Geometrically-Motivated Primary-Ambient Decomposition With Center-Channel Extraction

Jouni Paulus* ⓘ

Fraunhofer Institute for Integrated Circuits IIS, and
International Audio Laboratories Erlangen[†]
Erlangen, Germany

Matteo Torcoli ⓘ

Fraunhofer Institute for Integrated Circuits IIS, and
International Audio Laboratories Erlangen[†]
Erlangen, Germany

Abstract—A geometrically-motivated method for primary-ambient decomposition is proposed and evaluated in an up-mixing application. The method consists of two steps, accommodating a particularly intuitive explanation. The first step consists of signal-adaptive rotations applied on the input stereo scene, which translate the primary sound sources into the center of the rotated scene. The second step applies a center-channel extraction method, based on a simple signal model and optimal in the mean-squared-error sense. The performance is evaluated by using the estimated ambient component to enable surround sound starting from real-world stereo signals. The participants in the reported listening test are asked to adjust the audio scene envelopment and find the audio settings that pleases them the most. The possibility for up-mixing enabled by the proposed method is used extensively, and the user satisfaction is significantly increased compared to the original stereo mix.

Index Terms—center-channel extraction, listening test, primary-ambient decomposition, up-mixing

I. Introduction

Center-channel extraction (CE) and primary-ambient decomposition (PAD) are audio signal processing algorithms commonly used for decomposing a 2-channel input signal into a 3- or 4-channel output. The output can consist of estimates of the left, right, and center channel, or primary (or direct) components with a direction and ambient components without a clear direction. The methods are often motivated by the applications including dereverberation or controlling dry/wet-ratio [1]–[4], signal separation [5]–[7], up-mixing from stereo to a higher number of output channels [8]–[17], direct source re-panning without modifying the ambience [18], and audio coding by separating spatial cues from the audio itself [19].

With the exception of some more recent deep-learning based approaches, e.g., [15], [20], the CE and PAD solutions rely on the cross-correlation properties of the observed stereo signal channels and the signal model assumptions. One of the simplest forms is from [13], where the cross-channel coherence value of the input is mapped into an ambience separation mask weight. Other methods attempt to estimate the covariance properties of the target component signal with explicit equations, e.g., [1]–[3], [5], [9], [12], [21], [22], or by utilizing principal component analysis (PCA). The PCA

methods, e.g., [14] assume that the eigenvalues of the observation covariance matrix correspond to the levels of the primary and ambient components [19]. Assuming the larger eigenvalue to correspond to the primary component may lead to under-estimating the ambient component level and some works use scaling based on the ratio of the eigenvalues [4], [10], [23], [24]. The separation may rely on per-channel filters, e.g., [25], common filters for all channels, e.g., [2], or cross-channel filters, e.g., [8]. Targeting at signals with dominant ambient components, [26] formulates the PAD problem into a task of finding ambient component phase maximizing the sparsity of the primary component. This model was recently extended by introducing an assumption of the primary component having a harmonic spectrum [27]. Some works extend the definition from two input channels into multiple channels and defining the direct sound to be the part common in all channels [28]. Some methods have multiple methods running in parallel and use late fusion to outperform the single approaches [29]. The work [30] makes a good effort of unifying several approaches a common vocabulary and analysing the methods more closely.

The contribution of this paper is threefold. 1) We present a step-by-step derivation of a multi-channel Wiener-filter CE method based on estimating the signal statistics from a simple signal model. 2) This method is extended for a novel PAD solution by adding geometrically-motivated pre- and post-processing rotation steps. 3) The performance of the proposed PAD method is evaluated with the Adjustment/Satisfaction Test (A/ST) in a stereo-to-quad up-mixing application.

II. Background

The CE methods attempt to decompose a 2-channel (stereo) signal $\boldsymbol{x}_L$ and $\boldsymbol{x}_R$ into three components: left $\boldsymbol{l}$, center $\boldsymbol{c}$, and right $\boldsymbol{r}$ based on the signal model (the derivation assumes these being possibly complex-valued sub-band signals, e.g., in short-time Fourier transform (STFT) domain), and the time-dependency and band index are dropped in the following):

$$\begin{cases} \boldsymbol{x}_L = \boldsymbol{l} + \boldsymbol{c} \\ \boldsymbol{x}_R = \boldsymbol{r} + \boldsymbol{c} \end{cases}, \tag{1}$$

which in a matrix form is:

$$\begin{bmatrix} \boldsymbol{x}_L^\top \\ \boldsymbol{x}_R^\top \end{bmatrix} = \boldsymbol{D} \begin{bmatrix} \boldsymbol{l}^\top \\ \boldsymbol{r}^\top \\ \boldsymbol{c}^\top \end{bmatrix}, \tag{2}$$

where $x^\top$ is the transpose of x and D is the mixing matrix:

$$D = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}. \tag{3}$$

An assumption making the decomposition more approachable is that all the three component signals are statistically independent, i.e.,

$$E\{l^H r\} = E\{l^H c\} = E\{c^H r\} = 0, \tag{4}$$

where $E\{\cdot\}$ is the expected value operator over time and where x^H is the Hermitian transpose

The PAD methods use the signal model:

$$\begin{cases} x_L = a_L + p_L \\ x_R = a_R + p_R \end{cases}, \tag{5}$$

where a_L and a_R are the ambient components, and p_L and $p_R = g p_L$ the primary signal components, all to be estimated by the method, and g is a level-panning weight. Similar to (4), the component signals are assumed to be independent, i.e.,

$$E\{a_L^H a_R\} = E\{a_L^H p_L\} = E\{a_R^H p_R\} = 0. \tag{6}$$

Some earlier works, e.g., [3], [11] on PAD make the assumption that the ambient signal components have the same level, i.e., $E\{a_L^H a_L\} = E\{a_R^H a_R\}$. The PAD algorithm proposed in this paper relaxes this assumption.

III. PROPOSED METHOD

This section describes the estimation of the signal statistics from the observations, a minimum-mean-squared error (MMSE) method using the signal statistics for performing CE, and finally a novel geometry-based extension of the CE-solution for PAD.

A. Estimating Signal Statistics

The proposed CE-method needs to know the covariance matrix C_{obj} of the component signals. Remembering the assumption that the three component signals are independent (4), the covariance matrix consists of only the component energies $E\{l^H l\}$, $E\{r^H r\}$, and $E\{c^H c\}$ as

$$C_{obj} = \begin{bmatrix} E\{l^H l\} & 0 & 0 \\ 0 & E\{r^H r\} & 0 \\ 0 & 0 & E\{c^H c\} \end{bmatrix}. \tag{7}$$

The (auto- and cross-) covariance matrix of the two observed mixture signals is

$$C_x = \begin{bmatrix} c_{L,L} & c_{L,R} \\ c_{R,L} & c_{R,R} \end{bmatrix} = \begin{bmatrix} E\{x_L^H x_L\} & E\{x_L^H x_R\} \\ E\{x_R^H x_L\} & E\{x_R^H x_R\} \end{bmatrix}. \tag{8}$$

Here, $c_{i,j} = E\{x_i^H x_j\}$ are the covariance matrix entries between the two observed signals x_i and x_j, when we assume the two signals to have zero mean. With the assumption on the statistical independence of the component signals, the covariance matrix mixing model is

$$\begin{bmatrix} c_{L,L} \\ c_{L,R} \\ c_{R,L} \\ c_{R,R} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} E\{l^H l\} \\ E\{c^H c\} \\ E\{r^H r\} \end{bmatrix}. \tag{9}$$

Solving this provides us the estimates of the three component signal energies as

$$\begin{cases} E\{l^H l\} = c_{L,L} - \mathcal{R}\{c_{L,R}\} \\ E\{c^H c\} = \mathcal{R}\{c_{L,R}\} \\ E\{r^H r\} = c_{R,R} - \mathcal{R}\{c_{L,R}\} \end{cases}, \tag{10}$$

where $\mathcal{R}\{x\}$ returns the real part of the complex number. For the presentation clarity, the operator is omitted from the equations in the remainder of this paper.

B. Component Separation

The component signals can be obtained with a multi-channel Wiener filter G applied on the observed stereo signal as

$$\begin{bmatrix} \hat{l}^\top \\ \hat{r}^\top \\ \hat{c}^\top \end{bmatrix} = G \begin{bmatrix} x_L^\top \\ x_R^\top \end{bmatrix}. \tag{11}$$

The filter coefficients can be obtained using a well-known MMSE solution

$$G = C_{s,x} C_x^{-1}, \tag{12}$$

where $C_{s,x}$ is a covariance matrix between the components and mixture signals. From (2) this can be written as the product

$$C_{s,x} = C_{obj} D^H. \tag{13}$$

Taking the component signal covariance matrix elements from (7), this can be written as

$$C_{s,x} = \begin{bmatrix} c_{L,L} - c_{L,R} & 0 \\ 0 & c_{R,R} - c_{L,R} \\ c_{L,R} & c_{L,R} \end{bmatrix}. \tag{14}$$

This leads into the full un-mixing matrix solution of

$$G = \frac{\begin{bmatrix} c_{L,L} c_{R,R} - c_{R,R} c_{L,R} & c_{L,R}^2 - c_{L,R} c_{L,L} \\ c_{L,R}^2 - c_{L,R} c_{R,R} & c_{L,L} c_{R,R} - c_{L,L} c_{L,R} \\ c_{R,R} c_{L,R} - c_{L,R}^2 & c_{L,L} c_{L,R} - c_{L,R}^2 \end{bmatrix}}{c_{L,L} c_{R,R} - c_{L,R}^2}. \tag{15}$$

C. Extension into Primary-Ambient Decomposition

For applying the CE algorithm derived above for solving the PAD model of (5) we assume that our observation is a spatially rotated CE-scene in which the dominant sound direction corresponds to the primary / direct component. In other words, we find a rotation angle θ such that when applied on our 2-channel observation x_L, x_R, the resulting rotated scene fulfils the assumptions that a) the primary signal component is centered in the scene and b) uncorrelated with the two independent ambient signal components c) that have equal energies. We can apply CE in this rotated domain, and then apply inverse rotation on the extracted components.

The rotation of the scene is done with a rotation matrix

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{16}$$

The rotation angle θ can be solved from the covariance matrix of the rotated observations, being equal to applying the rotation on the covariance matrix of the observations:

$$C_{rot} = R_\theta C_x R_\theta^H = \begin{bmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{bmatrix}. \tag{17}$$

Requiring that the rotated observed signals have the same energy corresponds to requiring the main diagonal elements to be equal: $c_{1,1} = c_{2,2}$. Combining (8) and (17) and solving the rotation from this gives us the angle

$$\theta = \frac{1}{2} \tan^{-1} \frac{c_{L,L} - c_{R,R,}}{2c_{L,R}}. \tag{18}$$

Now we can use the earlier center-channel extraction algorithm for the decomposition. The full proposed PAD algorithm consists of the steps:

1) Compute input covariance matrix C_x with (8).
2) Determine the centering rotation θ with (18), and compute the rotated covariance matrix C_{rot} with (17).
3) Compute the un-mixing matrix G with (12) using the rotated covariance matrix C_{rot} in the place of the observation covariance matrix C_x.
4) Combine the centering rotation R_θ and the counter-rotation R_θ^H of the separated components (single-channel primary component duplicated in both outputs) with the un-mixing matrix G using

$$G_{full} = \begin{bmatrix} R_\theta^H & \mathbf{0} \\ \mathbf{0} & R_\theta^H \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} G R_\theta. \tag{19}$$

(The mixing matrix in the equation assigns the center of the rotated scene into the primary signal.)

5) Apply the combined un-mixing matrix on the observed signals with

$$\begin{bmatrix} a_L^\top \\ a_R^\top \\ p_L^\top \\ p_R^\top \end{bmatrix} = G_{full} \begin{bmatrix} x_L^\top \\ x_R^\top \end{bmatrix}. \tag{20}$$

The counter-rotation achieves that the primary component is located again in the original direction. A possible drawback of the counter-rotation applied on the ambient components is that even if the extracted components would be independent and of equal energy in the rotated coordinate system, the counter-rotation modifies both these properties.

It is worth noting that the computation of the un-mixing matrix (19) does not need to apply the rotations to the observed signals, but all operations can be performed in the covariance matrix domain. Further computational complexity optimizations can be obtained by inspecting the structure of the full un-mixing matrix more closely. We divide it into two parts corresponding to the ambient and primary components:

$$G_{full} = \begin{bmatrix} G_A \\ G_P \end{bmatrix}. \tag{21}$$

The part corresponding to the ambient signal component is

$$G_A = R_\theta^H \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} G R_\theta \tag{22}$$

$$= \begin{bmatrix} c_{R,R} & -c_{L,R} \\ -c_{L,R} & c_{L,L} \end{bmatrix} \frac{k - c_{L,L} - c_{R,R}}{2(c_{L,R}^2 - c_{L,L}c_{R,R})}, \tag{23}$$

with

$$k = \sqrt{(c_{L,L} - c_{R,R})^2 + 4c_{L,R}^2}. \tag{24}$$

In earlier publications, e.g., [2], [9], [25] this quantity corresponds to the total energy of the direct component in their signal model.

The part corresponding to the primary signal component is

$$G_P = R_\theta^H \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} G R_\theta = \begin{bmatrix} g_{P1,1} & g_{P1,2} \\ g_{P2,1} & g_{P2,2} \end{bmatrix} \tag{25}$$

with

$$\begin{cases} g_{P1,1} = \frac{2c_{L,R}^2 + c_{R,R}(c_{R,R} - c_{L,L} - k)}{2(c_{L,R}^2 - c_{L,L}c_{R,R})} \\ g_{P1,2} = \frac{-c_{L,R}(k - c_{L,L} - c_{R,R})}{2(c_{L,R}^2 - c_{L,L}c_{R,R})} \\ g_{P2,1} = g_{P1,2} \\ g_{P2,2} = \frac{2c_{L,R}^2 + c_{L,L}(c_{L,L} - c_{R,R} - k)}{2(c_{L,R}^2 - c_{L,L}c_{R,R})} \end{cases}. \tag{26}$$

It is possible to manipulate this representation further and to find the relationship

$$G_P = I - G_A, \tag{27}$$

where I is a 2-by-2 identity matrix.

Considering the covariance matrix of the rotated observed signals (17), and inserting the computed rotation angle from (18), the cross-terms can be simplified into

$$c_{1,2} = c_{2,1} = \frac{1}{2}\sqrt{(c_{L,L} - c_{R,R})^2 + 4c_{L,R}^2} = \frac{1}{2}k. \tag{28}$$

IV. EVALUATION

The proposed method is evaluated in a listening test focusing on the up-mixing application, similar to [8], [15]. The stereo input is split into primary and ambient components and some of the ambient component can be moved into the rear channels of a 5.1 setup (the center and LFE channels stay silent). The amount moved is controlled by the user with a dial over 15 steps from full signal in the front channels to only primary component in the front and ambient component in the rear channels. The range is extended by 10 steps in which the ambient component is further amplified up to $20\,\text{dB}$, and in the low end by 5 steps of gradual stereo image narrowing to dual mono, see Table I. All conditions are normalized to equal integrated loudness [31]. The adjustment is done with a continuously rotating dial without a tactile feedback of the steps. The test is conducted using the Adjustment Satisfaction Test (A/ST) [7], [32] with the main task of *"Please find the audio setting that pleases you the most with respect to audio scene envelopment and overall audio quality."* After adjusting the audio setting, the listener rates their satisfaction of the adjusted version compared to the original stereo mix on the scale *much worse, worse, slightly worse, same, slightly better,*

Table I

THE LOUDSPEAKER SIGNALS IN THE THREE OPERATING REGIONS IN THE UP-MIXING LISTENING TEST. IN THE *narrowing* RANGE THE STEREO IMAGE IS MADE NARROWER WITH THE CROSS-MIXING COEFFICIENT $a \in \{0.5, 0.57, 0.66, 0.76, 0.87\}$. IN THE *amb. relocation* REGION THE AMBIENT COMPONENT IN THE FRONT CHANNELS IS ATTENUATED BY A GAIN $g_a \in \{0, -1.5, -3, -5, -7.5, -10.5, -14, -18, -23, -28, -34, -41, -49, -59, -76, -96\}$ dB, AND THE MATCHING CONTRIBUTION IS MOVED TO THE REAR CHANNELS. IN THE *amb. boost* REGION, THE FULL AMBIENT SIGNAL IS LOCATED IN THE REAR CHANNELS AND FURTHER AMPLIFIED BY $b_a \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 20\}$ dB.

output channel	narrowing	amb. relocation	amb. boost
$\mathbf{y}_{L,front}$	$a\mathbf{x}_L + (1-a)\mathbf{x}_R$	$\mathbf{p}_L + g_a\mathbf{a}_L$	$\mathbf{p}_L$
$\mathbf{y}_{R,front}$	$(1-a)\mathbf{x}_L + a\mathbf{x}_R$	$\mathbf{p}_R + g_a\mathbf{a}_R$	$\mathbf{p}_R$
$\mathbf{y}_{L,rear}$	0	$(1-g_a)\mathbf{a}_L$	$b_a\mathbf{a}_L$
$\mathbf{y}_{R,rear}$	0	$(1-g_a)\mathbf{a}_R$	$b_a\mathbf{a}_R$

better, *much better*, each with 5 steps between two labelled positions on the scale.

The test stimuli are 16 items of 20–30 s in length, half are music-only (instrumental only, e.g., orchestral, and with singing, e.g., pop) and half are real-world broadcast content (TV dramas and magazine shows, where speech overlaps with background music, effects, or noise). The sampling rate is 48 kHz. The processing uses STFT with 1024-sample frames with 50% overlap, sine window, and 2x zero-padding. The observation cross-covariance matrix is computed for each frequency bin per-frame, and these are averaged with a 5-frame sliding mean for obtaining (8). The computed un-mixing matrices $\mathbf{G}_P$ and $\mathbf{G}_A$ are smoothed with a 3-frame sliding mean for reducing musical noise. The test takes place in a listening room with a 5.1 loudspeaker setup.

We compared informally several PAD algorithms from the literature, e.g., [1], [3], [13], [17], [24], [30], [33] in this same processing framework. Some systems performed sub-optimally, while a group of methods including the proposed one performed in general equally well with only small per-item differences. We expect this group of methods to perform comparably especially in application-oriented evaluation. Since we are more interested in showing that PAD-based up-mixing is ripe enough for applications rather than finding differences possibly stemming from unoptimized implementations, we include only the proposed method in the listening test.

V. RESULTS

The listening test involved 14 voluntary participants recruited among the research staff and interns at Fraunhofer IIS. The results of three of them were excluded in a post-screening phase, as they used the *worse* part of the satisfaction scale at least once, indicating a misunderstanding of the given task or problems during the adjustment phase (the initial stereo mix could be selected during adjustment, so the final satisfaction should be at least *same* as the initial condition). Figure 1 shows the main results after post-screening. The adjustment is measured using Rear-to-Front Ratio (RFR), that is the energy ratio of the rear loudspeakers to the front loudspeakers, measured in dB:

$$\text{RFR} = 10\log_{10}\frac{\sum_t \left(|\mathbf{y}_{L,rear}(t)|^2 + |\mathbf{y}_{R,rear}(t)|^2\right)}{\sum_t \left(|\mathbf{y}_{L,front}(t)|^2 + |\mathbf{y}_{R,front}(t)|^2\right)}. \quad (29)$$
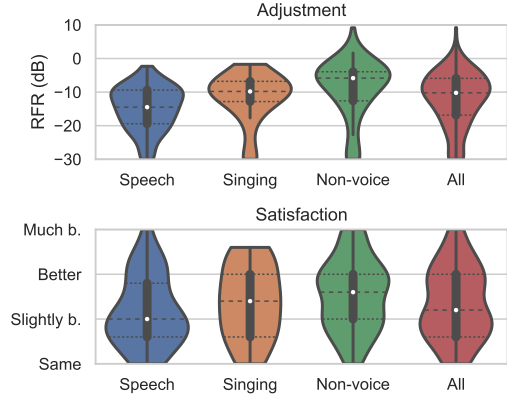

Figure 1. Violin plots of the Rear-to-Front Ratios (RFR) selected during adjustment (initial RFR= $-\infty$ dB truncated to -30 dB) and resulting satisfaction levels for test items including speech, singing, not including any voice, and overall. The quartiles are shown by dashed lines. Medians for the three item groups are statistically significantly different both in terms of selected RFR and of resulting satisfaction. Overall, a median RFR of -10.2 dB is selected by the participants and the user satisfaction is clearly increased in every case (Wilcoxon signed rank test for median = *Same*: $p < 0.00$).

When analysing the listening test results, we observed three signal classes: *speech* (7 items), *singing* (3 items), and *non-voice* (5 items) instead of the originally-assumed *music* and *broadcast*, and the result analysis uses this division.

It can be observed that the listeners made use of the possibility of re-locating some of the ambient component into the rear channels. Overall, a median RFR of -10.2 dB is selected and the resulting satisfaction is clearly increased. The resulting up-mixed signals are rated by the participants on average being *slightly better* to *better* compared to the original stereo mix. This suggests that the audio quality of the proposed approach is good enough for providing some level of up-mixing freedom that improves the satisfaction of the listeners compared to the original stereo audio.

Comparing items with speech, singing, or no voice at all, a trend appears. The highest RFRs (median -5.8 dB) are selected for non-voice items (instrumental music or applause), resulting in the highest satisfaction levels. Positive RFRs are also observed. Speech items show the lowest RFRs (median -14.5 dB). Music items with singing voice lie in between (median RFR -9.8 dB). Statistically significant difference between medians is tested by a Kruskal-Wallis test for both adjustment ($p < 0.00$) and satisfaction ($p = 0.026$), where non-parametric testing was selected because the Shapiro-Wilk test cast doubts on the normality of the considered groups.

The lower adjustment and satisfaction on speech items can be due to different factors. One explanation is that participants would desire higher RFRs also for these items, but had to compromise with decreasing quality due to inaccuracies introduced by the proposed algorithm (particularly well perceived on speech). An alternative is that higher RFRs are not aesthetically preferred on this kind of content. Reverberation associated with speech (or singing voice) can be a problem when it is moved as a part of the ambient component to the rear loudspeakers changing the perceived location of the talker (or

the singer). In fact, some participants reported that they did not want the talker position to change, while they enjoyed being surrounded by the scene background noise and music. Using a decorrelator on the signal moved in the rear channels could help on this issue, but this was omitted from the experiments for simplicity and reproducibility.

VI. CONCLUSION

This paper presented a step-by-step derivation of a center-channel extraction method using MMSE-filters for the separation of the signal components. The method was extended into primary-ambient-decomposition with rotations for creating a centred scene with the primary component being the center channel, and the overall algorithm was simplified into two equations. The proposed PAD method delivers results that were informally rated to be perceptually similar to solutions from the literature. The method was evaluated using A/ST methodology in the application of stereo-to-quad up-mixing. The test participants made use of the adjustment, especially for items without speech or singing, and their overall satisfaction to the spatial impression of the resulting up-mix signal was increased.

REFERENCES

[1] J. B. Allen, D. A. Berkely, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, Oct. 1977.

[2] A. Walther and C. Faller, "Direct-ambient decomposition and upmix of surround signals," in *Proc. of 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, Oct. 2011, pp. 277–280.

[3] C. Uhle and E. A. P. Habets, "Direct-ambient decomposition using parametric Wiener filtering with spatial cue control," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 36–39.

[4] M. M. Goodwin, "Primary-ambient decomposition and dereverberation of two-channel and multichannel audio," in *Proc. of 42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, USA, Oct. 2008, pp. 797–800.

[5] J. Geiger, P. Grosche, and Y. L. Parodi, "Dialogue enhancement of stereo sound," in *Proc. of 23rd European Signal Processing Conference*, Nice, France, Aug. 2015, pp. 874–878.

[6] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc. of 7th International Conference on Digital Audio Effects*, Naples, Italy, Oct. 2004, pp. 240–244.

[7] J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch, and H. Fuchs, "Source separation for enabling dialogue enhancement in object-based broadcast with MPEG-H," *Journal of the Audio Engineering Society*, vol. 67, no. 7/8, pp. 510–521, Jul./Aug. 2019.

[8] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2141–2150, 2007.

[9] C. Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 54, no. 11, pp. 1051–1064, Nov. 2006.

[10] Y.-H. Baek, S.-W. Jeon, Y.-C. Park, and S.-P. Lee, "Efficient primary-ambient decomposition algorithm for audio upmix," in *Proc. of 133rd Audio Engineering Society Convention*, San Francisco, California, USA, Oct. 2012.

[11] A. Härmä, "Estimation of the energy ratio between primary and ambience components in stereo audio data," in *Proc. of 19th European Signal Processing Conference*, Barcelona, Spain, Aug. 2011, pp. 1643–1647.

[12] A. Härmä and M. Park, "Extraction of voice from the center of the stereo image," in *Proc. of 130th Audio Engineering Society Convention*, New York, New York, USA, Oct. 2009.

[13] C. Avendano and J.-M. Jot, "Ambience extraction and synthesis from stereo signal for multi-channel audio up-mix," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, May 2002, pp. 1957–1960.

[14] R. Irwan and M. Aarts, Ronald, "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–926, Nov. 2002.

[15] K. M. Ibrahim and M. Allam, "Primary-ambient source separation for upmixing to surround sound systems," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Alberta, Canada, Apr. 2018, pp. 431–435.

[16] S. Kraft and U. Zölzer, "Low-complexity stereo signal decomposition and source separation for application in stereo to 3D upmixing," in *Proc. of 140th Audio Engineering Society Convention*, Paris, France, Jun. 2016.

[17] Y. Hashimoto, Y. Goto, and A. Ando, "Subjective evaluation of multichannel upmix method based on interchannel coherent component extraction," in *Proc. of 144th Audio Engineering Society Convention*, Milan, Italy, May 2018.

[18] A. Adami, M. Schoeffler, and J. Herre, "Re-panning of directional signals and its impact on localization," in *Proc. of 23rd European Signal Processing Conference*, Nice, France, Aug. 2015, pp. 549–553.

[19] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, USA, Apr. 2007, pp. 9–12.

[20] W. Lim, S. Beack, and T. Lee, "Multi-level stereo attention model for center channel extraction," in *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2019.

[21] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Jun. 2000, pp. 2985–2988.

[22] E. Vickers, "Frequency-domain two- to three-channel upmix for center channel derivation and speech enhancement," in *Proc. of 127th Audio Engineering Society Convention*, London, UK, May 2011.

[23] M. M. Goodwin, "Geometric signal decompositions for spatial audio enhancement," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, Apr. 2008, pp. 409–412.

[24] K. M. Ibrahim and M. Allam, "Primary-ambient extraction in audio signals using adaptive weighting and principal component analysis," in *Proc. of Sound and Music Computing Conference*, Hamburg, Germany, Sep. 2016, pp. 227–232.

[25] J. Merimaa, M. M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. of 123rd Audio Engineering Society Convention*, New York, New York, USA, Oct. 2007.

[26] J. He, W.-S. Gan, and E.-L. Tan, "Primary-ambient extraction using ambient phase estimation with a sparsity constraint," *IEEE Signal Processing Letters*, vol. 22, no. 8, Aug. 2015.

[27] H. Zhu, C. Shi, and Y. Wang, "F0-estimation-based primary ambient extraction for stereo signals," in *Proc. of 29th European Signal Processing Conference*, Dublin, Ireland, Aug. 2021, pp. 66–70.

[28] C. Uhle and E. A. P. Habets, "Subband center signal scaling using power ratios," in *Proc. of 53rd AES International Conference: Semantic Audio*, London, UK, Jan. 2014.

[29] J. He and W.-S. Gan, "Informed hybrid primary ambient extraction for spatial audio reproduction," in *Proc. of 2016 AES International Conference on Headphone Technology*, Aalborg, Denmark, Aug. 2016.

[30] J. He, E.-L. Tan, and W.-S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 505–517, Feb. 2014.

[31] International Telecommunication Union, *ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level*, Geneva, Switzerland, Oct. 2015.

[32] M. Torcoli, J. Herre, H. Fuchs, J. Paulus, and C. Uhle, "The Adjustment / Satisfaction Test (A/ST) for the evaluation of personalization in broadcast services and its application to dialogue enhancement," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 524–538, Jun. 2018.

[33] S. Kraft and U. Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *Proc. of 18th International Conference on Digital Audio Effects*, Trondheim, Norway, Nov. 2015.