SOUND SOURCE SEPARATION OF TRIO USING STEREO MUSIC SOUND SIGNAL BASED ON INDEPENDENT COMPONENT ANALYSIS

Satoru Morita and Yasuhito Nanri

Faculty of Engineering, Yamaguchi University
2-16-1 Tokiwadai, Ube, Japan

ABSTRACT

It is necessary that the number of the observation signals equals to the number of source signals, if independent component analysis is used to perform the sound source separation. It is difficult to perform sound source separation from a stereo music sound signal when the number of sound sources are more than two. We propose the technique to perform sound source separation from a stereo music sound signal that the number of sound sources is more than two using the frequency analysis and independent component analysis.

1. INTRODUCTION

Recently, some techniques of the sound source separation are proposed. Those techniques are classified into two. One is the sound source separation method using many microphones [1][2] [3][4], another one is the sound source separation method using two microphones[5][6][7][8][9][10][11]. As the music that we contact with television or a compact disc is recorded with monaural or stereo recording, the technique using two microphones is useful in the case of separating music sound sources. Recently independent component analysis is proposed to perform sound source separation[1]. Moreover, the method separating sound sources in the frequency domain with independent component analysis is proposed[4]. In general, it is necessary that the number of the observation signals equals to the number of sound sources. In the case of separating sound source based on independent component analysis with a stereo music sound signal, we can perform sound source separation when the number of sound source is two, but cannot perform sound source separation when the number of sound sources is more than two. On the other hand, the method separating sound source using the ration between right and left power spectrum is proposed[5][6]. The purpose of the study is not the sound estimation but the note estimation. In this paper, we propose sound source separation based on independent component analysis and the histogram of the ratio between right and left power spectrum in the frequency domain.

We perform sound source separation with a stereo music sound signal in each frequency, as We pay attention to the thing that the signal in a frequency tends to contain at most whether a single signal or two signals, even if many sound sources are mixed, We use the ratio between right and left observation power spectrum in a frequency, when a sound signal exist in a frequency. Even if many sound signals are mixed, a sound signal or two sound signals are included in a frequency. At first, we get plural local maxima from the histogram of the ratio between right and left power spectrum. As the possibility that the sound signal is not mixed is high except neighbor local maxima of the histogram, we estimate two sound sources using independent component analysis estimating a 2×2 mixing matrix. If two sound sources of the separation result have near two maxima, the possibility that two sound sources are mixed in the frequency is high. If not, we estimate a $m \times 2$ mixing matrix using the histogram gotten from many 2×2 matrixes in the frequency. We show the effectiveness to apply the method for trio in the “Spring” movement of Vivaldi’s The Four Seasons.

2. SOUND SOURCE SEPARATION OF STEREO MUSIC SOUND SIGNAL

2.1. Sound source separation by the ratio and left ratio of power spectrum

A stereo music sound signal is sampled with a frequency of 44.1kHz and is quantized by sixteen bits. We translate the observation signals to the Fourier space using FFT of 2^{16} bits for 256 sound corresponding to sixteen notes. We define the position $h(t, f)$ at the time t in the frequency f as

$$h(t,f) = \left(\begin{array}{ll} \frac{p_R(t,f)}{p_L(t,f)} & (p_R(t,f) <= p_L(t,f)) \\ 2 - \frac{p_R(t,f)}{p_L(t,f)} & (p_R(t,f) > p_L(t,f)) \end{array} \right. \quad (1)$$

where $p_L(t, f)$ and $p_R(t, f)$ is the left and right power spectrum in the frequency $f(Hz)$ at time $t(s)$ respectively. The ratio $h(t, f)$ is near 0, if the position of a sound source is left. The ratio $h(t, f)$ is near 1, if the position of a sound source is middle. The ratio $h(t, f)$ is near 2, if the position of a sound source is right. We make a histogram by voting every 0.025 from 0 to 2 for 256 sound corresponding to sixteen notes. If the number of instruments is m, we extract local

1-4244-0367-7/06/$20.00 ©2006 IEEE

maxima of m individual. Local maxima of m individual are defined as $pk_0(t)$,$pk_1(t)$,$\cdots$, $pk_{m-1}(t)$. If $h(t,f)$ is larger than $pk_i(t) - width$, and smaller than $pk_i(t) + width$, we define it as ith sound in the f frequency at time t . If not, as the possibility that a number of the mixed sound is more than two is high, we select the following process.

2.2. Sound source separation by independent component analysis estimating a 2×2 mixing matrix

The sound source separation can be performed to get two sound sources from two observation signals in the case of using independent component analysis for 256 sound corresponding to sixteen notes.

In this paper, we use ICA algorithm by maximum likelihood estimation[1]. I do centration of data x of 256 sound of each frequency and move the mean to zero. Correlation matrix $C = E\{xx^T\}$[1]is calculated.

We initialize the separating matrix B by a random number. As the mixing matrix corresponds to the inverse matrix of B, the mixing matrix is initialized by the random numbers. After that we use the inverse matrix of the mixing matrix as separating matrix. At first, the decorrelation and normalization can be done as

$$B \leftarrow (BCB^T)^{-\frac{1}{2}}B. \tag{2}$$

After that, the separation matrix is updated by

$$B \leftarrow B + diag(\alpha_i)[diag(\beta_i) + E\{g(y)y^T\}]B \tag{3}$$

where $g(y) = tahn(y)$,

$$y = Bx, \tag{4}$$

$$\beta_i = -E\{y_i \cdot g(y_i)\}(i = 0, 1), \tag{5}$$

$$\alpha_i = \frac{-1}{\beta_i + E\{g'(y_i)\}}(i = 0, 1). \tag{6}$$

It is repeated until B converges if B doesn't converge.

The mixing matrix A can be estimated from the inverse matrix of the estimated separation matrix as

$$A = B^{-1}. \tag{7}$$

Supposing the observation signal x, the sound source s and the mixing matrix A, the following is formed as

$$x = As \tag{8}$$

where

$$A = \left[\begin{array}{cc} a_{00} & a_{10} \\ a_{01} & a_{11} \end{array} \right] \tag{9}$$

.

The ratio between right and left observation signals containing separated sound sources can be calculated from the estimated matrix $A(t,f)$. The sound that the separated sound includes in the right and left observation signal is represented as $a_{i0}(t,f)$ and $a_{i1}(t,f)$ which is component of the mixing matrix $A(t,f)$ respectively. The ratio between right and left observation signal for a separated sound source can be calculated as we suppose the number of sound sources to be 2.

The position $h'_i(t,f)$ for the ith separation signal defined from the estimated mixing matrix at time t in the frequency f is defined as

$$h'_i(t,f) = \left(\begin{array}{ll} \frac{a_{i0}(t,f)}{a_{i1}(t,f)} & (a_{i0}(t,f) <= a_{i1}(t,f)) \\ 2 - \frac{a_{i0}(t,f)}{a_{i1}(t,f)} & (a_{i0}(t,f) > a_{i1}(t,f)) \end{array} \right.$$

$$(i = 0, 1). \tag{10}$$

We make a histogram by voting the position $h'_i(t,f)$ every 0.025 from zero to 2. We extract local maxima of m individual from the histgram when a number of the instrument is m. We represent local maxima of m individual as $pk'_0(t)$, $pk'_1(t)$, $\cdots$, $pk'_{m-1}(t)$.

Moreover, the performance improvement can be expected by combining it the method by the histogram of the ratio between right and left power spectrum more. Two sound sources can be gotten using independent component analysis estimating a 2×2 mixing matrix. If the position $h'_i(t,f)$ is larger than $pk'_j(t) - width$ and smaller than $pk'_j(t) + width$, we define it as jth sound in the f frequency at time t . In this experiment, we use $width = 0.1$. If $h'_i(t,f)$ in $i = 0$ is different from $h'_i(t,f)$ in $i = 1$, we judge that two sound sources are mixed. If not, as the possibility that more than two sounds are mixed is high, we select the following process. We calculate the mean both $a_{i0}(t,f)$ and $a_{i0}(t,f)$ for the local maxima $pk_0(t)$,$pk_1(t)$,$\cdots$, $pk_{m-1}(t)$ of m individual derived from the histogram of the ratio between left observation signal $a_{i1}(t,f)$ and right observation signal $a_{i0}(t,f)$ containing ith sound source. As a result, we can get both $a'_{j0}(t,f)$ and $a'_{j1}(t,f)$ for which is the jth position for sound sources of m individual .

2.3. Sound source separation by a $m \times 2$ mixing matrix

From a histogram of $h'_i(t,f)$ provided by independent component analyses estimating a 2×2 mixing matrix, I estimate a $m \times 2$ mixing matrix. Every each sound source after separation extracts a 1×2 mixing matrix from a 2×2 mixing matrix, and a $m \times 2$ mixing matrix $A'(t,f)$ is found by averaging it. The $m \times 2$ mixing matrix $A'(t,f)$ is composed of the elements $a'_{ij}(t,f), (i = 0, 1, \cdots, m-1), (j = 0, 1)$. Sound sources of m individual can be derived from two observation signals by a generalization inverse matrix as the mixing matrix in each position is estimate. If a $m \times 2$ mixing matrix A' is estimated, sound sources of m individual can be gotten from two observation signal as

$$y = A's \tag{11}$$

$$A'^{-1} = A'^T(A'A'^T)^{-1} \tag{12}$$

$$s = A'^{-1}y \tag{13}$$

where s is sound sources and y is observation signals. In the case separating sound sources using only general inverse matrix, sound sources can be estimate by multiplying the estimated general inverse matrix to observation signal. Moreover, improvement in the performance can be expected by combining it the mothods based on the histogram of the ratio between right and left power spectrum and the histogram generated from a 2×2 mixing ! ! matrix estimated using independent component analysis more. In other words, we use the method using histogram generated from the ratio between right and left power spectrum if a sound exists in the frequency. We use the method using histogram generated from the estimated a 2×2 mixing matrix, if two sounds exist in the frequency. If not, we use a $m \times 2$ mixing matrix. We can get the coefficients after the Fourier transform. Sound sources of m individual can be estimated by inverse Fourier transform.

2.4. Reconstruction of sound sources

We generate sound sources from the real and imaginary parts in all frequencies space for sound sources of m individual using inverse Fourier transform.

2.5. Evaluating the results of sound source separation

We evaluate the separation effect for each instrument by comparing the original sound that a fixed place is not defined with the sound derived from separation resolut as

$$E_j = \frac{\Sigma_{t=0}^{tmax-1} \frac{|as_j(t)-ak_j(t)|}{asmax_j} * 100.0}{tmax} \tag{14}$$

where the number of sound sources is m, the total number of sound during evaluating time is $tmax$, the volume of the jth original sound source at time t before positioning is $as_j(t)$ and the maximum volume of the jth original sound source during the evaluating time before positioning is $asmax_j$. We calculate whole separating effect by getting the mean after the evaluation of all sound as

$$E = \frac{\Sigma_{j=0}^{m-1} \Sigma_{t=0}^{tmax-1} \frac{|as_j(t)-ak_j(t)|}{asmax_j} * 100.0}{tmax * m}. \tag{15}$$

2.6. The flow for sound source separation of a stereo music sound signal

- We translate the observation signal to the Fourier space using FFT of 2^{16} bits for 256 sound corresponding to sixteen notes.
- We estimate local maxima $pk_j(t)$ of m individual by generating histogram of the ratio $h(f,t)$between right and left power spectrum. if $h(f,t)$ exists near loacal maxima $pk_j(t)$ of m individual, we define that a sound source exists in the frequency f at time t.
- We estimate a 2×2 mixing matrix using fast independent component analysis from a stereo music sound signal supposing two sound sources. We make a histogram of the ratio $h'_i(f,t)(i = 0,1)$ between right and left observed signals includes each sound source gotten from the estimated 2×2 mixing matrix. We get local ratio $h'_i(t)$ in $i = 0$ and $i = 1$ near local maxima $pk_j(t)$ of histogram, and the ratio $h'_i(t)$ in $i = 0$ is different from the ratio $h'_i(t)$ in $i = 1$.
- After we estimate the $m \times 2$ mixing matrix $A'(t,f)$,we can get sound sources using generalized inverse matrix if the sound source mixed with another sound source in a frequency.
- We estimate real and imaginary part in the frequency domain by the estimated frequency space. We generate sound sources of m individual in the estimated frequency space using the inverse Fourier transform.

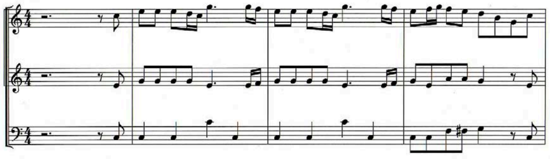
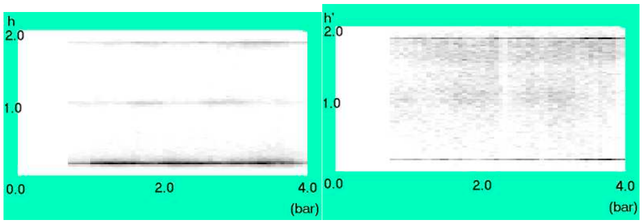


Fig. 1. The first 4 bar in the "Spring" movement of Vivaldi's The Four Seasons.

3. SOUND SOURCE SEPARATION OF TRIO

Figure 1 shows the first 4 bar in the "Spring" movement of Vivaldi's The Four Seasons. The top, middle and bottom of figure 1 corresponding to the music for the violin, cello and contrabass respectively. A stereophonic signal is generated by MIDI so that violin, cello and contrabass might be positioned at left, middle and right respectively. We perform sound source separation using the proposed method from a stereo music sound signal.

Figure 2(a) shows the histogram of the ratio $h(t,f)$ between the right and left power spectrum for the first four bar in the "Spring" movement of Vivaldi's The Four Seasons. Figure 2(b) shows the histogram of the ratio $h'_i(t,f)(i = 0,1)$ of the observed right and left signal included in sound sources derived from a 2×2 mixing matrix. The horizontal axis is bar and the vertical axis is $h(t,f)$,$h'_i(t,f)(i = 0,1)$, and the frequency is represented by gray level. If the voting number is many, the color is near black. in Figure 2(a) and Figure 2(b). Table 1 shows the evaluation using equation (14) and equation (15) in the first 4 bar in the "Spring" movement of Vivaldi's The Four Seasons. The separation performance is 66% in the method using the histogram of the ratio between right and left power spectrum in the frequency domain as shown in Table



(a) (b)

Fig. 2. (a) The histogram of the ratio $h(t,f)$ of the power spectrum for the first four notes in the "Spring" movement of Vivaldi's The Four Seasons. (b) The ratio $h'_i(t,f)(i = 0,1)$ between observation signals including a sound source estimated from the 2×2 mixed matrix for the first four notes in the "Spring" movement of Vivaldi's The Four Seasons.

1(I), and it is 83 % in the method using the independent component analysis estimating a 2×2 mixing matrix as shown in Table 1(II) and it is 86 % in the method using the generalized inverse matrix of the estimated $m \times 2$ mixing matrix as shown in Table 1(III). The separation performance is 91% in the method using three methods as shown in Table 1(IV). It is understood that performance improves in comparison with the technique so far. We used $C1 = 0.675, C2 = 1.375$ as the boundary to classify to each sound source from the histogram.

Table 1. The evaluation of the sound source separation performance for the first four notes in the "Spring" movement of Vivaldi's The Four Seasons.

	violin	cello	contrabass	average
I	59.7(%)	81.9(%)	91.6(%)	77.7(%)
II	85.1(%)	85.4(%)	79.4(%)	83.3(%)
III	88.6(%)	78.9(%)	91.1(%)	86.2(%)
IV	83.1(%)	94.3(%)	97.2(%)	91.2(%)

The instrument of the low-pitched tone is gotten well in the method I using histogram from the ratio between right and left power spectrum. The performance tends to go down when the instrument of the high-pitched tone is used. This cause is so that a low tone may contain many harmonics and the frequency that a sound source not mix to the other sound source is much. The separation performance is improved for the instrument of high tone but is not improved for the instrument of low tone in the method by independent component analysis estimating a 2×2 mixing matrix and the generalized inverse matrix after estimating a $m \times 2$ mixing matrix. The separation performance is improved for the instrument of the low tone but does not tend to come down for the instrument of the high tone and tend to be improved for the instrument of the middle tone in the method using the three methods. It is found that the performance improves in the whole.

4. CONCLUSIONS

We proposed the method to perform sound source separation from a stereo music sound signal for trio that the number of the sound source is more than two. The performance is improved by estimating a $m \times 2$ mixing matrix using the histogram of the ratio between right and left power spectrum in the frequency domain and the histogram of the ratio between right and left observation signal derived from the 2×2 mixing matrix estimated using the independent component analysis. We confirm that the performance is improved in comparison with the method so far. In this method, sound sources can be gotten even if the number of the sound source is more than three.

5. REFERENCES

[1] J. Karhunen A. Hyvarinen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., 2001.

[2] C. A. Ross O. M. E. Mitchell and H. Yatos G, "Signal processing for cooktail party effect," *J. Acoust. Soc. Am.*, vol. 50, no. 2, pp. 656–660, 1971.

[3] C. Jutton and J. Herault, "Blind separation of sources, part i:an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[4] S. Araki H. Sawada, R. Mukai and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Proc. of ICA2003*, pp. 505–510, 2003.

[5] A. Miwa ans S. Morita, "Sound source separation for stereo music signal recorded in an active environment," *proc. of ICME2001*, pp. 205–209, 2001.

[6] A. Miwa ans S. Morita, "Automatic music transcription for a trio using stereo musical sound signal," *proc. of ICMCS, Vol.2 1*, vol. 2, pp. 824–827, 1999.

[7] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, no. 4, pp. 911–918, 1976.

[8] A. Nohorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. on ASSP*, vol. 34, no. 5, pp. 1124–1138, 1986.

[9] M. Abe and S. Ando, "Application of loudness/ pitch/ timbre decomposition operators to auditory scene analysis," *Proc. of ICASSP*, pp. 1124–1138, 1986.

[10] M. Goto T. Nakatani and H. Okuno, "Localization by harmonic structure and its application to harmonic sound stream segregation," *proc. of Acoustics, Speech and Signal processing*, pp. 653–656, 1996.

[11] K. Kashino and H. Murase, "Sound source identification system for ensemble music based on template adaptation and music stream extraction,speech communication," *Speech Communication*, vol. 27, pp. 337–359, 1999.