Separation of Individual Instrumental Tones in Monaural Music Signals Applying a Modified Wiener Filter and the Gabor Wavelet Transform

Yukio Fukayama
Graduate School of Science and Technology, Hiroshima Institute of Technology
2-1-1 Miyake, Saeki-ku, Hiroshima city 731-5193, JAPAN
Email: fukayama@ieee.org

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. It is a modified Wiener filter for the mixture, the current statistics of which are sequentially evaluated in the chromatic scale with the Gabor wavelet transform.

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 so far well-known, however, the 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 it employs a modified Wiener filter so as to separate the mixture of time-variant and correlated random processes such as amplitude changing tones and same pitch tones by different instruments.

The proposed processor that is a modified Wiener filter combined with evaluation of current statistics in the chromatic scale [10] with the Gabor wavelet transform is intended to clear a long processing time problem of a same-purpose approach by us [4]. Our former approach evaluates the current statistics only in the time-domain in which heavy calculations for eigenvalues and eigenvectors of time-variant correlation matrices are required every time step. While, in the proposed approach, eigenvalues and eigenvectors of any correlation matrices are always power spectra and the sinusoidal functions in the frequency-domain. Unlike the classical Wiener filter, the processor in this report is allowed to use non-causal filter combined with input data buffer which is convenient for approaches in the frequency domain.

The model of music sounds is assumed to be a weighted mixture of each standard tone, which is a typical tone of every pitch played by a particular instrument with invariant power, where the weight is called the amplitude factor corresponding to the standard tone. Thus, the proposed processor, at first, evaluates the current amplitude factors which are the coordinates of the mixture where the standard tones form oblique basis. Then, according to evaluated coordinates of the mixture, its current autocorrelation and cross-correlation to each standard tone are calculated for obtaining the modified Wiener filter. Finally, the filter separates the mixture into individual tone, which is a component of a particular pitch played by an instrument with variant power, and combination of which provides a melody.

2. Modified Wiener filter
In this section a modified Wiener filter which deals with random processes with time-variant second order statistics is introduced in the discrete time domain. Firstly consider signal \{U(k)\} and noise \{W(k)\} at discrete time \(k\) with the sampling interval \(\Delta t\), and the mixture \{Y(k)\} of them is input to the filter:

\[
Y(k) = U(k) + W(k),
\]

Proceedings of the 50th ISCIE International Symposium on Stochastic Systems Theory and Its Applications
Kyoto, Nov. 1-2, 2018

– 153 –
where \( \{U(k)\} \), \( \{W(k)\} \) and \( \{Y(k)\} \) are zero-means and real-valued random processes. The autocorrelation \( r_{YY}^k(m) \), \( r_{UU}^k(m) \) and cross-correlation \( r_{UY}^k(m) \) of them, superscripts and subscripts of which respectively indicate variables and time, are changing slowly with \( k \). Thus the processes can approximately be regarded as the jointly wide sense stationary in the \( M \)-neighborhood of \( k \) or discrete time at \( k + m \), \( |m| \leq 2M \):

\[
\begin{align*}
\hat{r}_{YY}^k(m) &:= E[Y(k+m)Y(k)] = r_{YY}^k(-m) , \\
\hat{r}_{UU}^k(m) &:= E[U(k+m)U(k)] = r_{UU}^k(-m) , \\
\hat{r}_{UY}^k(m) &:= E[Y(k+m)U(k)] = r_{UY}^k(-m) .
\end{align*}
\]

Then, an autocorrelation matrix \( R_{YY}^k = (\hat{r}_{YY}^k)^T \in \mathbb{R}^{2M+1 \times 2M+1} \) and a cross-correlation vector \( R_{UY}^k \in \mathbb{R}^{2M+1} \) around \( k \) are defined as:

\[
R_{YY}^k := E[Y^k(Y^k)^T] =
\begin{pmatrix}
r_{YY}^k(0) & r_{YY}^k(1) & \cdots & r_{YY}^k(2M) \\
r_{YY}^k(-1) & r_{YY}^k(0) & \cdots & r_{YY}^k(2M-1) \\
\vdots & \vdots & \ddots & \vdots \\
r_{YY}^k(-2M) & r_{YY}^k(-2M+1) & \cdots & r_{YY}^k(0)
\end{pmatrix},
\]

(5)

where the input sequence \( \{Y(k)\} \) around \( k \) forms an input vector \( Y^k \in \mathbb{R}^{2M+1} \) as follows:

\[
Y^k := (Y(M) \ Y(M-1) \ \cdots \ Y(-M))^T.
\]

Now consider a discrete-time Winer filter with a rectangle window that obtains the estimate \( \hat{U}(k) \) of signal \( \{U(k)\} \) from input \( \{Y(k)\} \) in the following FIR filter form:

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

(8)

where coefficient vector \( h^k \in \mathbb{R}^{2M+1} \) is given by the statistical least squares criterion to be minimize the following cost function \( C \) of the estimating error under positive definite \( R_{YY}^k \):

\[
C(h^k) := E[(\hat{U}(k) - U(k))^T]
= E[(h^k)^T Y^k - (h^k)^T Y^k - U(k))^T]
= (h^k)^T R_{YY}^k h^k - (h^k)^T r_{UU}^k - (r_{UY}^k)^T h^k + r_{UU}^k (0)
= |h^k - (R_{YY}^k)^{-1} r_{UY}^k|^T R_{YY}^k (h^k - (R_{YY}^k)^{-1} r_{UY}^k)
+ |r_{UU}^k (0) - (r_{UY}^k)^T R_{YY}^k r_{UU}^k + (h^k - (R_{YY}^k)^{-1} r_{UY}^k)^T R_{YY}^k (h^k - (R_{YY}^k)^{-1} r_{UY}^k)|.
\]

The equality condition above gives the optimal coefficient the seeking filter is:

\[
h^k = (R_{YY}^k)^{-1} r_{UY}^k ,
\]

(10)

where \( r_{UY}^k \) in (6) and \( r_{YY}^k \) in (5) are respectively evaluated in sections 3.

3. Evaluation of correlations

In this section current autocorrelation and cross-correlation to each standard tone are obtained applying the model of music sounds. At first, a music sound \( Y(k+m) \) of the \( M \)-neighborhood of \( k \) is assumed to be a mixture of individual tones that are products of the \( l \)-th standard tones \( S_l(k+m) \) and the amplitude factors \( \sigma^l_z \), \( l = 1,\ldots,L \):

\[
Y(k+m) = \sum_{l=1}^{L} \sigma^l_z S_l(k+m) + V(k+m) ,
\]

(11)

where a particular pitch number \( i \) played by the \( p \)-th instruments is given a unique number \( l = l(i,p) \). In addition, the \( l \)-th amplitude factor which is changing slowly can be regarded as a positive constant in the neighborhood and a contaminating colored noise \( V(k+m) \) is independent of the standard tones. Also, assuming that \( S_l(k+m) \) and \( V(k+m) \) are zero-means random processes with the jointly wide sense stationary, the second order statics and power spectra of which can be evaluated previously.

Here the autocorrelation \( r_{YY}^k(m) \) and the power spectrum \( \Phi_{YY}^k(\theta) \) in the \( M \)-neighborhood of \( k \) form the Discrete time Fourier transform pair with sampling interval \( \Delta t \) and normalized angle frequency \( \theta = \omega \Delta t \), which is denoted by \( \triangleleft \Phi_{DTFT} \rightarrow \), as are follows:

\[
\begin{align*}
\hat{r}_{YY}^k(m) &:= \sum_{l=1}^{L} \sigma^l_z^2 r_{YY}^k(m) + r_{UY}^k(m) \\
&\triangleleft \Phi_{DTFT} \rightarrow \Phi_{YY}^k(\theta) = \Phi_{YY}^k(-\theta) \\
&= \sum_{l=1}^{L} \sigma^l_z^2 \Phi_{YY}^k(\theta) + \sum_{l=1}^{L} \sigma^l_z^2 \Phi_{YY}^k(\theta + \phi) + \Phi_{YY}^k(\theta) \\
&\triangleleft \Phi_{DTFT} \rightarrow \Phi_{YY}^k(\theta) \\
&= \sum_{l=1}^{L} \sigma^l_z \Phi_{YY}^k(\theta) + \Phi_{YY}^k(\theta) ,
\end{align*}
\]

(12)

where \( \rho^l_z(m) \) and \( \Phi_{YY}^k(\theta) \) are respectively a correlation and a spectrum between \( S_l(k+m) \) and \( S_j(k+m) \), \( l, j = 1,2,\ldots,L \). Also \( r_{YY}^k(m) \) and \( \Phi_{YY}^k(\theta) \) are those of \( V(k+m) \). In addition, the following quantities and properties have been applied to (12):

\[
\begin{align*}
\rho^l_z(m) &:= E[S_l(k+k)S_j(k)] = \rho^l_z(-m) \\
&\triangleleft \Phi_{DTFT} \rightarrow \rho^l_z(\theta) = \Phi_{YY}^k(\theta) ,
\end{align*}
\]

(13)

\[
\begin{align*}
\hat{r}_{YY}^k(m) &:= E[V(k+m)V^*(k)] \triangleleft \Phi_{DTFT} \rightarrow \Phi_{YY}^k(\theta) ,
E[S_l(k+m)V^*(k)] = 0 ,
\end{align*}
\]

(14)

(15)

where 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 \( \sigma^i \) shall be identified as a non-negative value in order to let \( \Phi'_{\nu}(\theta) \) in (12) fit for actually observed spectrum under previously evaluated \( P_{\nu}(0) \) and \( \Phi_{\nu'}(\theta) \):

\[
\Psi^k = f(\sigma^k) + \varepsilon^k,
\]

where \( \varepsilon^k \in \mathbb{R}^N \) and \( \Psi^k \in \mathbb{R}^N \) respectively denote the fitting error vector and observed spectra vector around \( k \) for which \( \Psi^k_n(\theta_n) \) at pitch number \( n = 1, \ldots, N \) in section 4 is applicable:

\[
\Psi^i := \left( \Psi^i_1(0) \ldots \Psi^i_N(0) \right)^T.
\]

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 \( f_{40} = 440.0 \) [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.

Furthermore a function \( f(\sigma^k) \in \mathbb{R}^N \), a vector with non-negative components \( \sigma^i \in \mathbb{R}^N \), and non-negative definite matrix \( P(0) \in \mathbb{R}^{N \times N} \) are defined:

\[
f(\sigma) := \left( (\sigma^i)P(0)\sigma^i + \Phi_{\nu'}(\theta_i) \right), \quad \sigma^i := (\sigma^i)^T, \quad P(0) := \left( \begin{array}{c} P_{II}(0) \quad \vdots \quad P_{LL}(0) \end{array} \right),
\]

Thus applying the Gauss-Newton method to obtain \( \sigma^i = \arg\min[J^i(\sigma)] \), \( \sigma_i \) gives the desired \( \sigma^k \) after having sufficed for a convergence condition with a proper threshold \( \delta \) through recursive procedure, the iteration counts of which is shown in subscript \( i \) and the initial condition of which is \( \sigma_0 \), as follows:

\[
\sigma_i = \sigma_{i-1} - G(\sigma_{i-1})[\Psi^k - f(\sigma_{i-1})],
\]

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

\[
J^i(\sigma) := (\varepsilon^k)^T\mathbf{B}\varepsilon^k,
\]

\[
G(\sigma) := [F^T(\sigma)B(\sigma)]^{-1}F^T(\sigma)B(\sigma),
\]

\[
F(\sigma) := \frac{\partial f(\sigma)}{\partial \sigma^i} = \left( \begin{array}{c} \sigma_i^T\mathbf{P}(\theta_i) \\ \vdots \\ \sigma_N^T\mathbf{P}(\theta_N) \end{array} \right).
\]

Concerning (21) and (22), the middle 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.

Therefore the cross-correlation \( \rho^k_{L} \) between music signal and the \( L \)-th standard tone is calculated from the obtained \( \sigma^k \) with multiplying (11) by \( S_j(k) \) and taking expectation:

\[
\rho^k_{L}(m) := \mathbb{E}\{Y(k+m)S_j(k)\} = \sum_{j=1}^{N} \sigma^j_{L} \rho_j(m);
\]

thus, the following cross-correlation is obtained for (10):

\[
r^k_{L}(m) = \sigma^j_{L} \rho_j(m),
\]

where the filter in (8) estimates the individual tone \( U(k) = \sigma^j_{L}S_j(k) \) as the signal to be distilled and the sum of the other tones is regarded as a noise \( W(k) \) in (1).

Finally \( r^i_{L}(m) \) in (5) is also obtainable as follows:

\[
r^i_{L}(m) = \sum_{j=1}^{N} \sum_{l=1}^{N} \sigma^i_{L} \sigma^j_{L} \rho_j(m) + r_{LL}(m).
\]

4. Applying Gabor Wavelet Transform

Power spectrum \( \Phi'_{\nu'}(\theta) \) of a random process \( Y(k) \) is evaluated applying the Gabor wavelet transform [6] in this section.

At first, a window \( \gamma_n(m) \), \( (n = 0, 1, \ldots, N) \) and its discrete-time Fourier transform \( \Gamma_{\nu}(e^{i\theta}) \) with normalized angle frequency \( \theta = \alpha \Delta t \) are considered as follows:

\[
\gamma_n(m) := \frac{1}{\sqrt{d_n}} \gamma_n(m) = \frac{1}{\sqrt{2\pi d_n}} e^{-\frac{1}{2} \frac{(\alpha \Delta t)^2}{d_n}}.
\]
where energy of the window \( y(n) \) is approximated as the same value for all \( n \):

\[
r_{r_0, \tau_0}(0) = \frac{1}{\sqrt{2\pi\Delta}} \int_{-\infty}^{\infty} y(t) e^{-\frac{i\sigma_0 t^2}{2\Delta}} dt.
\]  

Assuming sufficiently small \( \Delta t \), truncation \( M > 3/\Delta t \), and \( 0 < a_n < 1 \), autocorrelation \( r_{y_0}(m) \) and power spectrum of the window \( \gamma_y(m) \) are written as follows:

\[
r_{y_0}(m) := \sum_{n = -M}^{M} y(n+l) y(n) = \gamma_y(m) + \gamma_y(-m)
\]

\[
\leftrightarrow \text{DTFT} \rightarrow \gamma_y(e^{\phi}) \text{DTFT} \rightarrow \gamma_y(e^{\phi}) = \left| \gamma_y(e^{\phi}) \right|^2.
\]

where energy \( r_{y_0}(0) \) of the window is approximated as

\[
r_{y_0}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \gamma_y(e^{\phi}) \right|^2 d\phi = \frac{2\pi \Delta}{\Delta - a_n \Delta} e^{-\frac{i\sigma_0}{\Delta}} \int_{-\infty}^{\infty} \frac{e^{\frac{i\sigma_0 t^2}{2\Delta}}}{\sqrt{2\pi\Delta}} dt.
\]  

Then consider the following Gabor wavelet \( g_m(\omega) \) at \( \theta_m \), to which scaling factor \( a_n = \theta_n/\theta_m \) is applied:

\[
g_m(\omega) := \gamma_y(m)e^{j\theta_m} = \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{i\sigma_0 (\omega - \theta_m)^2}{2\Delta}}, \quad \theta_m < 0 < \theta_0 < \cdots < \theta_n < \pi
\]

where \( 0 < \theta_n < \theta_0 \) is assumed to prevent from aliasing.

Considering (12) and (31), the Gabor wavelet transform \( \Psi^*_n(\theta_m) \), \( n = 1, \cdots, N \) of a random process \( Y(k) \) provides its smoothed power spectrum which is convolution operation \( \ast \) between a window \( \left| \gamma_y(e^{\phi}) \right|^2 \) and \( \Phi^*_n(\theta_m) \) in the domain of normalized angle frequency \( \theta \) as follows:

\[
\Psi^*_n(\theta_m) := E \left\{ \sum_{m = -M}^{M} Y(k+m) \gamma_y(m) e^{-j\omega_0 m} \right\}
\]

\[
= \left\{ \sum_{m = -M}^{M} Y(k+m) \gamma_y(m) e^{-j\omega_0 m} \right\} \sum_{l = -M}^{M} \gamma_y(l) e^{j\omega_0 l}
\]

\[
= \sum_{m = -M}^{M} \sum_{l = -M}^{M} Y(k+m) \gamma_y(m) \gamma_y(l) e^{j\omega_0 (m+l)}
\]

\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \Phi^*_n(\theta_m) \right|^2 \left| \gamma_y(e^{\phi}) \right|^2 d\phi.
\]

where \( \Psi^*_n(\theta_m) \) consists of above \( \Psi^*_n(\theta_m) \).

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. In this section, for example, 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 former approach [4].

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 typically played on every pitch by different instruments.

2) Assuming a music sound model and identifying amplitude factors which are coordinates of current music sound with respect to the standard tones.

3) Evaluating autocorrelation of current sound and cross-correlations between the sound and each standard tone from the identified coordinates applying the Gabor wavelet transform under the least squares criterion.

4) Estimating individual tone of particular pitch and timbre from the music sound in the time domain applying a modified Wiener filter with the identified autocorrelation and cross-correlations of current music sound.

References

[1] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki: Combined method of array processing and independent component analysis for blind separation of acoustic signals, IEEE transactions on speech and audio processing, Vol.11, No.3, pp.204-215, 2003.

[2] H. L. Van Trees: Detection, Estimation, and Modulation Theory, John Wiley and Sons, New York NY, 1968.

[3] N. Wiener: Extrapolation, Interpolation, and Smoothing of Stationary Time Series, John Wiley and Sons, New York NY, 1949.

[4] Y. Fukayama, D. Tanaka, and T. Kataoka: Separation of Individual Instrumental Sounds in Monaural Music Signals, IJICIC, Vol.8, No.1, pp.2275-2283, 2012.

[5] R. A. Roberts and C. T. Mullis: Digital signal processing, Addison Wesley, Reading MA, 1987.

[6] I. Daubechies: Ten lectures on Wavelets, SIAM, Philadelphia PA, 1992.

[7] A. Papoulis: Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York NY, 1984.

[8] J. B. Thomas: An Introduction to Communication Theory and Systems, Springer-Verlag, New York NY, 1988.

[9] MIDI manufacturers association: The technology of MIDI, http://www.midi.org/aboutmidi/tutorials.php.

[10] M. Kennedy: Oxford dictionary of music, Oxford, London UK, 1985.

[11] Y. Fukayama, O. Fukayama, and S. Fukayama: An Introduction to Digital Signal Processing with Exercises on Excel, Corona, Tokyo JP, 2013.