TEMPORAL QUANTIZATION OF SPATIAL INFORMATION USING DIRECTIONAL CLUSTERING FOR MULTICHANNEL AUDIO CODING

Shigeki Miyabe,[1] *Keisuke Masatoki,*[2] *Hiroshi Saruwatari,*[2] *Kiyohiro Shikano,*[2] *Toshiyuki Nomura*[3]

[1]The University of Tokyo, Tokyo, Japan, miyabe@hil.t.u-tokyo.ac.jp
[2]Nara Institute of Science and Technology, Nara, Japan, {keisuke-m,sawatari,shikano}@is.naist.jp
[3]NEC Corporation, Kanagawa, Japan, t-nomura@da.jp.nec.com

ABSTRACT

Binaural cue coding, which is a representing low bit-rate coding of multichannel audio, generates large distortion when the audio data have complex spatial image, such as symphony. Such distortion caused by the low frequency resolution of spatial information because BCC quantizes the parameters of localization. In this paper we propose a new coding framework by quantizing the spatial information temporally. The single-channel sum signal is panned to the multiple channels by selecting the prototypes of the spatial filter. Optimization of the prototypes with minimum coding error is given by a k-means-like clustering of the angles whose centroids are given by the first principal components of the covariances in the classes. The efficiency of the proposed coding with high quality is verified both in the objective and subjective evaluations.

Index Terms— Multi-channel audio coding, binaural cue coding, k-means clustering, vector quantization

1. INTRODUCTION

With the recent spread of multichannel audio and diversification of the content delivery form, demand for the low bit-rate codec of multichannel audio is getting higher. While the main target of the conventional audio codecs is single channel signal, binaural cue coding (BCC) has been an epoch-making coding scheme to utilize correlation among channels explicitly [1]. Parametric stereo, an alternative of BCC, is adopted as a basis of MPEG-4 Surround [2]. Also, combining source separation techniques and MPEG-4 Surround framework, the MPEG has started the Spatial Audio Object Coding (SAOC) project to edit the sources freely on the user side, and many activities are conducted [3, 4, 5].

By quantizing the frequency resolution of the correlation among the channels, BCC compresses the multichannel signal into a singe channel signal so-called *sum signal* and low bit-rate parameters so-called *side information*. The side information is a parameterization of the spatial cues related with perception of source localization, and the decoder reconstructs the multichannel signal by panning of the sum signal according to the side information. The side information is composed of three parameters. Two are inter-channel level difference (ICLD) inter-channel time difference (ICTD), which are important for the perception of the sources' directions. The other is inter-channel coherence (ICC), which can express the sizes of the sources. By analyzing these parameters in a certain bandwidth of cochlear filter bank to resemble human auditory system, the number of the parameters are reduced. However, the analysis of such spatial parameters in the filter banks with certain bandwidths corresponds to the assumption that the signal in a filter bank is generated from a source in a certain direction. Thus, BCC degrades the quality of audio signals which are generated by many audio sources, such as symphony, because the underlying sparse source model does not fit to the signals well.

In this paper, we propose a new framework of multichannel audio coding by temporal quantization of the panning information with the frequency resolution maintained. The format is similar to that of BCC with the single channel sum signal and side information. However, the side information is not a perceptual parameterization but quantized vector angle prototypes of the level and phase difference among sources. To optimize the angles with minimum squared-error criterion, we propose a new k-means-like clustering algorithm which uses the principal component analysis in the calculation of the centroid of the angles. With such squared-error minimization, the proposed method realizes the coding with lower bit rate and higher quality than BCC. The effectiveness of the strategy to quantize the spatial information temporally is ascertained in both the objective and subjective evaluations.

2. CONVENTIONAL BINAURAL CUE CODING

Here we review the algorithm of BCC briefly. Note that in the experiments we used a precise implementation of the original paper of BCC [1] although the description here is simplified. The processing of BCC is illustrated in Fig. 1. The C-channel original signal is denoted by the C-dimensional column vector $\boldsymbol{r}(n)$ as

$$\boldsymbol{r}(n) = [r_1(n), \ldots, r_C(n)]^{\mathrm{T}}, \tag{1}$$

where n is the discrete time index and $\{\cdot\}^{\mathrm{T}}$ denotes transposition. Then $\boldsymbol{r}(n)$ is transformed to short-time complex spectrum $\boldsymbol{x}(k,m)$ by frame analysis and fast Fourier transform (FFT), denoted as

$$\boldsymbol{x}(k,m) = [x_1(k,m), \ldots, k_C(k,m)]^{\mathrm{T}}, \tag{2}$$

where $k = 0, \ldots, K-1$ is the discrete frequency index of K-point FFT, and m denote the frame index. The sum signal $s(n)$ is the inverse short-time FFT of the complex sum spectrum $\tilde{y}(k,m)$ given by

$$\hat{y}(k,m) = \alpha(k,m) \sum_{c=1}^{C} x_c(k,m), \tag{3}$$

where $\alpha(k,m)$ is a term to compensate the attenuation caused by cancellation of the channels. By dividing the spectra $\boldsymbol{x}(k,m)$ and $\hat{y}(k,m)$ into cochlear filter bank of about 20 subbands, the three parameters ICLD $\Delta\lambda_c\left(\beta(k),m\right)$, ICTD $\Delta\tau\left(\beta(k),m\right)$ and ICC $\Phi_c\left(\beta(k),m\right)$ for $c = 1, \ldots, C$. Here $\beta(k)$ denotes the subband block index to which the k-th frequency bin belongs. The

This work is partly supported by MIC Strategic Information and Communications R&D Promotion Programme in Japan.

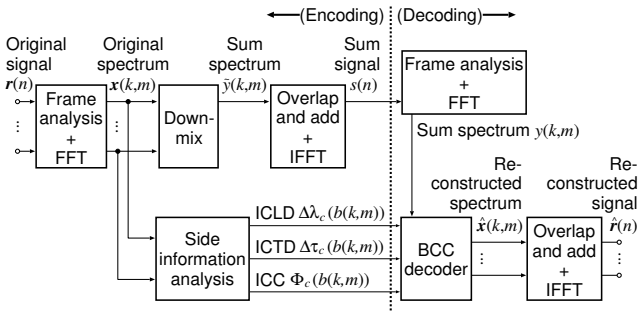

Figure 1: Configuration of binaural cue coding.

analyzed parameters here, $\Delta\lambda_c\left(\beta(k),m\right)$, $\Delta\tau\left(\beta(k),m\right)$, and $\Phi_c\left(\beta(k),m\right)$ are the level ratio, the sample time difference and the correlation coefficient between $x_c(k,m)$ and $\tilde{y}(k,m)$ in the $\beta(k)$-th subband block of the m-th frame.

On the decoder side, the sum signal $s(n)$ is converted to short-time spectrum $y(k,m)$ again, and reconstruction $\hat{\boldsymbol{x}}(k,m)$ of the original multichannel spectrum $\boldsymbol{x}(k,m)$ is obtained as time-varying panning of $y(k,m)$ to the C channels as

$$\hat{\boldsymbol{x}}(k,m) = \left[\hat{x}_1(k,m),\ldots,\hat{x}_c(k,m)\right]^{\mathrm{T}} = \boldsymbol{g}(k,m)y(k,m), \quad (4)$$

$$\boldsymbol{g}(k,m) = \left[g_1(k,m),\ldots,g_C(k,m)\right]^{\mathrm{T}}, \quad (5)$$

where $\boldsymbol{g}(k,m)$ is a multichannel filter to pan the sum spectrum $\boldsymbol{y}(k,m)$, given as

$$g_c(k,m) = \left\{\Delta\lambda_c\left(\beta(k),m\right) + \left(1-\Phi_c(\beta(k),m)\right)\delta\left(k\right)\right\} \cdot \exp\left[\frac{\jmath k}{2\pi K}\Delta\tau\left(\beta(k),m\right)\right], \quad (6)$$

where $\delta(k)$ is a constant generated by uniform random function and $\jmath = \sqrt{-1}$. Thus, $\boldsymbol{g}(k,m)$ is the combination of the intensity panning with ICLD and the phase panning with ICTD distorted by ICC. Finally, the reconstructed signal $\boldsymbol{r}(\hat{n})$ is obtained by frame overlap and add of the inverse FFT (IFFT) of $\boldsymbol{x}(k,m)$.

As a matter of fact of the relation in Eq. (6), the panning with single pair of ICLD and ICTD in a subband of a certain width of frequency assumes sparseness of the sources and signal arrive from only one source in each subband block. Although the model mismatch is relaxed with decorrelation by ICC, large model mismatch causes significant distortion.

3. PROPOSED MULTICHANNEL AUDIO CODING

3.1. Decoding and data format

The data used in the proposed decoding system is composed of the single-channel sum signal $s(n)$ and the side information. One component of the side information is C prototypes of the L-channel spatial filter $\boldsymbol{h}(k,l), l=1,\ldots,L$ in each frequency band;

$$\boldsymbol{h}(k,l) = \left[h_1(k,l),\ldots,h_C(k,l)\right]^{\mathrm{T}}, \ \|\boldsymbol{h}(k,l)\| = 1. \quad (7)$$

Unlike the memoryless coding/encoding of BCC, a certain number of frames in a time segment share the prototype $\boldsymbol{h}(k,l)$. By switching the L spatial prototypes $\boldsymbol{h}(k,l)$, time-variant spatial filtering similar to Eq. (4) is realized in each independent frequency band. The information of which prototype to use is recorded by $\log_2 L$ bits index $I(k,m)$. Thus, the multichannel signals are expressed by a single channel sum signal $s(n)$, the C-channel and

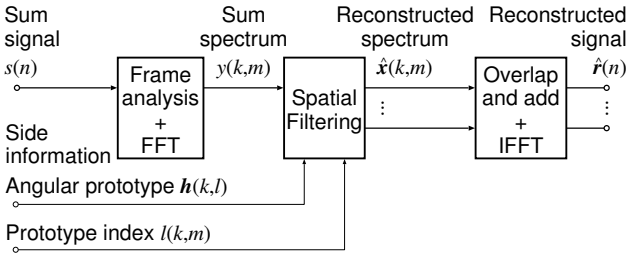

Figure 2: Configuration of the proposed decoding process.

L-frame signal length of filter $\boldsymbol{h}(k,l)$, and indices $I(k,l)$ of few bits.

The process of decoding is illustrated in Fig. 2. First, the sum signal $s(n)$ is transformed to short-time complex spectrum $y(k,m)$ by devision into frames and FFT analysis. Then we obtain the decoded C-channel spectrum $\hat{\boldsymbol{x}}(k,m)$ by selecting $\boldsymbol{h}(k,l)$ according to $I(k,m)$ and filtering $y(k,m)$, as

$$\hat{\boldsymbol{x}}(k,m) = \boldsymbol{h}(k,I(k,m))y(k,m). \quad (8)$$

Finally the reconstructed signal $\hat{\boldsymbol{r}}(n)$ of the original $\boldsymbol{r}(n)$ is obtained by IFFT and frame overlap and add of $\hat{\boldsymbol{r}}(n)$.

3.2. Encoding

A session of coding is conducted with a given segment of the signal. First, similarly to BCC, we obtain the short-time FFT spectrum $\boldsymbol{x}(k,m)$ of the original C-channel signal $\boldsymbol{r}(n)$. Next, all the frames of $\boldsymbol{x}(k,m)$ in a segment is analyzed by clustering of the angles of the vectors, described in Sect. 3.3, and obtain L vectors $\boldsymbol{h}(k,l)$, $l=1,\ldots,L$, which are prototypes of the angles of $\boldsymbol{x}(k,m)$ to describe the relations of the level and phase differences among the channels. Then, the amplitude $\tilde{y}(k,m)$ of $\boldsymbol{h}(k,l)$ to give the reconstruction of $\boldsymbol{x}(k,m)$ with the minimum squared error is given as the following projection using the condition $\|\boldsymbol{h}(k,l)\| = 1$;

$$\tilde{y}(k,m) = \arg\min_a \|\boldsymbol{x}(k,m) - \boldsymbol{h}(k,I(k,m))a\|^2 = \boldsymbol{h}(k,I(k,m))^{\mathrm{H}}\boldsymbol{x}(k,m). \quad (9)$$

Finally, by the frame overlap and add of the IFFT of $\tilde{y}(k,m)$, we obtain the sum signal $s(n)$. Note that the sum spectrum $\tilde{y}(k,m)$ at the encoder is different from the one $y(k,m)$ at the decoder. Although the smoothing in $y(k,m)$ blurs the spatial image, it also reduces the musical noise in the decoded signal $\hat{\boldsymbol{r}}(n)$ because the rapid switch of the spatial prototype sometimes generate isolated peaks in $\tilde{y}(k,m)$. Hereafter $\tilde{y}(k,m)$ is refered to as *prior sum spectrum.*

3.3. Prototype optimization with minimum coding error

In the signal reconstruction of the proposed coding, while the amplitude and phase of the signal is given by the scalar sum spectrum, the level and phase differences of the channels are expressed by the prototype vector $\boldsymbol{h}(k,l)$ of panning. Here we propose a k-means-like algorithm to optimize the angle of the vector $\boldsymbol{h}(k,l)$ with minimum-squared-error criterion. The squared error of the coding is written in the frequency domain as

$$E(k) = \sum_{\forall m}\|\boldsymbol{x}(k,m) - \hat{\boldsymbol{x}}(k,m)\|^2 = \sum_{\forall m}\|\boldsymbol{x}(k,m) - \boldsymbol{h}(k,I(k,m))y(k,m)\|^2. \quad (10)$$

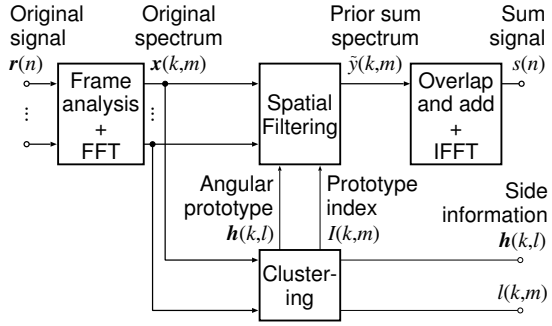

Figure 3: Configuration of the proposed encoding process.

Here, to simplify the processing on the decoding side, by ignoring the effect of window, we substitute $y(k,m)$ with $\tilde{y}(k,m)$ the prior as

$$\begin{aligned}\hat{E}(k) &= \sum_{\forall m} \|\boldsymbol{x}(k,m) - \boldsymbol{h}(k,I(k,m))\tilde{y}(k,m)\|^2 \\ &= \sum_{l=1}^{L} \sum_{m\in\Omega_l(k)} \hat{E}_l(k), \end{aligned} \tag{11}$$

where $\Omega_l(k)$ is a set of time-frequency slots (k,m) where the l-th prototype $\boldsymbol{h}(k,l)$ is selected, and

$$\hat{E}_l(k) = \sum_{m\in\Omega_l(k)} \|\boldsymbol{x}(k,m) - \boldsymbol{h}(k,I(k,m))y(k,m)\|^2 \tag{12}$$

is the total squared error of the l-th class. Since the spectrum $\tilde{y}(k,m)$ is given automatically by $\boldsymbol{h}(k,m)$ and $\Omega_l(k)$, the problem is to minimize $\hat{E}(k)$ over $\boldsymbol{h}(k,m)$ and $\Omega_l(k)$, as

$$\underset{\boldsymbol{h}(k,l),\Omega_l(k)}{\arg\min}\ \hat{E}(k) = \underset{\boldsymbol{h}(k,l),\Omega_l(k)}{\arg\min} \sum_{l=1}^{L} \hat{E}_l(k). \tag{13}$$

This objective function corresponds to the k-means clustering of the centroid $\boldsymbol{h}(k,l)$ in the sense of angle to minimize the sum of the inner-class errors $\hat{E}_l(k)$. In the following, we show that the first principal component of the l-th class gives $\boldsymbol{h}(k,l)$ to minimize $E_l(k)$, and derive the k-means clustering algorithm of angles.

Given a C-dimensional vector $\boldsymbol{b}$ with unit norm, i.e., $\|\boldsymbol{b}\| = 1$, the minimum squared error $\hat{e}(k,m;\boldsymbol{b})$ in the time-frequency slot (k,m) with the prototype $\boldsymbol{h}(k,m)$ is given using the projection in Eq. (9) as

$$\begin{aligned}\hat{e}(k,m;\boldsymbol{b}) &= \min_{a} \|\boldsymbol{x}(k,m) - \boldsymbol{b}a\|^2 \\ &= \left\|\boldsymbol{x}(k,m) - \boldsymbol{b}\boldsymbol{b}^{\mathrm{H}}\boldsymbol{x}(k,m)\right\|^2 \\ &= \|\boldsymbol{x}(k,m)\|^2 - \boldsymbol{b}^{\mathrm{H}}\boldsymbol{x}(k,m)\boldsymbol{x}(k,m)^{\mathrm{H}}\boldsymbol{b}. \end{aligned} \tag{14}$$

Thus, the total $\hat{E}_l(k;\boldsymbol{b})$ of the squared error is given by

$$\begin{aligned}\hat{E}_l(k;\boldsymbol{b}) &= \sum_{m\in\Omega_l(k)} \hat{e}(k,m;\boldsymbol{b}) \\ &= \sum_{m\in\Omega_l(k)} \|\boldsymbol{x}(k,m)\|^2 - \sum_{m\in\Omega_l(k)} \boldsymbol{b}^{\mathrm{H}}\boldsymbol{x}(k,m)\boldsymbol{x}(k,m)^{\mathrm{H}}\boldsymbol{b}. \end{aligned} \tag{15}$$

Since the first term in Eq. (15) is constant, the prototype $\boldsymbol{h}(k,l)$ to minimize inner-class squared error is given by the first principal component of the inner-class covariance matrix, given as

$$\begin{aligned}\boldsymbol{h}(k,l) &= \underset{\boldsymbol{b}\,|\,\|\boldsymbol{b}\|=1}{\arg\min}\ \hat{E}_l(k;\boldsymbol{b}) \\ &= \underset{\boldsymbol{b}\,|\,\|\boldsymbol{b}\|=1}{\arg\min} \left[- \sum_{m\in\Omega_l(k)} \boldsymbol{b}^{\mathrm{H}}\boldsymbol{x}(k,m)\boldsymbol{x}(k,m)^{\mathrm{H}}\boldsymbol{b} \right] \\ &= \underset{\boldsymbol{b}\,|\,\|\boldsymbol{b}\|=1}{\arg\max}\ \boldsymbol{b}^{\mathrm{H}} \left[\sum_{m\in\Omega_l(k)} \boldsymbol{x}(k,m)\boldsymbol{x}(k,m)^{\mathrm{H}} \right] \boldsymbol{b}. \end{aligned} \tag{16}$$

Using the error minimization under the fixed class membership described above and the class update of k-means algorithm, the iterative optimization of the objective in Eq. (13) is given as:

[Step 1] Set initial prototypes $\boldsymbol{h}(k,l)$ randomly, and calculate membership $\Omega_l(k)$ to minimize the projection error.

[Step 2] Update each prototype $\boldsymbol{h}(k,l)$ by substituting with the first principal component of the covariance matrix of the l-th class.

[Step 3] Update the membership $\Omega_l(k)$ to minimize the error.

[Step 4] Go back to the step 2 until convergence.

To reduce the frequency discontinuity of the sum signal $s(n)$, the average of the phase of $\boldsymbol{h}_l(k,l)$ is equalized. Also, the prototype index $I(k,m)$ is given by $m \in \Omega_l(k)$ as

$$I(k,m) = l \;\big|\; m \in \Omega_l(k). \tag{17}$$

3.4. Segment optimization for temporal quantization

While BCC is memoryless coding algorithm, the proposed method share a part of side information in a certain length of segment. Since the optimal spatial prototype $\boldsymbol{h}(k,l)$ depends on the auditory scene, i.e., transitions of instrumental arrangement in a music piece and set changes in a movie. Thus, if the switch of the prototypes does not match the change of the auditory scene, the coding becomes inefficient. The segmentation can be optimized with the dynamic programming problem to minimize squared-error criterion similarly to the coding algorithm. First we specify the number T of the segments, and make the combinations of $T-1$ segment partitions on the certain fixed intervals of frames. Then we try the coding for each segmentation, and chose the one with minimum coding error. Here we discussed the off-line segment optimization of long data with rich computation but the reduction of the computational cost and online segmentation is a remaining problem. Note that, if we can use large prototype number L, the degradation of quality is low even without the segmentation.

3.5. Value Compared with BCC

The following section demonstrats that the proposed method can compresses the multichannel audio signal with higher sound quality and lower bit rate than BCC. Also, the proposed method can be extended to multichannel sum signal coding easily by using multiple eigenvalues in the PCA. However, the proposed method has several drawbacks in its computation. The current proposed enconding requires long buffer for the training of angular prototypes, and its on-line processing is difficult. In addition, large computation is required in the dynamic programming for the optimization of the segmentation. Also, directional clustering has to solve eigenvalue problem. The proposed decoding has to stock the spatial prototypes and memory requirement is slightly higher.

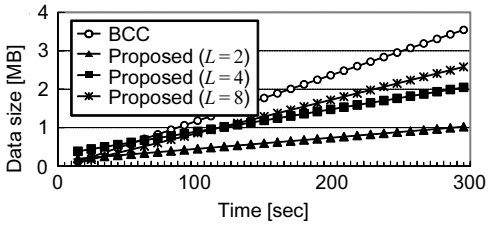

Figure 4: Comparison of the encoded data sizes.

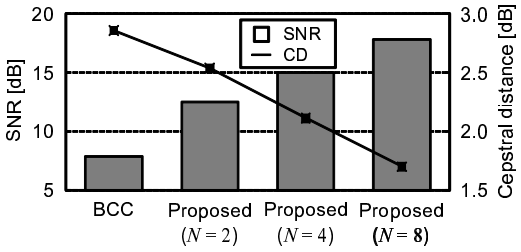

Figure 5: Results of SNR and CD in BCC and proposed method.

4. EXPERIMENTS AND RESULTS

In this experiment, we evaluate performances of our coding for stereo music signals chosen from commercially available compact discs, e.g., 10 pieces of full length each from classical music, popular music, and jazz music. The sampling frequency is 44.1 kHz and the resolution is 16 bits. The filter length and frame size of the proposed method is 4096 points with 4006-point frame shift. The window is combination of rectangular and Han window; the rectangular window with 3916 points is inserted in the middle of the von Hann window of 120 points, and 30 zeros are padded in the both ends of the window. We compared three numbers of prototypes of $L = 2, 4, 8$. We segmented each music piece into 10 segments with $L = 2, 4$, and did not segment each piece with $L = 8$. As shown in Fig. 4, all parameter settings of the proposed method have lower bit rates than BCC.

To evaluate the signal reconstruction quality objectively, we used to scores; signal-to-noise ratio (SNR) and cepstral distance (CD) [6], which are commonly used in speech coding. As a subjective evaluation, XAB test is conducted. Test subjects are five males and two females with normal audibility.

Figure 5 shows comparison of BCC and the proposed method from the viewpoint of averaged SNR and CD over all 30 music pieces. It is confirmed that the proposed method can reconstruct the signals with less distortion than BCC, and the distortion is reduced according to the increases of the number L of bases.

Next, we show the results of the subjective evaluation in Fig. 6. The higher score of the proposed method with any parameter setting is significant with the error bars of 95% confidence interval.

To investigated the robustness of the proposed method against unclarity of source sparseness, we compared the coding quality of three music pieces with two instruments and 13 pieces with more instruments. The number of the prototypes of the proposed method is $L = 2$, and its filter length is 1024 taps, whose bit rate is lower than that of BCC. Figure 7 shows the results of the comparison. If the number of the sources is more than two, distortion of BCC increases greatly. In contrast, the proposed method is robust regardless of the number of sources.

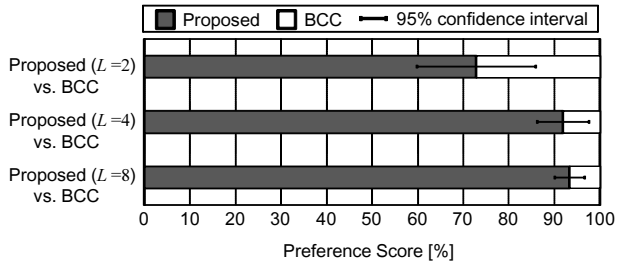

Figure 6: Results of subjective evaluation for each kind of sources.

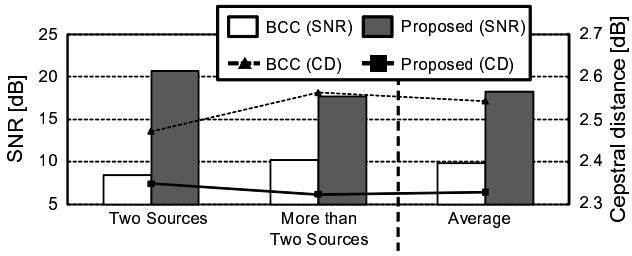

Figure 7: Results of robustness against non-sparseness.

5. CONCLUSIONS

By prototyping the vectors to express level and phase differences among channels in narrow frequency bands, we proposed a new coding framework of multichannel audio signals. The multichannel signals are reconstructed from a single channel audio data and low bit-rate information composed of the prototypes and its indices. The coding error corresponds to the projection error of the prototypes to the original signals, and its squared-error minimization is given by the k-means clustering with its centroid in the sense of the angle given by the first principal component of the inner-class covariance. Since the assumption about the sparseness among sources is weaker than BCC, it is more efficient and robust coding. The effectiveness of the proposed method is ascertained both in the objective and subjective evaluations.

6. REFERENCES

[1] F. Baumgarte and C. Faller, "Binaural Cue Coding—Part II: Psychoacoustic Fundamentals and Design Principles," *IEEE Trans. Speech Audio Process.,* vol. 11, no. 6, pp. 520–531, 2003.

[2] H. Purnhagen, "Low complexity parametric stereo coding in MPEG-4," *Proc. DAFx*, pp. 163-168, 2004.

[3] J. Herre, S. Disch, "New Concepts in Parametric Coding of Spatial Audio: From SAC to SAOC," *Proc. ICME,* pp. 2–4, 2007.

[4] S. Miyabe, T. Mihashi, T. Takatani, H. Saruwatari, K. Shikano, T. Nomura, "Compressive coding of stereo audio signals extracting sparseness among sound sources with independent component analysis," *Proc. WASPAA,* pp.331–334, 2007.

[5] Y. Haraguchi, S. Miyabe, H. Saruwatari, K. Shikano, T. Nomura, "Source-oriented localization control of stereo audio signals based on blind source separation," *Proc. ICASSP,* pp.177–180, 2008.

[6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition,* Upper Saddle River, NJ: Prentice Hall, 1993.