Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Sound Source Separation and Synthesis for Audio Enhancement based on Spectral Amplitudes of Stereo Signals

Masayuki Nishiguchi, Ayumu Morikawa, Yuya Ishii
Kanji Watanabe, Koji Abe, and Shouichi Takane

Department of Electronics and Information Systems
Akita Prefectural University
84-4 Ebinokuchi Tsuchiya Yurihonjo Akita Japan
nishiguchi@akita-pu.ac.jp

Received February 2017; revised September 2017

Abstract. *A sound source separation algorithm based on the spectral amplitudes of stereo signals has been developed for the up-mixing playback of the contents. Short-term Fourier transforms (STFT) of the signals on the left and right channels are first calculated. The coefficients of the discrete Fourier transform (DFT) are used to calculate the ratio of the spectral amplitudes of the left and right channels, which is termed the channel level difference (CLD). The DFT coefficients are then divided into multiple groups on the basis of the CLD, with each group representing a separated sound source. The signal-to-distortion ratio (SDR) is used to evaluate the signal separation performance. It was found that a rough estimate of the CLD threshold yielding the best SDR could be obtained by cross-correlating the separated sounds. The proposed method allows separation of a mixture of more than two sound sources. For playback on a headset, each separated signal is convolved with head-related transfer functions (HRTF) that represent the direction of that particular sound source. Subjective listening tests showed that the sound synthesized by this method is more realistic than that synthesized with HRTFs that represent only left and right loudspeakers.*
Keywords: Head Related Transfer Function, Source Separation, Short-term Fourier transform, Up-mix

1. **Introduction.** The popularity of audiovisual content for DVDs, Blu-ray Discs, and other media for consumer electronics appliances is driving the demand for multichannel audio playback systems (5.1ch, 7.1ch, virtual-3D headphones, etc.) that provide highly realistic audio representation. On the other hand, although most media are still encoded in the 2-channel (2ch) stereo format, music distribution services are moving away from that format and from compact discs. To obtain more realistic sound quality with multichannel-playback systems and virtual-3D headphone systems, it is desirable to convert 2ch content into a multichannel format. That requires appropriate up-mixing technologies. If the sound sources recorded on a 2ch CD could be separated into discrete sources, then it would be easy to up-mix the sound to a greater number of channels (e.g., 5.1ch) by using virtual microphones in a computer. Also, it is possible to apply HRTFs to each sound source based on its direction to enable virtual-3D playback for headphones. Thus, a sound source separation technology that employs the spectral amplitudes of the left and right channels of the source signals has been developed to accomplish that.

Sound source separation technologies are based on a variety of principles, such as independent-component analysis (ICA) [1], non-negative matrix factorization (NMF) [2][3], or deep neural networks (DNN) [4][5]. ICA requires at least N channels of observed sound to separate N source signals, provided that each source signal is statistically independent of the others. So, if we start off with 2ch stereo signals, it can distinguish at most 2 source signals, which is inadequate for our purposes. On the other hand, NMF creates a factorization of a spectrogram of captured sound material, where a spectrogram is represented by a sum of multiple products of spectral bases and their activation functions. Repeated operations are needed to extract appropriate spectral bases. Finally, if training data were appropriately selected, a DNN could be trained to extract a certain instrument from a mixture of signals.

The method proposed here uses only the ratio of the spectral amplitudes of the signals on the left and right channels. Spectral coefficients are clustered into several groups based on that ratio, and each group is taken to be a discrete sound source. This method roughly divides the sound space into multiple directions, which can be considered to be a spatial allocation of sound sources. The reason this approach is taken is that the first goal is to use HRTFs to generate a virtual-3D sound field. There is a set of HRTFs for every 10° sector of the horizontal plane, which together cover a full 360°. Thus, if the sound space is divided into multiple directions, these HRTFs can be used to synthesize a virtual-3D sound field for headphone playback.

The algorithm to segregate desired speech signals from concurrent sounds is shown in [6], where the differences in spectral amplitudes and phases received by two microphones are utilized. Our proposal is, on the other hand, to separate sound sources using 2ch stereo signals manually mixed down from many tracks with panning operation. Therefore, our proposed method does not use differences of phases but amplitudes, since panning operation involves only amplitude modifications for most of the cases. More importantly, our proposal is to provide an algorithm to find the best threshold values to cluster the amplitude differences into multiple groups to allow best separation of the source signals from each other on condition that sound source locations are unknown, while [6] does not discuss anything about optimizing such threshold values since it assumes sound source locations are known.

The paper is organized as follows: Chapter 2 defines the channel level difference (CLD), which is the ratio of the spectral amplitudes of the left- and right-channels. Chapter 3 shows the separation algorithm. Chapter 4 presents a criterion for evaluating the separation performance and defines the signal-to-distortion ratio (SDR). Chapter 5 explains how to generate test signals used to calculate the SDR of the separated signals. Chapter 6 shows how to find a suboptimal CLD threshold for creating a near-optimal separation in terms of SDR. Chapter 7 presents separation algorithm for signals consisting of more than two sound sources. Chapter 8 explains how to apply HRTFs to the separated sources and use the proposed separation algorithm for highly realistic headphone playback. Chapter 9 presents the results of listening tests, and Chapter 10 makes some concluding remarks.

2. **Channel level difference.**

Several tens of original tracks, and sometimes more than a hundred, are usually used to generate the content for CDs. They are mixed down to 2 channels. In this process, mastering engineers allocate different instruments in the sound space by panning with the left and right channels. Thus, in most cases, the difference in panning level can be used as a cue to how to separate individual instruments. Based on this idea, our method calculates the level differences of the spectral amplitudes of the left and right channels. This is termed the channel level difference (CLD). It is written either as a function of frequency or as a function of the index of the frequency bin, k, in the

DFT domain:

$$CLD(k) = 20 \log \left(\frac{|X_L(k)|}{|X_R(k)|} \right) , \tag{1}$$

where

$$|X_L(k)| = \sqrt{r_L(k)^2 + i_L(k)^2} \tag{2}$$

$$|X_R(k)| = \sqrt{r_R(k)^2 + i_R(k)^2} \ . \tag{3}$$

$r_L(k)$ is the real part and $i_L(k)$ is the imaginary part of the k-th DFT coefficient of the left channel, and $r_R(k)$ and $i_R(k)$ are for the right channel. $CLD(k)$ roughly indicates the direction that the k-th DFT coefficient is coming from. It is assumed that an identical instrument or sound source generates a group of DFT coefficients coming from about the same direction. Thus, grouping the DFT coefficients according to CLD enables the instruments and sound sources on a 2ch track of a CD to be separated. More precisely, this method separates or clusters the directions of sound sources rather than separating music sources themselves. However, this is reasonable for our application. The goal is to apply HRTFs to sound sources from multiple directions. The sources may not necessarily be independent of each other. So, sound direction separation, rather than sound source separation, should work better for our application.

The graph of CLD vs. frequency (Fig. 1) shows the CLD distribution over a segment (length: 2 seconds) of a music track. Each dot represents a CLD obtained from a DFT snapshot of the left or right channel for a particular frequency. The dots can be divided into two groups: those with CLDs above the range $-3 \sim 0dB$ and those with CLDs below that range. This implies that the sound comes mainly from two directions. The DFT coefficients for the two groups of CLDs also form two groups, each of which represents a direction of sound.

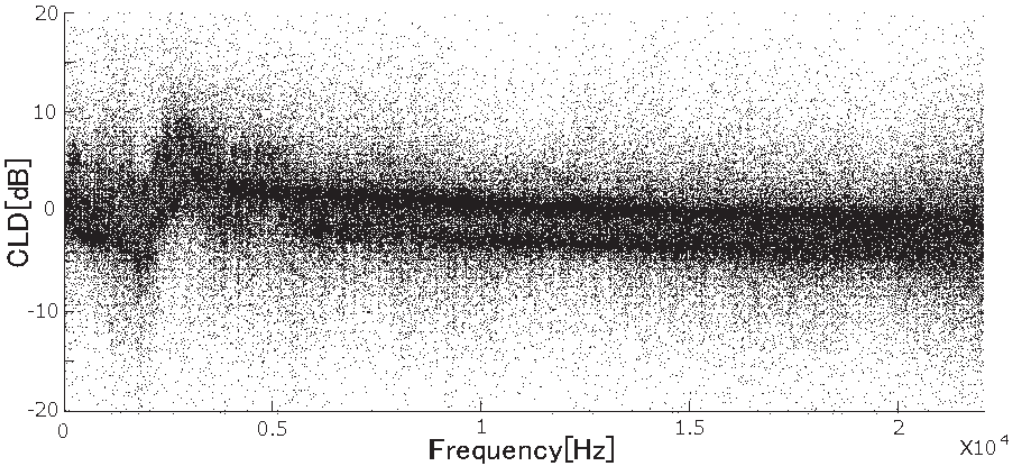

FIGURE 1. CLD distribution for 2 seconds of music

3. **Separation algorithm.** Let's assume a signal under consideration is composed of two sound sources A and B. Given a CLD threshold T, DFT coefficients of the left and right channel signal, $X_L(k)$ and $X_R(k)$, are clustered into two groups of DFT coefficients, $X_{LA}(k,T), X_{LB}(k,T)$, and $X_{RA}(k,T), X_{RB}(k,T)$, respectively as,

$$X_{LA}(k,T) = \begin{cases} X_L(k) & (CLD(k) < T) \\ 0 & (otherwise) \end{cases} \tag{4}$$

$$X_{LB}(k,T) = \begin{cases} X_L(k) & (CLD(k) > T) \\ 0 & (otherwise) \end{cases} \tag{5}$$

$$X_{RA}(k,T) = \begin{cases} X_R(k) & (CLD(k) < T) \\ 0 & (otherwise) \end{cases} \tag{6}$$

$$X_{RB}(k,T) = \begin{cases} X_R(k) & (CLD(k) > T) \\ 0 & (otherwise) \end{cases} , \tag{7}$$

where $X_{LA}(k,T), X_{RA}(k,T)$ are considered as DFT coefficients of left and right channel of sound source A, and $X_{LB}(k,T), X_{RB}(k,T)$ are considered as DFT coefficients of left and right channel of sound source B. Separated audio signals in time domain can be obtained by computing the inverse short-term Fourier transform (ISTFT) with an appropriate windowing (e.g. hanning window) and overlap-add (OLA) using the DFT coefficients $X_{LA}(k,T), X_{RA}(k,T)$, and $X_{LB}(k,T), X_{RB}(k,T)$. It is desirable to have a particular value of the CLD threshold, called the optimal CLD threshold, that separates the data points appropriately.

4. **Optimal CLD threshold and signal-to-distortion ratio.** As you can see Figure1, the location of the valley of the CLD distribution depends on frequencies. Therefore, ideally, the CLD threshold to separate the data points should be changed depending on the frequencies. However, in order to evaluate the validity of the base line of the proposed algorithm, the optimal CLD threshold is assumed to be constant over frequency and time. As explained in Chapter 6, the proposed algorithm cross-correlates the separated signals generated from each group of the data points to find the appropriate CLD threshold value. Therefore, CLD threshold is determined so that the major part of the data points will be appropriately classified, even with a fixed CLD threshold value. Before a way of determining it is discussed, the signal-to-distortion ratio (SDR) [7], which is used to evaluate the sound separation performance, is explained. The SDR is defined to be

$$SDR = 10 \log \left(\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} (x(n) - g \cdot x'(n))^2} \right), \tag{8}$$

where $x(n)$ is an original source signal, $x'(n)$ is a separated signal, N is the number of samples of the source signal, and g is a scaling factor that compensates for the change in gain during the separation process. g is given by

$$g = \frac{\sum_{n=0}^{N-1} x(n) \cdot x'(n)}{\sum_{n=0}^{N-1} x'(n)^2} . \tag{9}$$

The optimal CLD threshold is defined to be the CLD that gives the best SDR for the separated signals. It can be determined by calculating the SDR for every possible CLD threshold and selecting the CLD that gives the highest SDR. Clearly, the SDR can only be calculated if the original source signal is known. So, in practical situation, we need to infer a suboptimal CLD threshold from observed data to make a proper separation,

which is explained in Chapter 6. Known mixed music signals from some independent music source material are used to evaluate the proposed separation algorithm.

5. **Generation of test signals.** Mixed music signals are generated by using appropriate panning levels. In the generation process (Fig. 2), the coefficients α and β determine the panning levels. Two music sources A and B are used. Each of them is approximately 20 seconds long with the sampling rate of 44.1kHz. The music source A is a violin solo, and B is a guitar solo. It is assumed that the sound sources A and B are independent of each other, and that the total energy is the same before and after mixing. $(\alpha+\beta)^2$ is therefore set to 1.

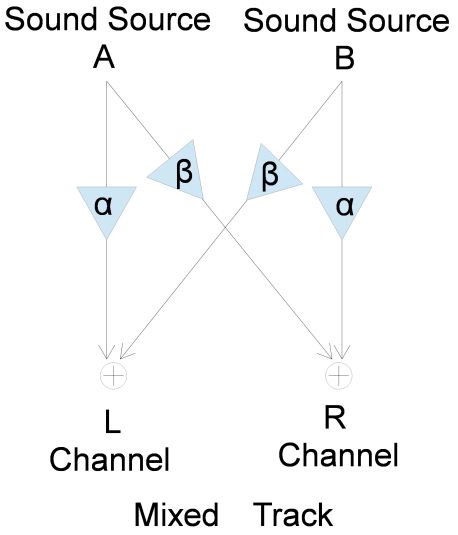

FIGURE 2. Mixing of sound source signals

6. **Suboptimal CLD threshold from cross-correlation.** Since the SDR can only be calculated when the original source signal is known, in real-world situations, the optimal CLD threshold cannot be found by computing the SDR. Our solution to this problem is to infer a suboptimal CLD threshold from observed data. This is accomplished by computing cross-correlations of time domain waveforms of the separated signals for different CLD thresholds. The CLD threshold that generates the local minimum in cross-correlation is assumed to be the best one and is called the suboptimal CLD threshold. If sound sources are appropriately separated, we can expect the value of the cross-correlation of the separated signals get smaller since common components included in both are eliminated by the separation while both of the separated signals have sufficient energy. If we separate the source signals with too high or too low CLD threshold, the energy of the either of the separated signals is close to zero, which is not appropriate separation in spite of small value of the cross-correlation. We therefore use a CLD value that gives local minimum between peaks in cross-correlation as the suboptimal threshold.

Let us explain this cross-correlation method in detail. Let $x_{LA}(n,T), x_{LB}(n,T), x_{RA}(n,T)$, and $x_{RB}(n,T)$ be the time domain signals obtained by computing the ISTFT, hanning windowing, and overlap-add (OLA) with sequences of DFT coefficients $X_{LA}(k,T)$, $X_{LB}(k,T)$, $X_{RA}(k,T)$, and $X_{RB}(k,T)$. The cross-correlation of the source A and source B in the left channel is defined as $\phi_L(\tau_L,T)$,

$$\phi_L(\tau_L, T) = \begin{cases} \sum_{n=0}^{N-1-\tau} x_{LA}(n,T)x_{LB}(n+\tau_L,T) & (0 \le \tau_L < N) \\ \sum_{n=-\tau_L}^{N-1} x_{LA}(n,T)x_{LB}(n+\tau_L,T) & (-N < \tau_L < 0) \end{cases} , \quad (10)$$

where N is the number of the samples of source signals in time domain, and τ_L is the number of sample shift between source A and source B. The cross-correlation of the right channel $\phi_R(\tau_R, T)$ is also computed in the same manner from $x_{RA}(n,T)$ and $x_{RB}(n,T)$. Now we define similarity $d(T)$ of the separated source A and B as,

$$d(T) = \max\{\phi_L(\tau_L, T), (-N < \tau_L < N)\} + \max\{\phi_R(\tau_R, T), (-N < \tau_R < N)\} , \quad (11)$$

where $\max\{\phi(\tau, T), (-N < \tau < N)\}$ returns the maximum value of $\phi(\tau, T)$ within $-N < \tau < N$ for a given T. Figure 3 shows the relationship between CLD threshold T vs. similarity $d(T)$, where the number of the DFT points in STFT is 16384.

Test signals (see Chapter 5) were used to determine whether or not the suboptimal CLD threshold, which is given by the local minimum in similarity $d(T)$, was close to the optimal CLD threshold, which is given by the highest SDR. The graph of SDR vs. CLD threshold and similarity $d(T)$ vs. CLD threshold for a music sample (Fig. 3) shows that the two are fairly close, and that the suboptimal CLD threshold can be obtained by looking at similarity $d(T)$. As you can see, the local minimum of the similarity $d(T)$ is given at around -0.3 dB in CLD threshold, and maximum SDR of the separated signals are given at around 0.4 dB. However, the SDR values at -0.3dB are almost the same as the best SDR values given at 0.4 dB, and the differences of the SDR values at these two CLD thresholds are about 0.2 dB or so. This means the CLD threshold value given by the local minimum of the similarity $d(T)$ provides fairly close SDR values to the best SDR values, and therefore we can say that the CLD threshold is a good estimation of the optimal CLD threshold.

The SDRs obtained from the optimal and suboptimal (i.e., estimated) CLD thresholds were plotted against panning level for two samples (Fig. 4). The differences of the SDRs obtained from the optimal and suboptimal CLD thresholds are approximately $0.2 \sim 0.5$ dB for most panning levels, and 1 dB or so in the worst case. This endorses the suboptimal CLD thresholds are good estimation of the optimal CLD thresholds.

This method is possible because an overlap-add (OLA) method with 75% overlap is applied to the ISTFT to synthesize the time domain waveform. Although the DFT coefficients are divided into multiple groups in any given time frame, a DFT coefficient in the next time frame at the same frequency as that of the current frame may be put into a group different from that of the current frame. Since the time domain frames overlap by 75% in the OLA, the same frequency component may exist in different groups in the time domain waveforms. There is, therefore, a certain correlation between the different groups of the synthesized waveforms. This happens mainly when a CLD of a DFT coefficient is close to the CLD threshold being used. So when a CLD threshold clearly separates the DFT coefficients into different groups, the cross-correlation is small, leading to a small similarity $d(T)$.

7. **Separation of more than two sound sources.** For the simplicity of the explanation, source separation of two sound source have been discussed so far. However, the proposed method can separate more than two sound sources. Fig. 5 shows similarity

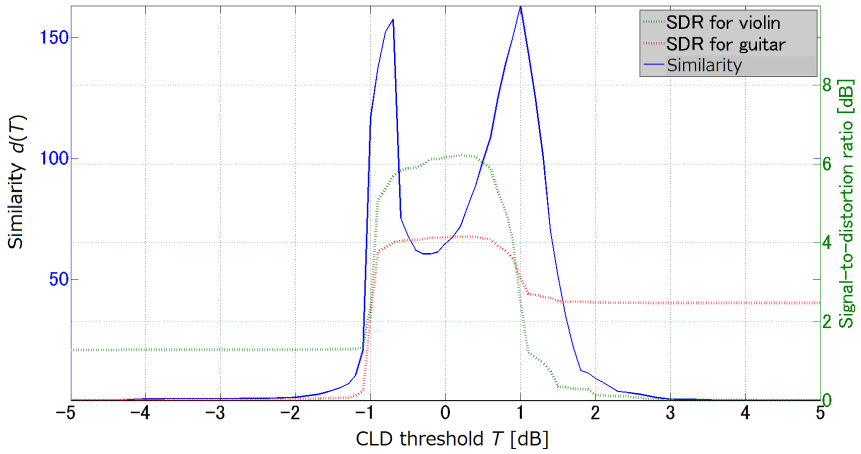

FIGURE 3. Similarity and SDR vs. CLD threshold for violin and guitar.

$d(T)$ of the signals separated into two groups along with the change of the CLD threshold T. The sound source used here is a rock music including vocal, sub-vocal, and guitar. As you can see, there are three peaks in similarity, each of which corresponds to each sound source. The left peak represents guitar, the middle peak represents vocal, and the right peak represents sub-vocal. Setting two CLD thresholds at the local minimum of the similarity, which are in between the three peaks, the DFT coefficients are clustered into three groups. Each of the group corresponds to the three sound source objects.

In general, when we have $M-1$ local minimum of similarity $d(T)$ at CLD thresholds T_{min} $(1 \leq min < M)$, the DFT spectra $X_L(k)$ and $X_R(k)$ can be separated into M groups, $X_{Lm}(k)$ and $X_{Rm}(k)$ $(0 \leq m < M)$ respectively, representing M sound source objects as,

$$X_{Lm}(k) = \begin{cases} X_L(k) & (T_m < CLD(k) < T_{m+1}) \\ 0 & (otherwise) \end{cases} \tag{12}$$

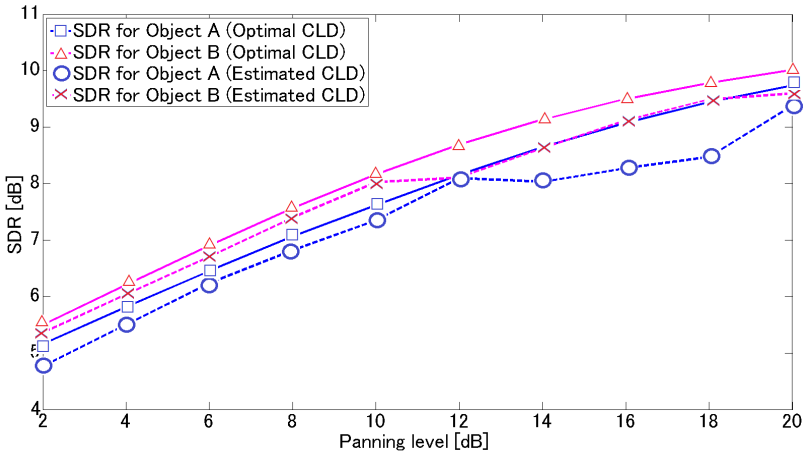

FIGURE 4. SDRs of optimal and suboptimal (i.e. estimated) CLDs vs. panning level.

$$X_{Rm}(k) = \begin{cases} X_R(k) & (T_m < CLD(k) < T_{m+1}) \\ 0 & (otherwise) \end{cases} , \tag{13}$$

where

$$T_m = \begin{cases} -\infty & (m = 0) \\ T_{min} & (1 \le m, min < M) \\ +\infty & (m = M) \end{cases} . \tag{14}$$

Separated signals in time domain are then synthesized from these DFT coefficients $X_{Lm}(k)$ and $X_{Rm}(k)$ $(0 \le m < M)$, in the same manner as the case of two sound sources.

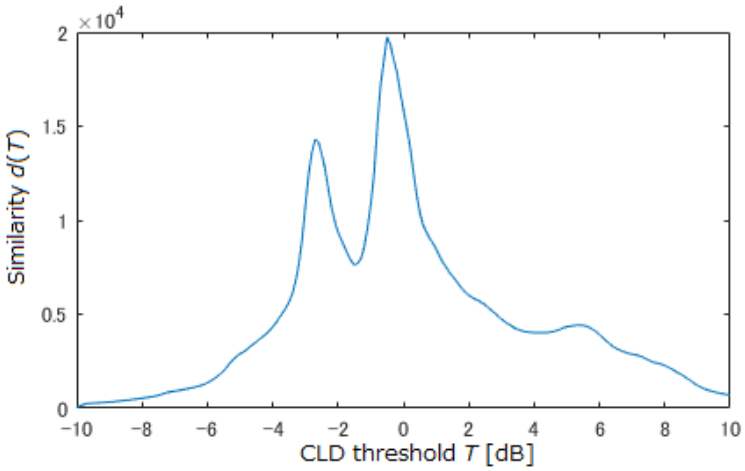

FIGURE 5. Similarity d vs. CLD threshold T for three sound source objects

8. **Application to virtual-3D headphones.** It was shown above that a suboptimal CLD threshold for separating source signals can be obtained by checking the cross-correlations between separated signals. This method was used to separate signals for the highly realistic playback of audio content. HRTFs were used for playback over headphones, and there was a group of HRTFs for each 10° sector of the horizontal plane. HRTFs were applied to each separated sound source based on the direction of the incoming sound from the standpoint of the listener. We therefore need to know the relationship between (a) the difference in signal level between the left and right ears, which is termed the interaural level difference (ILD), and (b) the directional angle from which the sound is coming. For this purpose, the difference in the energy of the head-related impulse response (HRIR) between the left and right ears is computed for all the angles in the database. Note that an HRIR is a time domain representation of an HRTF. Now we compute

$$ILD(\theta) = 10 \log \frac{\sum_n HRIR_R(\theta, n)^2}{\sum_n HRIR_L(\theta, n)^2} , \tag{15}$$

where $ILD(\theta)$ is the ILD as a function of the direction of the sound source, θ, in degrees; and $HRIR_L(\theta, n)$ and $HRIR_R(\theta, n)$ are the HRIRs of the left and right ears, respectively, for the direction θ and time index n. We used 256 samples of HRIR data $(0 \le n < 256)$ for left and right ears respectively in computing $ILD(\theta)$.

The dependence of $ILD(\theta)$ on the direction θ (Fig. 6) tells us how the difference in the level of a source signal between the left and right ears is related to its direction. This information, together with the CLD that gives the peak similarity $d(T)$ of the separated sound sources, enables us to estimate the directional angle of a separated sound source in the horizontal plane. Then, the HRTF filter appropriate to that angle is applied to the separated sound object.

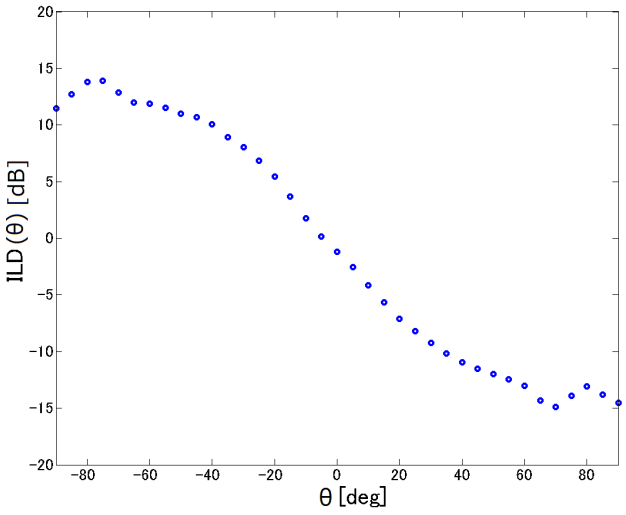

FIGURE 6. $ILD(\theta)$ vs. θ

9. **Listening tests.** To assess the performance of our method, headsets were used to compare two samples of reproduced audio.

1. The left- and right-channel signals of the original 2ch stereo were convolved with HRTFs for angles of $-45°$ and $+45°$, respectively. It should be noted that each HRTF consisted of HRTFs for the left and right ears.
2. The method described above was used to separate 2ch stereo signals into multiple (typically three or four) sources. Each sound source was convolved with HRTFs for the directional angle of the separated source or wider angle. Again, each HRTF consisted of HRTFs for the left and right ears.

Seven subjects listened to the above pair 10 times with a randomized order. The subjects were students of Akita Prefectural University and most of them were working on HRTF related research topics. Therefore, they were familiar with the sound of spatial audio synthesized with HRTFs. They were asked to mark which one in the pair provides better sound images. Just their preference was asked, and no more detailed explanation was provided in the instruction for the evaluation. The results of the paired comparison test (Fig. 7) shows that the preference for Sample 2 (our method) was higher than that for Sample 1 (conventional method), meaning that the proposed source separation algorithm with HRTFs appropriate for headphone playback reproduces highly realistic audio.

10. **Conclusions.** An audio source separation algorithm based on the ratio of the amplitudes of the spectral coefficients of the left and right channels has been developed. Cross-correlation of the synthesized time domain waveforms of the separated source signals is a good indicator for finding an appropriate threshold for grouping spectral coefficients based on their amplitude ratio during the separation process. Tests showed that applying

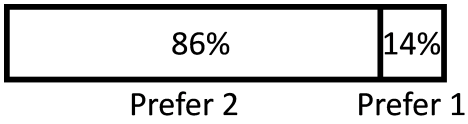

FIGURE 7. Results of paired-comparison test of proposed algorithm

appropriate HRTFs to each separated source signal provides more realistic reproduction quality in headsets than conventional virtual-3D playback does.

REFERENCES

[1] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, 2001.

[2] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.

[3] C. Fevotte, N. Bertinand J.-L. Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis, *Neural Computation*, vol. 21, no. 3, pp. 793-830, 2009.

[4] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, *Signal Processing Magazine*, 2012.

[5] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Deep learning for monaural speech separation, *ICASSP*, 2014.

[6] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,*Acoustical Science and Technology*, vol. 22, no. 2, pp. 149-157, 2001.

[7] E. Vincent, R. Gribonval, and C. Fevotte, Performance measurement in blind audio source separation, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.