MULTICHANNEL-TO-WAVE FIELD SYNTHESIS UPMIXING TECHNIQUE BASED ON SOUND SOURCE SEPARATION

KEUNWOO CHOI, TAE JIN PARK, JEONGIL SEO, AND KYEONGOK KANG

Electronics and Communications Research Institute (ETRI), Daejeon, Republic of Korea
{gnu, inctrl, seoji, kokang}@etri.re.kr

In this research, we propose a multichannel-to-loudspeaker array upmixing algorithm. To take advantage of a loudspeaker array, we introduce an approach based on audio source separation. During the analysis phase, multichannel signals are separated into element signals in stereo format through stereo-channel extraction. They are then separated into source signals by the Laplacian Mixture Models of the features from the stereo signals. During the synthesis phase, the sources are rendered as virtual sources through Wave Field Synthesis, or as focused sources. Subjective tests show that the proposed algorithm outperforms the comparison algorithm in terms of localization quality.

INTRODUCTION

A loudspeaker array is considered an advanced configuration for the reproduction of immersive sound fields, providing an enlarged listening spot and a convenient set-up method. These qualities give loudspeaker arrays an edge over discrete multichannel sound reproduction systems such as 5.1 or 7.1 channel surround, which provide a small sweet spot with a complicated installation process. Using an object-based sound format, loudspeaker arrays can locate the sound sources at the correct positions to the listeners. Although Spatial Audio Object Coding was proposed in this regard [1], most commercial audio content such as TV shows or movie DVDs are mixed in a channel-based sound format when delivered to the users. To reproduce such content effectively through loudspeaker arrays, an upmixing process should first be performed.

One simple algorithm for the playback of 5.1-channel audio through a loudspeaker array might be to create imaginary sources at the position of the 5.1 loudspeakers with the corresponding channel signals. This simplest upmixing method, however, does not improve the problem of a narrow sweet-spot for a discrete loudspeaker configuration. Another approach, as proposed by YAMAHA, is to create multiple beams in different directions, and has turned out to have some market success. Still, localizations may remain far from the reference positions. In other words, this approach does not guarantee the original sound scene.

In this research, we propose another approach for the playback of a 5.1-channel audio signal using a frontal loudspeaker array. A 5.1-channel layout is selected as the source because it is the most widespread multichannel sound format on the market. We selected a single and frontal loudspeaker array as the target system of the upmixing algorithm, considering the practical aspects for a home installation.

We also adopted a source separation to achieve a correct and clear localization. Based on this same motivation, Cobos et al. and Kamado et al. introduced algorithms using source separation [2, 3] and [4], although multichannel sources were not considered, which makes achieving both separation and reproduction more complex.

We describe the background of the problem and the proposed algorithm in Section 1. In Section 2, the proposed algorithm is described in detail. An assessment is then given in Section 3. Finally, we provide some concluding remarks regarding this research in Section 4.

1 BACKGROUND

1.1 Multichannel Surround Format

Among various multichannel audio formats, the 5.1-channel surround format has been the most popular. 5.1-

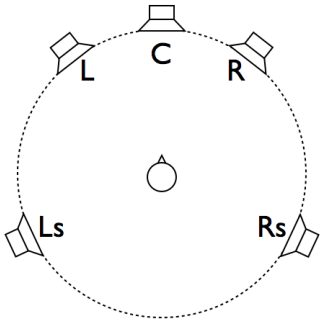

Figure 1. The standard layout of 5.1-channel surround loudspeakers

channel surround comprises channels for the front-left (L), and front-right (R), front-centre (C), low-frequency enhancement (Lfe), surround-left (Ls), and surround-right (Rs). Among the existing multichannel sound formats, 5.1-channel surround is most popular layout for commercial and home theatres. The standard setup for 5.1-channel surround was clearly addressed by ITU-R. Bs.775 [5].

When compared to stereo sound, an additional centre channel helps the localization in front of the listeners and stabilizes the sound images. For cinema sound, this channel is mainly used for dialogue, while various instruments are mixed into the music. Surround channels (left and right surround) are commonly used for an enhancement of the immersive effect. As a result, surround channels include ambience sources or audience applause most of the time, which is referred to as a *direct/ambient approach* [6]. In contrast, for an *in-the-band approach*, these channels also contain direct sound components to localize behind the listeners. For movies, they are often combined with front channels to produce moving sources.

The signals are generally mixed in a studio with a pan-pot, which uses a gain panning technique. With five correctly installed loudspeakers, a listener located at the sweet spot can perceive the sound source localizations on the horizontal plane.

However, installing five loudspeakers and a woofer at the appropriate positions is very inconvenient and complex for the majority of cases. In addition, localizations for out-of-sweet-spot listeners are very unnatural and biased. For this reason, 5.1-channel surround systems have failed for home use, despite their many useful features.

1.2 Applications of Loudspeaker Arrays

1.2.1 WaveField Synthesis

WaveField Synthesis (WFS), an innovative reproduction algorithm, was proposed by Berkhout et al. [7]. The theory approximates a 3D space into a 2D plane, and then conducts sampling in the spatial domain to compute the rendering coefficients, i.e., the gains and delays for each loudspeaker. Using these coefficients, it is possible to reproduce the same sound field as the real sound source within a certain area enclosed by the loudspeakers for frequencies under the aliasing frequency, which depends on the distance between adjacent loudspeakers. In short, WFS can create imaginary sources.

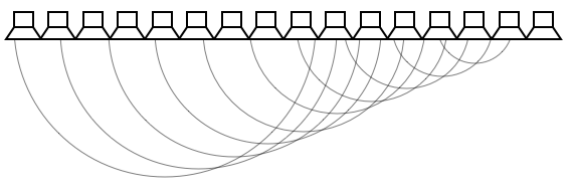

Figure 2. An illustration of WFS: Generation of a plane wave with a linear loudspeaker array

We can choose a sound field produced by an imaginary source as either a plane or spherical wave. Using a spherical wave, an imaginary source can be located at the same position for listeners in the sweet-spot area. This also means that we can control the sense of distance, as well as the sense of direction [8]. On the other hand, plane waves can be used when all listeners should recognize the angle of sound source constantly within the listening area.

With a linear array, i.e., not an enclosing array, sound fields from imaginary sources located behind the array can be constructed.

1.2.2 Focused Sources

Focusing, or near-field beamforming, is a method for creating an imaginary source between a listener and the loudspeaker arrays. A time-reversal technique is used to compute the gains and delays of the loudspeakers [9].

Focused sources are generally used to generate virtual sources between the listeners and loudspeaker array. Owing to such practical limitations as spatial aliasing and truncation effects, perceptual aspects should be considered when generating the focused sources [10]. Moreover, in [11], focused sources are located at walls to generate virtual loudspeakers. An example of this is illustrated in Figure 3, where the absorption coefficient of the wall is 0.05.

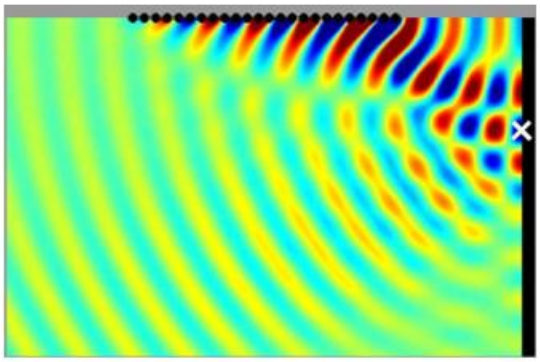

Figure 3. An example of a sound field generated by a focused source at a wall[1]

[1] Figure courtesy of Nara Hahn, University of Rostock.

1.3 Sound Source Separation Algorithm

A lot of researches on the problems of audio source separation have been conducted. To be specific, these problems are defined by such variables as the number of observations, number of estimated sources, and the mixing techniques.

Researches have focused on underdetermined cases in which the number of mixtures is smaller than the number of sources [12]. In such cases, the parameter values of some parametric source models are typically estimated. The sources are then separated by masking in the time-frequency domain [13], through a calculation and multiplication of the unmixing matrices, and so on.

One approach for estimating the parameters is to use the location of the sources in the mixture [14]. During the mixing process, the audio sources are 'located' using various kinds of panning techniques, and distributed at different positions between the loudspeakers. These result in an Inter-channel Intensity Difference (IID) and Inter-channel Time Difference (ITD) in the mixture, and we use the IID to separate the sources in this research.

1.3.1 ADRess

Barry introduced ADRess (Azimuth Discrimination and Resynthesis), which is a source-separation technique for stereo mixtures that assumes an instantaneous stereo mixing [15]. First, a stereo signal is modelled through Eqs. (1) and (2), where the indices of the sources $s(t)$ are omitted for simplicity. The intensity ratio of the sources is then calculated using Eq. (3).

$$L(t) = \sum P_l S(t) \tag{1}$$

$$R(t) = \sum P_r S(t) \tag{2}$$

$$g = \frac{P_l}{P_r} \tag{3}$$

With g, a source can be cancelled out by $L(t)-gR(t)$. In this regard, the sources can be separated from a mixture in the time-frequency domain, where $L(t)$ and $R(t)$ are substituted with the STFT coefficients from $L(t)$ and $R(t)$. In short, an IID-related feature is extracted from a stereo mixture.

To extract a certain source from a mixture, the concept of an *azimuth subspace width* is adopted in ADRess. The components within an azimuth subspace width are regarded as parts of the same source.

1.3.2 Laplacian-Mixture Models

A more generalized method for clustering features than the use of an azimuth subspace width is to model the distribution of features. Gaussian Mixture Models are commonly used for this as one can easily take advantage of existing computational techniques [16, 17].

In many cases, however, the feature distributions are more like super-Gaussian or Laplacian. In [18], Mitianoudis and Stathaki proposed an Expectation-Maximization (EM) algorithm to estimate the parameters of the Laplacian-Mixture Models (LMM).

An LMM is defined in Eq. (4), where a_i , θ_i , and c_i are the weights, centers, and widths of each Laplacian, and N and T are the number of Laplacians and features. The update rules are then proposed through Eqs. (5) and (6), where $p(i \mid \theta_n)$ indicates the probability of θ_i belong to the *i*-th Laplacian.

$$p(\theta) = \sum_{i=1}^{N} a_i c_i e^{-2c_i|\theta-\theta_i|} \tag{4}$$

$$\theta_i^{+} \leftarrow \frac{\sum_{n=1}^{T} \frac{\theta_n}{|\theta_n - \theta_i|} p(i \mid \theta_n)}{\sum_{n=1}^{T} \frac{1}{|\theta_n - \theta_i|} p(i \mid \theta_n)} \tag{5}$$

$$c_i^{+} \leftarrow \frac{\sum_{n=1}^{T} p(i \mid \theta_n)}{2\sum_{n=1}^{T} |\theta_n - \theta_i| \, p(i \mid \theta_n)} \tag{6}$$

After an estimation of the parameters, the time-frequency bins are clustered into the respective sources to which they belong. As each Laplacian is overlapped, the threshold method is optional, and depends on the choice of continuous or binary mask type, or whether artifacts or interferences are of focus.

2 ALGORITHM

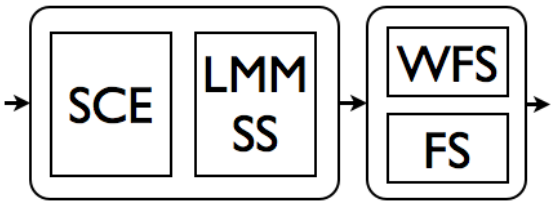

Figure 4. A block diagram of the proposed algorithm

The proposed algorithm consists of analysis and synthesis phases. The analysis phase consists of two stages, Stereo-Channel Extractions (SCE) and source separation, based on Laplacian-Mixture Models (LMM SS). In SCE, 5.1-channel audio signals are separated into six element signals, each in stereo format, and LMM-based source separations from the stereo mixtures are performed in an LMM SS.

During the synthesis phase, WFS or focused sources are chosen to reproduce the sound field of the separated signals. For the sources behind a loudspeaker array,

WFS renders virtual sources that generate a spherical wave from the estimated source position. Focused sources are used for the other sources.

2.1 Signal Model

2.1.1 Signal Model of 5.1-Channel Content

We assume that 5.1-channel signals are mixed with pan-pot, which uses two adjacent loudspeakers to localize the sound sources. Therefore, the signal can be described as a sum of six-channel signals, as shown in Eq. (7), which have four silent or null channels and two audio content channels. The signals, $s_{5.1}{}^{p}(t)$, are 6-by-1 vectors, where p is an index for the possible pairs in the assumption. In other words, $s_{5.1}(t)$ is a virtual stereo-channel audio signal in a 6-channel format.

$$s_{5.1}(t) = \sum_{p=1}^{6} s_{5.1}{}^{p}(t) \tag{7}$$

According to the mixing model, there are six kinds of possible signals, as shown in Table 1. There are five signals with adjacent channel pairs, as p = 2 - 6. Pair 1, which consists of channels 1 (left) and 2 (right), should be included, as the sources can be panned with L and R loudspeakers, excluding channel 3 (center).

p						
1	$S_L{}^1(t)$	$S_R{}^1(t)$	0	0	0	0
2	$S_L{}^2(t)$	0	$S_C{}^2(t)$	0	0	0
3	0	$S_R{}^3(t)$	$S_C{}^3(t)$	0	0	0
4	$S_L{}^4(t)$	0	0	0	$S_{Ls}{}^4(t)$	0
5	0	$S_R{}^5(t)$	0	0	0	$S_{Rs}{}^5(t)$
6	0	0	0	0	$S_{Ls}{}^6(t)$	$S_{Rs}{}^6(t)$

Table 1. Six signals organizing 5.1-channel audio signals

2.1.2 Signal Model of Element Signals

The six signals mentioned above can be regarded as a stereo audio signal, excluding the four null channels. We call these stereo signals *element signals*, which are the basic elements of a 5.1-channel audio signal. For an element signal, we assume an instantaneous mixing as

$$s_i{}^{p}(t) = \sum_j g_{ij} x_j(t), \tag{8}$$

where $i = 1 \text{ or } 2$, $p = 1-6$, and $j \geq 1$.

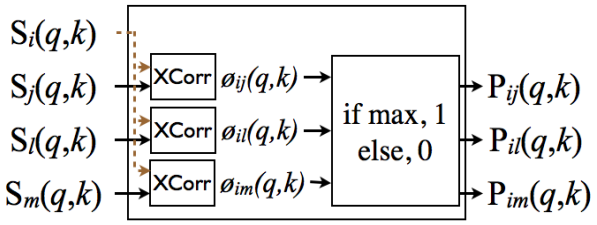

Figure 5. A block diagram of the Stereo-Channel Extraction stage

2.2 Source Separation

The analysis phase consists of stereo-channel extraction (SCE) and LMM-based source separation.

2.2.1 Stereo-Channel Extraction

The SCE is a process for resolving the ambiguity in each channel of a 5.1-channel signal. Based on the proposed mixing model, different components from stereo-channel signals are mixed in each channel. In this stage, for example, from the left channel signal, $S_L{}^1(t)$, $S_L{}^2(t)$, and $S_L{}^3(t)$ should be separately extracted.

These extractions are performed using binary masks determined from comparing the cross-correlations of each of the other channels. First, we perform a Short-Time Fourier Transform (STFT) and obtain time-frequency representation $S_i(q,k)$, where q and k denote the frequency and time index. We then calculate the cross-correlation (Eq. (9)) between the channels of the possible element signals. For example, when i = 1, i.e., analyzing the left channel, j, l, and m are set to 2, 3, and 5, respectively, as the left channel consists of components from elements 1, 2, and 4. In another example, if i = 5, we compute the correlation using only j = 1 and l = 6.

$$\phi_{ij}(q,k) = \lambda S_i(q,k) S^*{}_j(q,k) + (1-\lambda)\phi_{ij}(q-1,k) \tag{9}$$

By comparing the cross-correlation values, masks for SCE are obtained through Eq. (10), where q and k, the time and frequency indices, are omitted for the simplicity. The mask value $P_{ij}(q,k)$ in a time-frequency bin is assigned to 1 if, among the j-th, l-th, and m-th channels, the j-th channel is the one most correlated with the i-th channel. As a result, through Eq. (11), we can calculate $S_{j|i}(q,k)$, which represents a component from the i-th channel and is mostly correlated with the j-th channel.

$$P_{ij} = \begin{cases} 1 \text{ if } \phi_{ij} = \max(\phi_{ij}, \phi_{il}, \phi_{im}) \\ 0 \text{ else} \end{cases} \tag{10}$$

$$S_{j|i}(q,k) = S_i(q,k) \times P_{ij}(q,k) \tag{11}$$

Through masking, six element signals with a stereo-channel are estimated, as shown in Table 2, where the elements are listed in the same order as in Table 1. Finally, they are reconstructed into a time-domain signal through an inverse-STFT.

#	Channel 1	Channel 2
1	$\mathbf{S}_{2\|1}$	$\mathbf{S}_{1\|2}$
2	$\mathbf{S}_{3\|1}$	$\mathbf{S}_{1\|3}$
3	$\mathbf{S}_{3\|2}$	$\mathbf{S}_{2\|3}$
4	$\mathbf{S}_{5\|1}$	$\mathbf{S}_{1\|5}$
5	$\mathbf{S}_{6\|2}$	$\mathbf{S}_{2\|6}$
6	$\mathbf{S}_{6\|5}$	$\mathbf{S}_{5\|6}$

Table 2. Estimated stereo signal elements using stereo-channel extractions

2.2.2 LMM-Based Source Separation

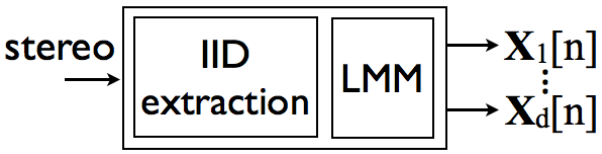

Figure 6. A block diagram of the LMM-based source separation stage

LMM-based source separation is performed as explained in Section 1.3.2. As a result, the sources and their positions for each frame are estimated during this stage.

In this research, we assume four sources in a stereo element. The update equations are iterated 40 times in each frame. In addition, the parameters resulting from the (q-1)-th frame are used as the initial values of the estimation of the q-th frame. This resolves the permutation problem of the source separation process. A hard threshold is employed during the clustering. For the position information, the means of each Laplacian are used during the synthesis phase.

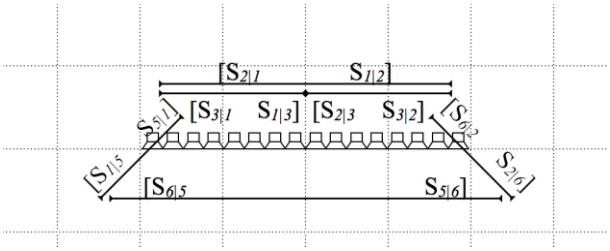

Figure 7. Illustration of virtual sources with WFS and focused sources

2.3 Synthesis

We adopt WFS to render the element signals 1–3, as they can be regarded as sources behind the loudspeaker array. For example, they can be rendered as shown in Figure 7, where six lines are associated with each element signal. The separated sources from the element signals are located on the associated lines.

However, the other element signals related with the surround channels raise a problem of how to synthesize the sources not behind the loudspeaker array. We use the focused source technique to reproduce element signals 4–6, as shown in Figure 7.

3 EVALUATION

3.1 Overview

Subjective tests were performed to evaluate the performance of the proposed algorithm. Three audio excerpts were selected, i.e., one music excerpt and two movie excerpts. They were sampled at 48,000 Hz and have a bit depth of 16. The STFT was performed using hamming windows, 43 ms frames, and a 43 ms hop length. Six experienced participants evaluated the listening room at the Electronics and Telecommunications Research Institute.

We compared the proposed algorithm with a comparison algorithm and 5.1-channel reference. The layouts of the proposed and comparison algorithms are illustrated in Figure 7 andFigure 8, where the grid has a length of 50 cm. For the reference system, five loudspeakers were installed as shown in Figure 1 with a radius of 1.5 m. For the proposed algorithm, the separated signals were reconstructed at the positions shown in Figure 7. For the comparison algorithm, virtual sources were rendered at the different positions shown in Figure 8, considering the 5.1-channel signals as five virtual sources. As the perception of the focused sources can be significantly affected by the precedence effect when the sources are located behind the listeners, the virtual sources for the surround channels are located at the front side.

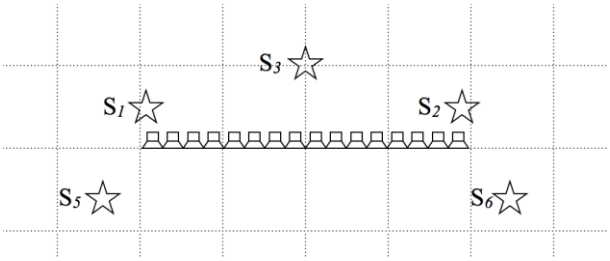

Figure 8. The layout of the comparison algorithm

A custom-made 32-channel linear loudspeaker array with a gap of 0.061 m was used to reproduce the sound. The array was made using two-inch full-range drivers from *Vifa*. For a reproduction of the reference signal, we used *ELAC 330 CE* passive loudspeakers. The reference signals were filtered with the impulse response of the driver of the loudspeaker array to minimize the difference in the frequency response.

3.2 Off-Center Configuration

For the experiments, we set the standard ITU 5.1-surround loudspeaker layout as the reference system. Each participant sat at the sweet-spot, i.e., at the center of x- and y-axes. As one of the merits a loudspeaker array is an enlarged sweet-spot, we designed this test to compare the performances of the off-center listening condition.

As a result, we moved the loudspeaker array 0.5 m to the left, as shown in Figure 9. This makes the listener remain at the sweet-spot of the 5.1 loudspeakers, but also off-center from the loudspeaker array. During the localization tests, the participants first listened to a reference signal, based upon which they are asked to imagine a reference sound scene at the sweet-spot. They then evaluated how correctly the loudspeaker array re-creates the sound scenes, considering they are located off-center. For example, a source from the center channel should be perceived at [-0.5, 1.5] when using the loudspeaker array, where the position of the participant is [0, 0].

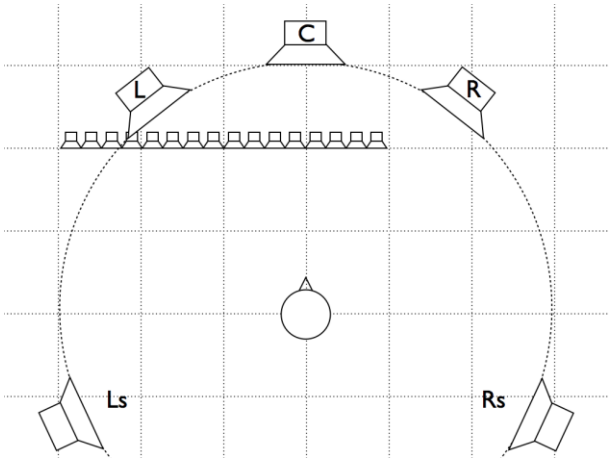

Figure 9. The experiment configuration

3.3 Attributes

The proposed algorithm consists of a separation process, which includes novel methods, and a reproduction process, which uses conventional methods. In principle, the separation must therefore be evaluated both individually and based on the whole procedure. Table 3 shows the attributes that should be evaluated to verify the proposed algorithm.

#	Attribute	Measure
1	Artifacts due to extraction and separation	Calculation of SAR or listening test
2	Interferences among sources	Calculation of SIR or listening test
3	Estimated positions of each sources	Position (or angle) difference or listening test

Table 3. Attributes to be evaluated and their corresponding measures

However, to evaluate the attributes in Table 3, the original sources and mixing matrices of real 5.1 excerpts are required. Owing to a lack of such database, we choose an alternative evaluation method that can be conducted using 5.1 surround audio signals only. We performed subjective tests with the sound reproduced through the loudspeaker array after the whole procedure was conducted, as shown in Table 4.

#	Attributes
1	Overall Sound Quality
2	Frontal Localization Quality

Table 4. Selected attributes evaluated during this research.

To evaluate the overall sound quality, artifacts and noises introduced by the extraction and separation were first evaluated. The overall localization quality includes the estimated positions of the sources. In addition, as the interferences among the sources result in vague or incorrect localizations, they were evaluated in terms of the overall localization quality.

In the overall sound quality test, the participants were asked to focus on the timbral quality and the sound artifacts. In the front localization quality test, the participants were then asked to assess the localization accuracy and precision of the sources. We paid particular attention to the front side, as the sources from the surround channels are not reproduced effectively through only the front loudspeaker array.

Along with these tests, the participants were asked to compare two algorithms (the proposed and comparison algorithms) with a hidden reference, and give them a score of 0 to 100.

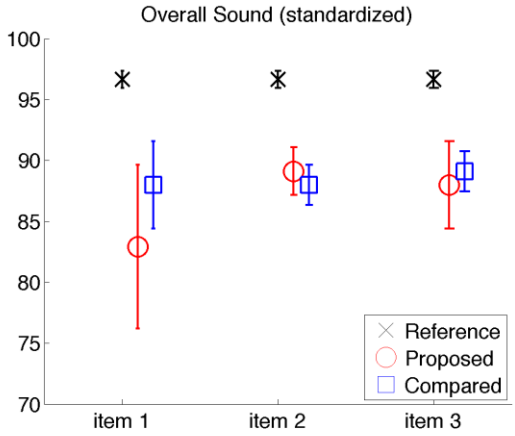

Figure 10. The results of timbral quality

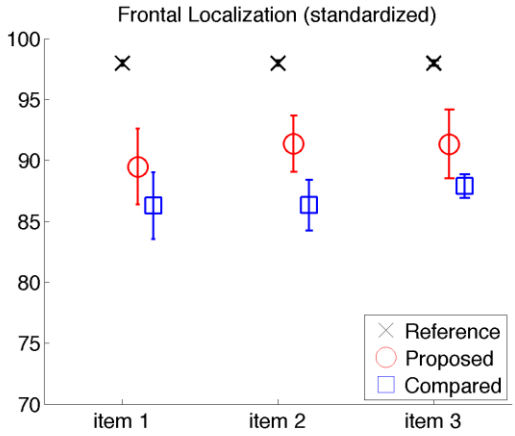

Figure 11. The results of localization quality.

3.4 Results and Discussions

Figure 10 and Figure 11 show the results of subjective tests using the mean and 95% confidence interval (ci). The scores were standardized to compensate the differences between participants. First, the scores were normalized to $z = (x - m_u)/\sigma_u$, where x, m_u, and σ_u are the score, mean participant score, and standard deviation of the participant scores, respectively. The normalized score, x_n is then computed as $z \times \sigma_g + m_g$, where m_g and σ_g are the global mean and standard deviation of the scores, respectively.

In short, the proposed algorithm showed both an advanced performance and a degraded sound quality. For item 1, which is a music excerpt, the overall sound quality showed a large gap with a large ci, while the ci values of the localization quality are overlapped. As the sources are separated and then reconstructed, the timbral quality is degraded, resulting in unintended low-frequency enhancements. An active downmix, or an advanced source separation with more accurate source-number estimation, may solve this problem

This timbral distortion is more critical for music, as the overall quality showed similar figures for items 2 and 3. As these excerpts consisted of sound effects from moving sources, fewer sound sources exist concurrently. This causes less timbral distortion and a more sensible difference in the localization qualities for these items.

4 CONCLUSIONS

We introduced a novel algorithm for upmixing multichannel audio signals into signals for loudspeaker arrays using source separation and WFS. For the separation of 5.1 audio-surround signals, we model the signals as a sum of stereo signals, which are called element signals. A Stereo Component Extraction technique is then introduced to extract the element signals. Laplacian-Mixture Models based on IID-panning are then assumed to separate the sources from the element signals. Finally, the sources are rendered through WFS and focused sources. During the reported experiment, we found improvements in the localization quality but a degradation of the sound quality.

Many aspects of this study can be further researched. As the timbral quality is a very important attribute, it should be improved for the practical use of the proposed approach. In addition, the complexity should also be considered for this purpose. During the synthesis phase, an active downmix can be a solution for a better timbral quality, as mentioned earlier. A fine reproduction technique for the sources outside a loudspeaker array is necessary to retain the convenience of such an array over a discrete loudspeaker system.

5 ACKNOWLEDGEMENTS

This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.

REFERENCES

[1] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, and E. Schuijers, "Spatial audio object coding (SAOC)-The upcoming MPEG standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*, 2008.

[2] M. Cobos, J. J. López, A. Gonzalez, and J. Escolano, "Stereo to wave-field synthesis music up-mixing: An objective and subjective evaluation," in *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*, 2008, pp. 1279-1284.

[3] M. Cobos and J. Lopez, "Resynthesis of wavefield synthesis scenes from stereo mixtures using sound source separation algorithms," *Journal of the Audio Engineering Society*, 2009.

[4] N. Kamado, M. Hirata, H. Saruwatari, and K. Shikano, "Object-based stereo up-mixer for wave field synthesis based on spatial information clustering," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 594-598.

[5] "ITU-R. BS. 775-2,Multichannel stereophonic sound system with and without accompanying picture," *International Telecommunications Union, Geneva, Switzerland*, 2006.

[6] T. Holman, *5.1 surround sound: Up and running*: Focal, 2000.

[7] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, p. 2764, 1993.

[8] H. Wittek, S. Kerber, F. Rumsey, and G. Theile, "Spatial perception in wave field synthesis rendered sound fields: Distance of real and virtual nearby sources," in *Audio Engineering Society Convention 116*, 2004.

[9] S. Spors, H. Wierstorf, M. Geier, and J. Ahrens, "Physical and perceptual properties of focused sources in wave field synthesis," in *127th AES Convention*, 2009.

[10] H. Wierstorf, A. Raake, M. Geier, and S. Spors, "Perception of Focused Sources in Wave Field Synthesis," *Journal of the Audio Engineering Society*, vol. 61, pp. 5-16, 2013.

[11] H. Chung, H. Shim, N. Hahn, S. B. Chon, and K.-M. Sung, "Sound reproduction method by front loudspeaker array for home theater applications," *Consumer Electronics, IEEE Transactions on*, vol. 58, pp. 528-534, 2012.

[12] X.-R. Cao and R.-w. Liu, "General approach to blind source separation," *Signal Processing, IEEE Transactions on*, vol. 44, pp. 562-571, 1996.

[13] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on*, vol. 52, pp. 1830-1847, 2004.

[14] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, pp. I-881-I-884.

[15] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discriminiation and resynthesis," 2004.

[16] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1564-1578, 2007.

[17] M. Kim, S. Beack, K. Choi, and K. Kang, "Gaussian mixture model for singing voice separation from stereophonic music," in *Audio Engineering Society Conference: 43rd International Conference: Audio for Wirelessly Networked Personal Devices*, 2011.

[18] N. Mitianoudis and T. Stathaki, "Overcomplete source separation using Laplacian mixture models," *Signal Processing Letters, IEEE*, vol. 12, pp. 277-280, 2005.