) R_{yy}^{-1} R_{uy}(0; k). \tag{10}
\]

Since \( R_{yy}(k) \) is positive definite, the optimal vector is:

\[
h(k) = R_{yy}^{-1}(k) r_{uy}(k), \tag{11}
\]

where \( r_{uu}(m; k) \) of \( r_{uy}(k) \) in (6) and \( r_{uy}(m; k) \) of \( R_{yy}(k) \) in (5) are respectively evaluated in sections 3 and 4.

### 3 Evaluation of cross-correlation

The amplitude factor \( \sigma_j(t) \) that corresponds to the \( l \)-th standard tone \( S_j(t) \) is changing slowly with continuous time \( t \) where \( l = l(p,i) \), \( l = 1, \cdots, L \) is a unique number for a particular pitch number \( i \) played by the \( p \)-th instruments. In this section, the cross-correlation \( r_{uy}(m; k) \) for (11) is evaluated through estimation of the amplitude factors which are regarded as positive constants in a short interval of \( \tau \).

Firstly the music signal \( Y(t) \) is assumed to be a mixture of individual tones, which are products of the standard tones and the amplitude factors, and a contaminating colored noise \( V(t) \):

\[
Y(t) = \sum_{j=1}^{L} \sigma_j(t) S_j(t) + V(t), \tag{12}
\]

where \( S_j(t) \) and \( V(t) \) are zero-means random processes with the jointly wide sense stationary at large, the second order statics and power spectra of which can be evaluated previously. Here the autocorrelation \( \Phi_{yy}(\tau) \) and the power spectrum \( \Phi_{yy}(j\omega) \) form the Fourier transform pair with time difference \( \tau \), which is denoted by \( \text{FT} \rightarrow \), as are follows:

\[
\Phi_{yy}(\tau) := \text{E}[Y(t) Y(t + \tau)]
\]

\[
= \sum_{j=1}^{L} \sum_{i=1}^{L} \sigma_j(t) \sigma_i(t) \rho_{ji}(\tau) + \Phi_{yy}(\tau)
\]

\[
\text{FT} \rightarrow \Phi_{yy}(j\omega) = \sum_{j=1}^{L} \sum_{i=1}^{L} \sigma_j(t) \sigma_i(t) P_{ji}(j\omega) + \Phi_{yy}(j\omega), \tag{13}
\]

where \( \rho_{ji}(\tau) \) and \( P_{ji}(j\omega) \) are respectively a correlation and a spectrum between \( S_j(t) \) and \( S_i(t) \), \( l, j = 1,2,\cdots, L \). Also \( \Phi_{yy}(\tau) \) and \( \Phi_{yy}(j\omega) \) are those of \( V(t) \).

In addition, the following quantities and properties have been applied:

\[
\rho_{ji}(\tau) := \text{E}[S_j(t + \tau) S_i(t)] = \rho_{ji}(-\tau)
\]

\[
= \text{FT} \rightarrow P_{ji}(j\omega) = P_{ji}(-j\omega) = P_{ji}(j\omega), \tag{14}
\]

\[
\Phi_{yy}(\tau) := \text{E}[V(t + \tau) V(t)] \xrightarrow{\text{FT} \rightarrow} \Phi_{yy}(j\omega), \tag{15}
\]

\[
\text{E}[S_j(t + \tau) V(t)] = 0, \tag{16}
\]

where \( \text{Re} \) and superscript * respectively denote the real part and the conjugate of a complex number.
Then the amplitude factors are evaluated through the weighted least square method in the frequency domain. Concretely, \( \sigma(t) \) shall be identified as a non-negative value in order to let \( \Phi_{\nu Y}(j\omega_{n}t) \), \( n=1,\cdots,N \) in (13) fit for actually observed spectrum under known Re\{\( P_{\nu Y}(j\omega_{n}) \)\} and \( \Phi_{\nu Y}(j\omega_{n}) \):

\[
\Psi_{Y} = f(\sigma) + \varepsilon ,
\]

where \( \varepsilon \in \mathbb{R}^{N} \) and \( \Psi_{Y} \in \mathbb{R}^{N} \) respectively denote the fitting error vector and observed spectra vector \( \Psi_{Y} \in \mathbb{R}^{N} \) around \( t \) for which \( |\nu_{n}(t)| \) in section 4 is applicable:

\[
\Psi_{Y} = \begin{bmatrix} |\nu_{1}(t)| & \cdots & |\nu_{N}(t)| \end{bmatrix}^{T} .
\]

In addition, a function \( f(\sigma) \in \mathbb{R}^{N} \), a vector \( \sigma(t) \in \mathbb{R}^{N} \), and non-negative definite matrix \( P(j\omega_{n}) \in \mathbb{R}^{N \times N} \) are defined:

\[
f(\sigma) := \begin{bmatrix} \sigma^{T} P(j\omega_{n}) \sigma + \Phi_{\nu Y}(j\omega_{n}) \end{bmatrix} ,
\]

where \( \sigma(t) := (\sigma_{1}(t) \quad \cdots \quad \sigma_{N}(t))^{T} \),

\[
P(j\omega_{n}) := \begin{bmatrix} \text{Re}\{P_{11}(j\omega_{n})\} & \cdots & \text{Re}\{P_{1N}(j\omega_{n})\} \\
\vdots & \ddots & \vdots \\
\text{Re}\{P_{N1}(j\omega_{n})\} & \cdots & \text{Re}\{P_{NN}(j\omega_{n})\} \end{bmatrix} .
\]

Thus applying the Gauss-Newton method to obtain \( \sigma(t) = \arg\min[J(\sigma)] \), \( \sigma^{(0)} \) gives the desired \( \sigma(t) \) after having suffered for a convergence condition with a proper threshold \( \delta \) through recursive procedure \( (i = 1,2,\cdots) \) with an initial \( \sigma^{(0)} \):

\[
\sigma^{(i)} = \sigma^{(i-1)} - G(\sigma^{(i-1)})|\Phi_{\nu Y} - f(\sigma^{(i-1)})| \leq \delta ,
\]

where the following cost function \( J(\sigma) \) with positive definite weight matrix \( B \in \mathbb{R}^{N \times N} \), correction gain matrix \( G(\sigma) \in \mathbb{R}^{N \times N} \) and derivative matrix \( F(\sigma) \in \mathbb{R}^{N \times N} \) are defined:

\[
J(\sigma) = \varepsilon^{T} B \varepsilon ,
\]

\[
G(\sigma) := F(\sigma) B F(\sigma)^{T} + F(\sigma) ,
\]

\[
F(\sigma) := \frac{\partial f(\sigma)}{\partial \sigma} = 2 \begin{bmatrix} \sigma^{T} P(j\omega_{n}) \\
\vdots \\
\sigma^{T} P(j\omega_{n}) \end{bmatrix} .
\]

Concerning (22) and (23), the medium value in an applicable range of each amplitude factor is given to the corresponding component of initial \( \sigma^{(0)} \) and negative value components of \( \sigma^{(0)} \) are replaced with 0 through the recursive procedure.

Finally the cross-correlation \( \phi_{\nu Y}(t;\tau) \) between music signal and the \( n \)-th standard tone is calculated from the obtained \( \sigma(t) \) with multiplying (12) by \( S_{n}(t-\tau) \) and taking expectation:

\[
\phi_{\nu Y}(t;\tau) := \mathbb{E}\{Y(t)S_{n}(t-\tau)\} = \sum_{\nu=1}^{L} \sigma_{\nu}(t) p_{\nu Y}(t) ;
\]

thus, the following cross-correlation is applied to the section 2:

\[
r_{\nu Y}(m;\kappa) = \sigma_{\nu}(m\Delta t) \phi_{\nu Y}(m\Delta t;\kappa\Delta t),
\]

in case that the individual tone \( U_{\nu} = \sigma_{\nu}(m\Delta t) S_{n}(m\Delta t) \) is the signal to be distilled while the sum of the other tones is a noise.

### 4 Evaluation of autocorrelation

In this section, a power spectrum \( \Phi_{\nu Y}(j\omega_{n}t) \) of a random process \( Y(t) \) which contains equivalent information of autocorrelation \( \phi_{\nu Y}(t;\tau) \) is evaluated from a sample process \( y(t) \) of \( Y(t) \) through the Gabor wavelet transform [6].

Since a music sound contains the individual tones which is periodic in time, \( y(t) \) can be described as follows:

\[
y(t) = 2 \sum_{n=1}^{N} |c_{n}(t)| \cos(\omega_{n} t + \Delta c_{n}(t))
\]

\[
= \sum_{n=1}^{N} |c_{n}(t)| e^{j\omega_{n} t} + c_{n}^{*}(t)e^{-j\omega_{n} t} | \mathcal{F}^{-1}Y(j\omega) \]

\[
= \sum_{n=1}^{N} |C_{n}(j\omega - \omega_{n}) + C_{n}^{*}(-j\omega + \omega_{n})| ,
\]

where \( \Delta \) denotes argument of a complex number, and the magnitude terms in the summations form Fourier transform pairs \( c_{n}(t) \leftarrow \mathcal{F}^{-1} \rightarrow C_{n}(j\omega) \). In addition fundamental and harmonic frequencies of each individual tone are assumed to be on some sampling points at \( \omega_{n} \), \( n=1,\cdots,N \) in frequency.

Then the Gabor wavelet which has windows at \( t \) in the time-domain and at \( \omega_{n} \) in the frequency-domain is considered:

\[
g_{\nu}(t) := \frac{1}{\sqrt{\Delta t}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-T_{n})^{2}}{2\Delta t}} e^{-j\omega_{n}T_{n}}
\]

\[
\leftrightarrow \mathcal{F}^{-1} \rightarrow G_{\nu}(j\omega) = \sqrt{\Delta t} e^{-\frac{(\omega - \omega_{n})^{2}}{2\Delta t}} ,
\]

where the mother wavelet \( g(t) \), factors \( a_{\nu} \) and \( \lambda \) are applied:

\[
g(t) := \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{t^{2}}{2\lambda^{2}}} ,
\]

\[
a_{\nu} = \frac{\omega_{n}}{\omega_{s}} ,
\]
Now obtain the Gabor wavelet transform $\eta_n(t)$ that is a projection of $y(t)$ onto the time-frequency plane $(t, \omega_n)$:

$$\eta_n(t) := \left[ g_n(\tau - t)y(\tau) d\tau = g_n(t)^* y(t) \right]$$

$$\leftarrow \text{FT:} \quad G_n(j\omega)Y(j\omega)$$

$$\approx \sqrt{\mu_n} e^{j \omega_n t} C_n((\omega - \omega_n)),$$  \hspace{1cm} (33)

where operator $*$ denotes convolution and the above approximation are provided by a window effect of $G_n(j\omega)$ in frequency; furthermore, the inverse Fourier transform of the last expression gives $\Phi_{yy}(j\omega_n; t)$ smoothed by a convolution:

$$[\eta_n(t)] = \left[ g_n(t)^* c_n(t)e^{j\omega_n t} \right]$$

$$= c_n(t)^* \frac{1}{\sqrt{2\pi|\mu_n|}} e^{-j\omega_n t} \approx \Phi_{yy}(j\omega_n; t).$$  \hspace{1cm} (34)

Finally the desired autocorrelation $\phi_{yy}(\tau; t)$ that is the inverse Fourier transform of $\Phi_{yy}(j\omega_n; t)$ is obtained as:

$$\phi_{yy}(\tau; t) = \frac{1}{N} \sum_{n=1}^{N} \psi_n(\tau) \cos\omega_n \tau;$$  \hspace{1cm} (35)

thus, the following autocorrelation is applied to section 2:

$$r_{yy}(m; k) = \phi_{yy}(m\Delta t; k\Delta t).$$  \hspace{1cm} (36)

5 Example of the processing

Performance of the foregoing signal processing method has been studied through the cross-validations where sounds for obtaining a priori statistics and mixtures to be separated are from different sources. Concretely examples in this section, the inputs to be separated are generated by ROLAND VXC MIDI system which is different from YAMAHA’s system for a priori statistics of every pitch and instrument. Listening to the separated tones including the same pitch case, the sufficient performance are evaluated. In addition, processing time of the proposed method is almost one tenth of the prior art [4].

Here MIDI (Musical Instrument Digital Interface) standard [9] assigns pitch number $n \in \mathbb{N}$ to frequency $f_n = f_0 2^{n/12}$ with $f_0 = 8.18$ [Hz] where the chromatic scale of equal temperament is based on the standard pitch in 1939 [10]. For instance, a piano has pitches in $21 \leq n \leq 108$ in which $f_{108} = 440.0$ [Hz] is the A tone directly above the middle C at $n = 60$. Then spectra in this section are projected by Gabor wavelet transform on the time-frequency plane where horizontal and vertical axes are respectively with time in second and the pitch numbers.

The first example is separation of a mixture of a continuous tone at pitch number $n = 81$ by a trumpet and interrupted tones, which consists of notes and rests alternately, at $n = 73$ by a violin. Spectra of the trumpet, the violin and mixture of them are shown in Fig.1, Fig.2 and Fig.3 in order. Having applied the proposed method to the mixture, Fig.4 and Fig.5 are spectra of separated tones of the trumpet and violin respectively where they seem to be sufficient estimations of Fig.1 and Fig.2. In addition the separated tones sound well comparing to the original tones.
The second example is separation of a mixture of a decreasing tone at $n=60$ by a trumpet and increasing tone at the same pitch by a violin. Spectra of the trumpet, the violin and mixture of them are shown in Fig.6, Fig.7 and Fig.8 in order. Having applied the proposed method, Fig.9 and Fig.10 are spectra of separated tones of trumpet and violin respectively where they seem to be sufficient estimations of Fig.6 and Fig.7. In addition the separated tones sound well comparing to the original tones.

6 Conclusions

A signal processor with the following features suitable for separating tones of a music ensemble, the combination of which provides melodies played by an instrument, is proposed:

1) Based on a priori knowledge of autocorrelations of and cross-correlations among standard tones that are played by different instruments on every pitch.

2) Identifying amplitude factors as coordinates of current music signal with respect to the standard tones that form a basis.

3) Evaluating cross-correlations between the current signal and each standard tone from the identified coordinates.

4) Estimating particular pitch and timbre components of the signal under the least squares criterion in the time domain applying the identified coordinates and the a priori statistics.

References


