SOUND LOCALIZATION ANALYSIS OF STEREO AUDIO SIGNALS BASED ON BLIND SOURCE SEPARATION

Yuuki Haraguchi†, Shigeki Miyabe †*, Hiroshi Saruwatari†, Kiyohiro Shikano†,Toshiyuki Nomura‡

†Nara Institute of Science and Technology
Nara, 630–0192, Japan
Phone:+81-743-72-5287
Email: yuuki-h@is.naist.jp

‡Common Platform Software Research Laboratories
NEC Corporation
Kanagawa, 211–8666, Japan

Abstract

In this paper, utilizing blind source separation (BSS) based on independent component analysis (ICA), we propose a method to analyze and control sound localization for each of the sound sources with only information of the multichannel signals as mixture of multiple sound sources. In the conventional BSS, the demixing filter has distortion caused by ambiguity of amplitude. To obtain the separated signals without distortion, the reconstruction of the original sound localization at audio channels using inverse filter of the demixing filter is proposed. The inverse filter of the demixing filter is useful for analysis of sound localization, however, the inverse filter has not only the effect of sound-localization reconstruction but also the effect of the distortion compensation. The compensation of distortion operates as imposition of distortion to the signals other than the one distorted by BSS. Thus, when the inverse filter is used for control of sound localization directly, the quality of the controlled sound source degrades. In this paper, we propose a method to extract the information of sound localization without containing the function of the compensation. By obtaining the demixing filter without distortion, the effect of the compensation of distortion is removed from the inverse filter. The obtained inverse filter is a replica of source-to-channel transfer function which determines sound localization. By modifying the inverse filter, sound localization of source can be controlled individually and freely.

1. Introduction

The recent advance and price-reduction of DSP have spread to various audio effector systems, which can achieve not only simple tone control but also 3D audio effects to control reverberation, width of chamber and many other spatial impressions. However, they are merely modifications of ready-made multichannel audio, and they are not sufficient for user-controllable audio reedit. Our research purpose is to construct a system in which users can reedit each of the sources as if the users can manipulate the mixing console by themselves and achieve

- customizable spatial re-allocation of audio objects,
- selectable enhancement of specific sources, and
- listener's virtual movement in primary sound field.

As one piece of evidence that the user-controllable audio reedit is in strong demand, the ISO/IEC Moving Picture Experts Group (MPEG) has started the Spatial Audio Object Coding (SAOC) project, which aims to standardize user-controllable audio technology. The most attention-getting technology of SAOC is binaural cue coding (BCC) [ ], which has been adopted as the basis of MPEG Surround codec standardized before SAOC. This codec can encode multichannel signal with a low bit-rate by parameterizing some spatial characteristics. However, this method analyzes characteristics of mixed audio signals but not localization of those sources, and is insufficient for the source reedit. Although some researchers have proposed a system to extract and edit vocal and drums parts [ ], they utilize characteristics of specific instruments and cannot be applied to general audio signals.

In this paper, we propose analysis and modification of source localization. When available information is only the stereo audio signal itself, which is a mixture of multiple sources, we have to extract information on the objective sources to control localization. For this purpose, we focus on blind source separation (BSS), especially that based on independent component analysis (ICA) [ ] because of its high-quality separability. In the conventional BSS, optimized ICA outputs distorted monaural estimation of each separated source. To compensate the distortion, the distorted monaural output is reconstructed as stereo source signal with its information on localization recovered. We analyze information of localization by dividing BSS into two steps; the first step is monaural source separation with low distortion, and the second step is reconstruction of the sources' spatial information. In addition, we introduce an application of audio object control, which is achieved by processing the extracted information of sound localization. The effectiveness of the proposed methods is ascertained in subjective evaluations. Finally, subjective evaluation results show that the proposed method can control sound localization of each sound source flexibility and individually.

*Research Fellow of the Japan Society for the Promotion of Science.

2. CONVENTIONAL BLIND SOURCE SEPARATION

2.1. Mixing Model

In this paper, we assume that the number of sound sources is L and the number of audio channels is M, and we deal with the case of $L = M$.

The source signal of the L sources in the time-frequency domain is denoted by an L-dimensional vector $\boldsymbol{S}(f,t) = [S_1(f,t),\ldots,S_L(f,t)]^{\mathrm{T}}$, where f is the index of the frequency bin and t is the index of the analysis frame. In addition, a linear time-invariant transfer system is denoted by an $M \times L$ mixing matrix $\boldsymbol{A}(f) = [A_{ml}(f)]_{ml}$, where $A_{ml}(f)$ is the transfer function from the l-th source to the m-th channel, and $[x]_{ml}$ denotes the matrix that includes the element x in the m-th row and the l-th column. Then, the observed signal $\boldsymbol{X}(f,t) = [X_1(f,t),\ldots,X_M(f,t)]^{\mathrm{T}}$ is written approximately as

$$\boldsymbol{X}(f,t) = \boldsymbol{A}(f)\boldsymbol{S}(f,t). \tag{1}$$

2.2. Frequency Domain ICA

Assuming the source signals are statistically independent mutually and no more than one source is Gaussian, ICA learns the demixing filter in an unsupervised manner. The condition of successful separation with the demixing filter is equivalent to independence among output signals. Here we describe frequency domain ICA (FDICA) [ ] where ICA is processed in the frequency domain. In this method, by using the demixing matrix $\boldsymbol{W}(f) = [W_{lm}(f)]_{lm}$, the separated signal $\boldsymbol{Y}(f,t) = [Y_1(f,t),\ldots,Y_L(f,t)]^{\mathrm{T}}$ is given by

$$\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t). \tag{2}$$

Here $\boldsymbol{W}(f)$ can be optimized by the following iterative updating formula [ ]:

$$\boldsymbol{W}^{[i+1]}(f) = \mu\left[\boldsymbol{I} - \langle\boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}(f,t)^{\mathrm{H}}\rangle_t\right]\boldsymbol{W}^{[i]}(f) + \boldsymbol{W}^{[i]}(f), \tag{3}$$

where $\boldsymbol{I}$ denotes the identity matrix, $\langle\cdot\rangle_t$ denotes the time-averaging operator, H shows conjugate transposition, $[i]$ is used to express the value of the i-th step in the iterations, μ is the step-size parameter, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function [ ].

2.3. Projection Back

Since the criterion of independence does not specify amplitude and order of signals, FDICA itself is insufficient as filter learning. Ambiguity of amplitude, the so-called *scaling problem*, randomizes spectral characteristics and it results as distortion in the output signals. Ambiguity of order is known as the *permutation problem*, and without identifying correspondence between the sources and separated outputs, broad-band estimation of the source signals cannot be obtained. Here we discuss the solution of the scaling problem using projection back (PB) [ ]. Under the assumption that the demixing matrix $\boldsymbol{W}(f)$ separates source components accurately, and permutation ambiguity is aligned by some means [ ], $\boldsymbol{W}(f)$ can be expressed as follows:

$$\boldsymbol{W}(f) = \mathrm{Diag}(\boldsymbol{C}(f))\boldsymbol{A}(f)^{-1}, \tag{4}$$

where $\boldsymbol{C}(f) = [C_1(f),\ldots,C_L(f)]^{\mathrm{T}}$ is a constant vector which denotes gain ambiguity of ICA, $\mathrm{Diag}(\cdot)$ is the diagonal matrix whose diagonal component is each element of the argument column vector. To compensate the effect of $\boldsymbol{C}(f)$, PB applies the inverse matrix of the demixing matrix

$$\boldsymbol{H}(f) = \boldsymbol{W}(f)^{-1}. \tag{5}$$

The inverse matrix $\boldsymbol{H}(f)$ reconstructs the amplitude of the separated signals at each of the audio channels, and its output signal $Z_{ml}(f,t)$ of the l-th separated signal at the m-th channel can be given as follows:

$$\begin{aligned}
&[Z_{ml}(f,t)]_{ml}\\
&= \boldsymbol{H}(f)\mathrm{Diag}(\boldsymbol{Y}(f,t))\\
&= [\mathrm{Diag}(\boldsymbol{C}(f))\boldsymbol{A}(f)^{-1}]^{-1}\mathrm{Diag}(\mathrm{Diag}(\boldsymbol{C}(f))\boldsymbol{A}(f)^{-1}\boldsymbol{X}(f,t))\\
&= \boldsymbol{A}(f)\mathrm{Diag}(\boldsymbol{C}(f))^{-1}\mathrm{Diag}(\boldsymbol{C}(f))\mathrm{Diag}(\boldsymbol{A}(f)^{-1}\boldsymbol{X}(f,t))\\
&= \boldsymbol{A}(f)\mathrm{Diag}(\boldsymbol{A}(f)^{-1}\boldsymbol{A}(f)\boldsymbol{S}(f,t))\\
&= \boldsymbol{A}(f)\mathrm{Diag}(\boldsymbol{S}(f,t)).
\end{aligned} \tag{6}$$

Thus, the scaling problem is solved in the form of the reconstruction of the transfer system $\boldsymbol{A}(f)$.

In general, $Z_{ml}(f,t)$ is obtained directly by the filter $\boldsymbol{H}(f)\boldsymbol{W}_l(f)$ instead of obtaining $\boldsymbol{Y}(f,t)$ where $\boldsymbol{W}_l(f)$ denotes the demixing matrix replacing all coefficients by zero except the l-th row of $\boldsymbol{W}(f)$ (see Fig. 1).

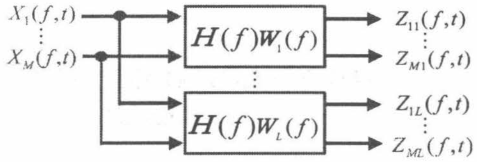

Figure 1: Separation step based on conventional ICA (FDICA+PB).

3. ANALYSIS OF SOUND LOCALIZATION

In this section, we propose an analysis method of sound localization. Since sound localization is determined individually for each of the sound sources, analysis of sound localization is inextricably linked to source separation.

3.1. Extraction of Sound-Localization Information

From Eq. (1), the transfer system $\boldsymbol{A}(f)$ contains all information on sound localization for each of the source signals in $\boldsymbol{S}(f,t)$. Assuming $\boldsymbol{A}(f)$ is known, the following processing is possible by using $\boldsymbol{A}(f)$.

First, from Eq. (1), the source separation is entirely achieved as follows:

$$\begin{aligned}
\boldsymbol{Y}(f,t) &= \boldsymbol{A}(f)^{-1}\boldsymbol{X}(f,t)\\
&= \boldsymbol{S}(f,t).
\end{aligned} \tag{7}$$

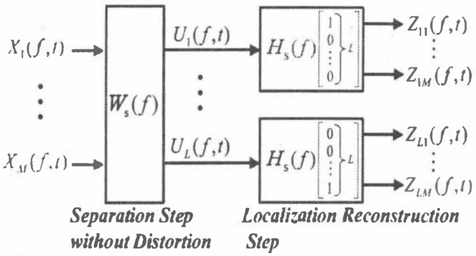

Figure 2: Procedure step of proposed method

Second, the same sound localization that the l-th sound source has can be given to another monaural sound source $R(f,t)$ as

$$\tilde{\boldsymbol{X}}(f,t)$$
$$= \boldsymbol{A}(f)[Y_1(f,t),\dots,Y_{l-1}(f,t),R(f,t),Y_{l+1}(f,t),\cdots,Y_L(f,t)]^{\mathrm{T}}, \quad (8)$$

where $\tilde{\boldsymbol{X}}(f,t)$ is the signal replacing $Y_l(f,t)$ with $R(f,t)$. We call such a substitution of the sources *punch in*.

Third, we can control sound localization individually and freely by modifying the relation among the channels of $\boldsymbol{A}(f)$. Using modified transfer system $\hat{\boldsymbol{A}}(f)$, the localization-controlled audio signal $\hat{\boldsymbol{X}}(f,t) = [\hat{X}_1(f,t),\dots,\hat{X}_M(f,t)]^{\mathrm{T}}$ is given by

$$\hat{\boldsymbol{X}}(f,t) = \hat{\boldsymbol{A}}(f)\boldsymbol{Y}(f,t). \quad (9)$$

However, the transfer system $\boldsymbol{A}(f)$ is generally unknown in practical situations and should be estimated by some means.

3.2. Problem of Conventional Projection Back

From Eqs. (4) and (5), the inverse matrix $\boldsymbol{H}(f)$ of $\boldsymbol{W}(f)$ estimated by ICA is given as follows:

$$\boldsymbol{H}(f) = \boldsymbol{A}(f)\mathrm{Diag}(\boldsymbol{C}(f))^{-1}. \quad (10)$$

If the punch-in process was implemented by using $\boldsymbol{H}(f)$, another monaural sound source $R(f,t)$ would be affected not only by $\boldsymbol{A}(f)$ but also by $\mathrm{Diag}(\boldsymbol{C}(f))^{-1}$. Therefore, $\boldsymbol{H}(f)$ is inadequate to substitute for the transfer system $\boldsymbol{A}(f)$. However, it is very difficult for the deconvolution to be achieved without information on source signals[ ] .

3.3. Proposed Algorithm

If the separation is achieved without distortion, its inverse filter plays only the role of reconstructing localization. Then the inverse filter can be used as an approximation of the transfer system $\boldsymbol{A}(f)$, and can achieve punch in without distorting the substituted source. Thus our strategy is to divide the separation process into two steps: a separation step without distortion and a localization reconstruction step (see Fig. 2). The localization control can be achieved by modifying the localization reconstruction step.

3.3.1. Separation step without distortion

In this section, to separate the observed signals into each of the monaural source signals without distortion, we obtain the demixing filter $\boldsymbol{W}_{\mathrm{s}}(f)$, which intentionally scales each of its separated signals to an average value of the channels.

It is easy to obtain the average value of the channels with respect to each sound source at the audio channels by using PB. Furthermore, it can be said that the average value of the channels is a monaural signal with little distortion. By using Eq. (6), the channel-averaged source estimation is given by

$$\frac{1}{M}\cdot\left[\sum_{m=1}^{M} Z_{m1}(f,t),\dots,\sum_{m=1}^{M} Z_{mL}(f,t)\right]^{\mathrm{T}}$$
$$= \frac{1}{M}\cdot\mathrm{Diag}([\boldsymbol{H}(f)]^{\mathrm{T}}[\underbrace{1,\dots,1}_{M}]^{\mathrm{T}})\boldsymbol{W}(f)\boldsymbol{X}(f,t). \quad (11)$$

Therefore, the demixing filter $\boldsymbol{W}_{\mathrm{s}}(f)$ is defined as follows:

$$\boldsymbol{W}_{\mathrm{s}}(f) = \frac{1}{M}\cdot\mathrm{Diag}\Big([\boldsymbol{H}(f)]^{\mathrm{T}}[\underbrace{1,\dots,1}_{M}]^{\mathrm{T}}\Big)\boldsymbol{W}(f). \quad (12)$$

Thus, by using $\boldsymbol{W}_{\mathrm{s}}(f)$, the average value of the channels with respect to each sound source $\boldsymbol{U}(f,t) = [U_1(f,t),\dots,U_L(f,t)]^{\mathrm{T}}$ is given by

$$\boldsymbol{U}(f,t) = \boldsymbol{W}_{\mathrm{s}}(f)\boldsymbol{X}(f,t). \quad (13)$$

3.3.2. Localization reconstruction step

Here, $\boldsymbol{H}_{\mathrm{s}}(f) = \boldsymbol{W}_{\mathrm{s}}(f)^{-1}$ is defined as the localization reconstruction filter. This filter only takes charge of reconstructing sound localization to the separated signal $\boldsymbol{U}(f,t)$.

By applying $\boldsymbol{H}_{\mathrm{s}}(f)$ to $\boldsymbol{U}(f,t)$, the output signals are equivalent to the output signals of PB as follows:

$$\boldsymbol{H}_{\mathrm{s}}(f)\mathrm{Diag}(\boldsymbol{U}(f,t)) = [Z_{ml}(f,t)]_{ml}\,. \quad (14)$$

This indicates that $\boldsymbol{H}_{\mathrm{s}}(f)$ reconstructs the inter-channel level and phase differences to the monaural separated signal $\boldsymbol{U}(f,t)$, which has sound reverberation caused by the transfer system $\boldsymbol{A}(f)$. Therefore, $\boldsymbol{H}_{\mathrm{s}}(f)$ can be approximated to play only the role of reconstructing sound localization of $\boldsymbol{U}(f,t)$.

4. MODIFICATION OF ANALYZED LOCALIZATION INFORMATION

In this section, we introduce an application of sound-localization control for stereo audio signals consists of two sources by modifying the localization reconstruction filter described in Sect. 3.3.

4.1. Inverse Localization

Inverse localization is a process to exchange the spatial characteristics of the two sources in the stereo audio signals. Such inversion can be considered as a variation of the punch-in process because this process is achieved by modifying Eq. (8). The localization-inverted stereo signals $\bar{\boldsymbol{X}}_{\mathrm{prop}}(f,t) = [\bar{X}_{\mathrm{L}_{\mathrm{prop}}}(f,t),\bar{X}_{\mathrm{R}_{\mathrm{prop}}}(f,t)]^{\mathrm{T}}$ are given by

$$\bar{\boldsymbol{X}}_{\mathrm{prop}}(f,t) = \boldsymbol{H}_{\mathrm{s}}(f)\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\boldsymbol{U}(f,t). \quad (15)$$

Note that inverse localization cannot be achieved by a simple exchange of the left and right channels.

Table 1: Experimental condition

Source set 1	Trombone + guitar
Source set 2	Piano + guitar
Source set 3	Singing voice + guitar
Sampling frequency	44.1 kHz
Quantization bit	16 bit
Filter length	512 point
ICA iterations	500

5. SUBJECTIVE EVALUATION

In this section, we describe the details and the results of subjective evaluation to verify the effectiveness of inverse localization introduced in Sect. 4.1 by comparing the proposed method and the competitive method described in the following.

5.1. Competitive Method

As the competitive method to verify quality of localization-inversed audio signal comparing the proposed method, inverse localization is processed by using PB filter $\boldsymbol{H}(f)$. Localization-inversed signal $\bar{\boldsymbol{X}}_{\mathrm{comp}}(f,t) = [\bar{X}_{\mathrm{L_{comp}}}(f,t), \bar{X}_{\mathrm{R_{comp}}}(f,t)]^{\mathrm{T}}$ by FD-ICA and PB is given by

$$\bar{\boldsymbol{X}}_{\mathrm{comp}}(f,t) = \boldsymbol{H}(f)\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\boldsymbol{Y}(f,t). \qquad (16)$$

5.2. Experimental Condition

The signals used in the experiment are three kinds of signal, and each of signals is real audio stereo signal which consists of two sources packed in the commercial-release compact disc. The detail of the signals is denoted in Table 1. Sound quality of localization-inversed audio signal by the proposed method and the competitive method is evaluated in an XAB test comparing the localization-inversed audio signal and original audio signal. Nine subjects participated in this experiment.

5.3. Experimental Result

The subjective evaluation result is denoted in Fig. 3. For each of the source sets, the proposed method successfully maintained better sound quality of original signal than the competitive method. This shows that, the localization reconstruction filter $\boldsymbol{H}_{\mathrm{S}}(f)$ in the proposed method takes only charge of reconstructing sound localization to the separated signal, which is an imitation of the transfer system $\boldsymbol{A}(f)$, and makes punch-in process available.

6. CONCLUSION

In this paper, to control sound localization for each of sound sources individually by using BSS based on ICA, we proposed an analysis of sound localization by dividing the conventional source separation process into the two steps: the monaural separation step without distortion and the localization reconstruction step. In addition, we indicated an application of sound-localization control of stereo signals with two sources, which are achieved by modifying the proposed analysis of sound localization. The result of subjective evaluation about the inverse localization shows that the localization reconstruction filter in the proposed method takes only charge of reconstructing sound localization to a monaural signal, approximates the transfer system, and makes punch-in process available.

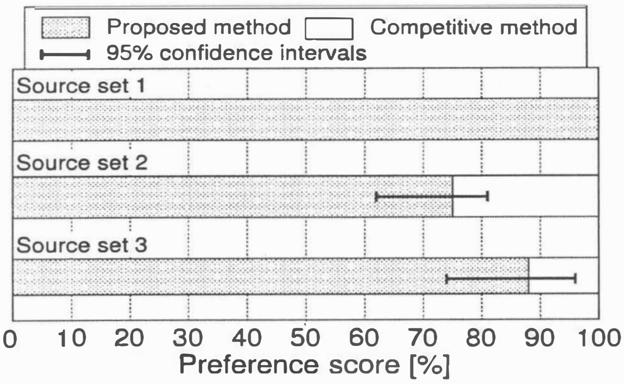

Figure 3: Experimental result.

References

[1] C. Faller and F. Baumgarte, "Binaural Cue Coding—Part II: Schemes and Applications," IEEE Trans Speech and Audio Processing, vol. 11, no. 6, pp. 520–531, 2003.

[2] O. Gillet and G. Richard, "Extraction and remixing of drum trucks from polyphonic music signals," Proc. WASPAA, pp. 315–318, 2005.

[3] P. Comon, "Independent component analysis—A new concept?," Signal Processing, vol. 36, pp. 287–314, 1994.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol. 22, pp. 21–34, 1998.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," IEICE Trans. Fundamentals, vol. E86-A, no. 3, pp. 590–596, 2003.

[6] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," Proc. NOLTA'98, pp. 923–926, 1998.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," IEEE Trans. Speech and Audio Processing, vol. 12, no. 5, pp. 530–538, 2004.

[8] H. Saruwatari, H. Yamajo, T. Takatani, T. Nishikawa, and K. Shikano, "Blind separation and deconvolution for convolutive mixture of speech combining SIMO-model-based ICA and multichannel inverse filtering," IEICE Trans. Fundamentals, vol. E88-A, no. 9, pp. 2387–2400, 2005.