DIRECT-AMBIENT DECOMPOSITION USING PARAMETRIC WIENER FILTERING WITH SPATIAL CUE CONTROL

Christian Uhle and Emanuël A. P. Habets

International Audio Laboratories Erlangen*, Erlangen, Germany

ABSTRACT

A method for decomposing audio signals into direct signals and ambient signals is described that can be applied to sound post-production and reproduction. The proposed method is based on a parametric multichannel Wiener filter (MWF) that enables a trade-off between the attenuation of the interfering signal and the distortion of the desired signal. We show that the MWF leads to distortions of the spatial cues of the ambient output signal, namely inter-channel correlation and inter-channel level difference. Our proposed solution is to control the trade-off parameter of the parametric MWF to ensure that these spatial distortions are inaudible.

Index Terms— direct-ambient decomposition, parametric Wiener filter

1. INTRODUCTION

Audio signals can be modeled as mixtures of direct sounds and ambient sounds. Direct sounds commonly lead to coherent signals, whereas ambient sounds (like room reverberation, applause or babble noise) lead to signals that are at least partially incoherent. The separation of direct and ambient signals can be applied to manipulate the amount of reverberation in an audio recording and for upmixing, i.e., for creating an output signal given an input signal with fewer channels. The ambient signals can be used for producing, for example, surround signals that are fed into the rear loudspeakers of a surround sound setup. The process of separating direct and ambient signals can be accomplished using a direct-ambient decomposition (DAD) method.

Various DAD methods have been proposed using spectral weighting that are based on i) the inter-channel correlation (ICC) [1, 2], and ii) estimates of the ambience power [3]. A method for extracting the uncorrelated signals using an adaptive filter algorithm for predicting the direct signal component and obtaining the ambience as the residual signal is described in [4]. For the processing of multichannel signals, a method based on estimates of the power spectral density (PSD) of the ambient signal using a two-channel downmix and Wiener filtering of the input signal [5] and a method using pairwise correlations [6] have been proposed. Approaches for the processing of single-channel signals are based on non-negative matrix factorization [7], spectral weighting using feature extraction and supervised learning [8], or on the estimation of the magnitude transfer function of the reverberant system which has generated the ambient signal [9].

Other approaches compute output signals as linear combinations of the input channels, e.g. the method proposed by Faller [10] that is based on the MWF, and methods using principal component analysis (PCA) [3, 11, 12]. An extensive comparison of the MWF and PCA approaches, a generic linear estimation framework, and means for adjusting the performance with respect to various distortion measures have been recently presented in [13]. In this work, we extend this idea and propose to achieve such adjustment signal dependently. To this end, we present a method based on the parametric MWF (PMWF), show potential distortions of the spatial cues of the output ambient signal introduced by the MWF, and propose a means for limiting these distortions to be below psychoacoustic thresholds.

The paper is organized as follows: in Sec. 2 the problem of DAD is formulated, Sec. 3 describes the PMWF for DAD, Sec. 4 introduces the control of the PMWF using psychoacoustic parameters, in Sec. 5 the proposed method is evaluated, and in Sec. 6 the conclusions are given.

2. PROBLEM FORMULATION

The signal model is represented in the time-frequency domain where m denotes the time index and k denotes the subband index. The i-th input signal is denoted by $Y_i(m,k)$ and consists of an additive mixture of a direct signal $D_i(m,k)$ and an ambient signal $A_i(m,k)$, which are both are assumed to be Gaussian random variables with zero mean. An input signal with 2 channels can then be written as[1]

$$\mathbf{y} = \mathbf{d} + \mathbf{a}, \tag{1}$$

with $\mathbf{y} = [Y_1\, Y_2]^T$, $\mathbf{d} = [D_1\, D_2]^T$, and $\mathbf{a} = [A_1\, A_2]^T$.

The objective of the DAD is to estimate $\mathbf{d}$ and $\mathbf{a}$, which in the following are obtained using

$$\hat{\mathbf{d}} = \mathbf{H}_D^H \mathbf{y}, \tag{2}$$

$$\hat{\mathbf{a}} = \mathbf{H}_A^H \mathbf{y}, \tag{3}$$

where $(\cdot)^H$ denotes the conjugate transpose operation, $\mathbf{H}_D$ is the filter matrix to estimate $\mathbf{d}$, and $\mathbf{H}_A$ is the filter matrix to estimate $\mathbf{a}$. The filter matrices are computed from estimates of the PSD matrices for $\mathbf{y}$, $\mathbf{d}$ and $\mathbf{a}$ given by $\mathbf{\Phi}_\mathbf{y} = E\{\mathbf{y}\mathbf{y}^H\}$, $\mathbf{\Phi}_\mathbf{d} = E\{\mathbf{d}\mathbf{d}^H\}$, and $\mathbf{\Phi}_\mathbf{a} = E\{\mathbf{a}\mathbf{a}^H\}$, where $E\{\cdot\}$ is the expectation operator.

As in [10, 13], we assume that

1. D_i and A_j are uncorrelated $\forall i, j$,
2. A_1 and A_2 are uncorrelated,
3. The ambience power is equal in all channels,
4. D_1 and D_2 are correlated and time aligned.

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

[1]For brevity, the time and subband indices are omitted when possible.

Consequently, the PSDs of the audio signals and ambient signals can be expressed as

$$\mathbf{\Phi_y} = \mathbf{\Phi_d} + \mathbf{\Phi_a}, \tag{4}$$

$$\mathbf{\Phi_a} = \phi_A \, \mathbf{I}, \tag{5}$$

where $\mathbf{I}$ is the identity matrix.

We aim at ambient output signals with similar spatial cues as the ambient input signals. In particular, we focus on the inter-channel level difference (ICLD) and ICC of the ambient output signal that are respectively given by

$$\text{aICLD} = \phi_{\hat{A}_1\hat{A}_1}\phi^{-1}_{\hat{A}_2\hat{A}_2} \, , \tag{6}$$

$$\text{aICC} = \phi_{\hat{A}_1\hat{A}_2}(\phi_{\hat{A}_1\hat{A}_1}\phi_{\hat{A}_2\hat{A}_2})^{-1/2} \, , \tag{7}$$

where $\phi_{\hat{A}_i\hat{A}_j}$ refers to elements of the PSD matrix $\mathbf{\Phi_{\hat{a}}} = E\{\hat{\mathbf{a}}\hat{\mathbf{a}}^H\}$. To measure the separation performance for the estimated ambient signals, we define the output direct-to-ambient ratio (DAR) that is obtained using the filter matrix $\mathbf{H}_A$:

$$\text{aDAR} = \frac{\text{tr}\{\mathbf{H}_A^H \mathbf{\Phi_d} \mathbf{H}_A\}}{\text{tr}\{\mathbf{H}_A^H \mathbf{\Phi_a} \mathbf{H}_A\}}, \tag{8}$$

where $\text{tr}\{\cdot\}$ is the trace operator.

3. DIRECT AND AMBIENT SIGNAL EXTRACTION

3.1. Estimation of the ambient signals

A minimum mean squared error (MMSE) estimate of the ambient signals, $\mathbf{a}$, can be obtained by finding a filter matrix $\mathbf{H}_A$ that minimizes

$$J_A = E\left\{\left\|\mathbf{a} - \mathbf{H}_A^H\mathbf{y}\right\|_2^2\right\} = \text{tr}\left\{E\left\{\mathbf{r}_d\mathbf{r}_d^H\right\} + E\left\{\mathbf{q}_a\mathbf{q}_a^H\right\}\right\}, \tag{9}$$

where $\mathbf{q}_a = [\mathbf{I} - \mathbf{H}_A]^H \, \mathbf{a}$ is the distortion of the ambient signals, and $\mathbf{r}_d = \mathbf{H}_A^H\mathbf{d}$ are the residual direct signals. Clearly, the attenuation of the direct signals comes at the expense of distorting the ambient signals. In this work, we like to control the trade-off between the amount of direct signal reduction and ambient signal distortion. Therefore, we introduce a trade-off parameter β ($\beta > 0$) in (9), i.e.,

$$J_A(\beta) = \text{tr}\left\{E\left\{\mathbf{r}_d\mathbf{r}_d^H\right\} + \beta \, E\left\{\mathbf{q}_a\mathbf{q}_a^H\right\}\right\}. \tag{10}$$

By equating the derivative of $J_A(\beta)$ with respect to $\mathbf{H}_A^H$ to zero, we find the PMWF matrix

$$\mathbf{H}_A(\beta) = [\mathbf{\Phi_d} + \beta \, \mathbf{\Phi_a}]^{-1} \, \beta \, \mathbf{\Phi_a}. \tag{11}$$

Given the assumption that lead to (4) and (5) it is known from [3, 10] that an estimate of ϕ_A can be obtained using

$$\hat{\phi}_A = \frac{1}{2}\left(\text{tr}\{\mathbf{\Phi_y}\} - \sqrt{(\phi_{Y_1Y_1} - \phi_{Y_2Y_2})^2 + 4\text{Re}\{\phi_{Y_1Y_2}\}^2}\right). \tag{12}$$

The PSD matrix of the input signals, $\mathbf{\Phi_y}$, can be estimated using recursive averaging

$$\mathbf{\Phi_y}(m) = \alpha \, \mathbf{\Phi_y}(m-1) + (1-\alpha) \, \mathbf{y}(m)\mathbf{y}^H(m), \tag{13}$$

where α ($0 \leq \alpha < 1$) is the forgetting factor that relates to the integration time.

3.2. Estimation of the direct signals

In a similar way as for the ambient signals, we can obtain an estimate of the direct signals, $\mathbf{d}$, and provide a trade-off between the reduction of the ambient signals and the distortion of the direct signals. Therefore, we define the cost function

$$J_D(\beta) = \text{tr}\left\{\beta \, E\{\mathbf{r}_a\mathbf{r}_a^H\} + E\{\mathbf{q}_d\mathbf{q}_d^H\}\right\}, \tag{14}$$

where $\mathbf{q}_d = [\mathbf{I} - \mathbf{H}_D]^H \, \mathbf{d}$ is the distortion of the direct signals, $\mathbf{r}_a = \mathbf{H}_D^H\mathbf{a}$ are the residual ambient signals, and β ($\beta > 0$) is the trade-off parameter. The solution is given by

$$\mathbf{H}_D(\beta) = [\mathbf{\Phi_d} + \beta \, \mathbf{\Phi_a}]^{-1} \, \mathbf{\Phi_d} \, . \tag{15}$$

Using (11) and (15), we can easily verify that

$$\mathbf{H}_D(\beta) + \mathbf{H}_A(\beta) = \mathbf{I}, \tag{16}$$

such that $\hat{\mathbf{d}} + \hat{\mathbf{a}} = \mathbf{y}$.

The i-th column of the filter matrix can be written as

$$\mathbf{h}_{D,i}(\beta) = [\mathbf{\Phi_d} + \beta \, \mathbf{\Phi_a}]^{-1} \, \mathbf{\Phi_d} \, \mathbf{u}_i \, , \tag{17}$$

where $\mathbf{u}_1 = [1\,0]$ and $\mathbf{u}_2 = [0\,1]$ and can be used to estimate the direct signal of the i-th channel. This filter is similar to the PMWF used for speech enhancement.

Assuming that at most one direct sound source is active per time-frequency instant, such that the rank of $\mathbf{\Phi_d}$ can be assumed to be one, we can use the matrix inversion lemma to write (15) as

$$\mathbf{H}_D(\beta) = \frac{\mathbf{\Phi_a}^{-1}\mathbf{\Phi_d}}{\beta + \lambda} = \frac{\mathbf{\Phi_a}^{-1}\mathbf{\Phi_y} - \mathbf{I}}{\beta + \lambda}, \tag{18}$$

where λ is the multichannel direct-to-ambient ratio (DAR)

$$\lambda = \text{tr}\{\mathbf{\Phi_a}^{-1}\mathbf{\Phi_y}\} - 2 \, . \tag{19}$$

Instead of estimating the ambient signals directly, we can first estimate the direct signals using (18) and then use (1) to compute the ambient signals. The main advantage of this procedure is the reduced computational load when inverting the diagonal matrix $\mathbf{\Phi_a}$ instead of the full rank matrix $[\mathbf{\Phi_d} + \beta \, \mathbf{\Phi_a}]$.

4. CONTROL OF THE TRADE-OFF PARAMETER

4.1. Rationale

The derived filters (11) and (15) are equivalent to the MWF used in [10] for $\beta = 1$. Figure 1 illustrates the aICLD and aICC as function of the ICLD of the direct input signal (denoted by iICLD) and of the DAR of the input signal (denoted by iDAR). It is apparent that when iDAR is large, the output signals are panned to the opposite direction of the input signals (upper plot) and have negative ICC of large magnitude (lower plot), which is in contrast to the ideal case according to our model assumptions (aICLD = 0 dB, aICC = 0).

Figure 2 illustrates the spatial cues (aICLD and aICC) obtained for an iDAR of 12 dB as function of iICLD and β. It is apparent that increasing the trade-off parameter to e.g. $\beta = 7$ improves the spatial cues such that aICLD approaches 0 dB and aICC approaches 0. Since the ideal case can not be achieved in all cases or comes at the cost of high residual direct power, our rationale is to control the trade-off parameter such that the absolute values of both, aICLD and aICC, are below perceptually motivated thresholds.

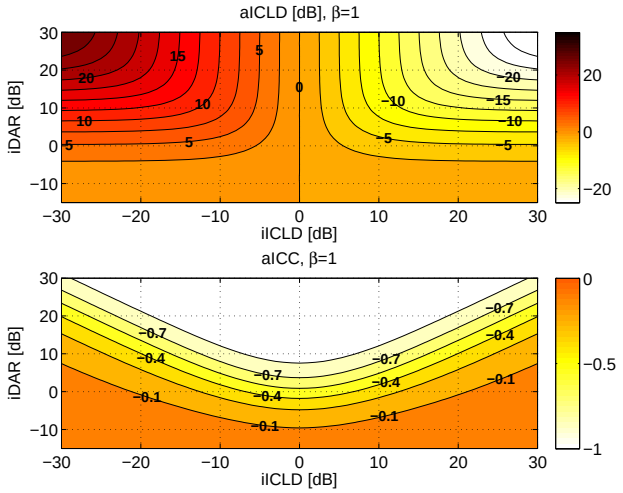

Fig. 1. aICLD (upper plot) and aICC (lower plot) for $\beta = 1$.

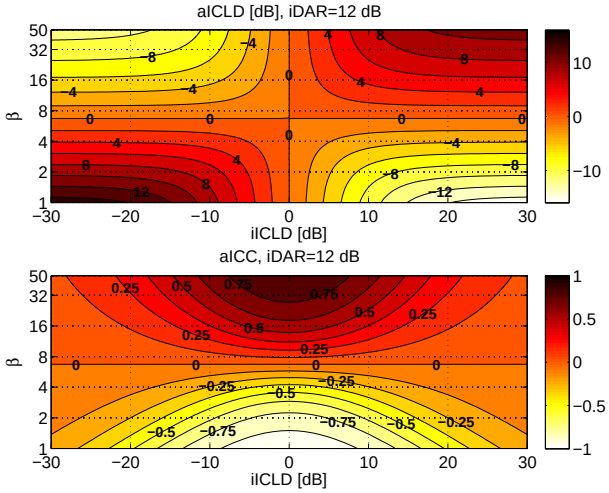

Fig. 2. aICLD (upper plot) and aICC (lower plot) for iDAR=12dB as function of β and the iICLD.

4.2. Psychoacoustic thresholds

The proposed psychoacoustic thresholds are based on the *just-noticeable level variation* (JNLV) for the aICLD and the *just-noticeable difference of interaural correlation* (JNDIC) for the aICC.

Research on the JNLV using amplitude-modulated stimuli indicates that the JNLV largely depends on the stimulus (tone versus noise, center frequency and bandwidth), absolute level and modulation frequency. It ranges between 0.2 dB (for a 1-kHz tone at 100 dB SPL) and slightly below 4 dB (at the threshold of hearing) and is approximately 0.8 dB for white noise when presented with an SPL larger than 30 dB [14].

The JNDIC depends on the ICC of the reference condition and is markedly larger for uncorrelated reference stimuli than for correlated stimuli. The correlation sensitivity for narrowband stimuli in diffuse sound field reference conditions depends on the center frequency and the bandwidth of the stimulus and is smaller for negative than for positive deviations from the reference condition [15].

Experiments with pink noise whose correlation above 500 Hz has been varied between -1 and 1 with a step size of 0.2 reveals that deviations larger than 0.4 were discriminable [16].

4.3. Implementation

As computing the trade-off parameter such that aICLD and aICC are below defined limits is computational complex, we propose in this paper to determine the trade-off parameter using numerical simulation. To this end, we increase the trade-off parameter β until aICLD and aICC are below the psychoacoustic limits $\Lambda_{\text{ICLD}} = 1$ dB (for aICLD) and $\Lambda_{\text{ICC}} = 0.2$ (for aICC). The obtained value for β that fulfills both requirements is stored in a lookup table and referred to as β_{opt}.

To illustrate the effects of the threshold values on β_{opt} separately, we compute β_{ICLD} by considering only Λ_{ICLD}, and similarly we compute β_{ICC} by considering only Λ_{ICC}. Figure 3 shows β_{opt}, β_{ICLD} and β_{ICC} for different iICLD as a function of the iDAR and shows that $\beta_{\text{opt}} = \max(\beta_{\text{ICC}}, \beta_{\text{ICLD}})$. For very small iICLD, the parameter Λ_{ICC} is the deciding factor for β_{opt}. For iICLD = 3 dB, both spatial cues have a similar effect on β_{opt}, whereas for large iICLD $\beta_{\text{opt}} = \beta_{\text{ICLD}}$.

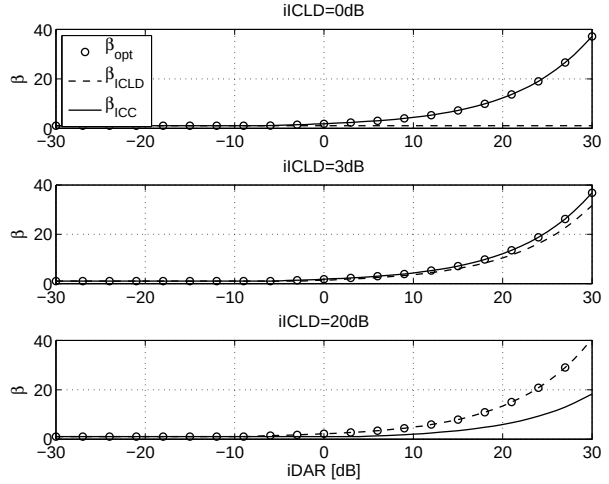

Fig. 3. Trade-off parameters for iICLD = 0 dB, 3 dB, and 20 dB: β_{ICLD} (dashed line), β_{ICC} (solid line) and $\beta_{\text{opt}} = \max(\beta_{\text{ICC}}, \beta_{\text{ICLD}})$ (open circles).

5. PERFORMANCE EVALUATION

Figure 4 shows aICLD, aICC and aDAR for $\beta = \beta_{\text{opt}}$. It confirms that the spatial cues are below the defined thresholds. The aDAR increases when the iDAR increases, and the DAR improvement (i.e. the difference aDAR-iDAR) is better for large iDAR than for small iDAR.

Figure 5 illustrates the example of a commercial recording of 4 seconds length with singing, instrumental accompaniment panned off-center to both sides, and reverberation. The first 60 subbands are shown, corresponding to a frequency range of 2584 Hz. The sum of the auto-PSDs ($\text{tr}\{\mathbf{\Phi}_{\mathbf{y}}\}$), ICC and ICLD of $\mathbf{y}$ are shown in the upper row of plots. The following rows of plots illustrate $\text{tr}\{\mathbf{\Phi}_{\hat{\mathbf{a}}}\}$, the aICC and the aICLD of the output signals obtained using the MWF, the PMWF with $\beta = 10$ and the proposed method with β_{opt}.

The distortions of the spatial cues can be observed for the MWF. The attenuation of the singing (the loud direct signal panned to the

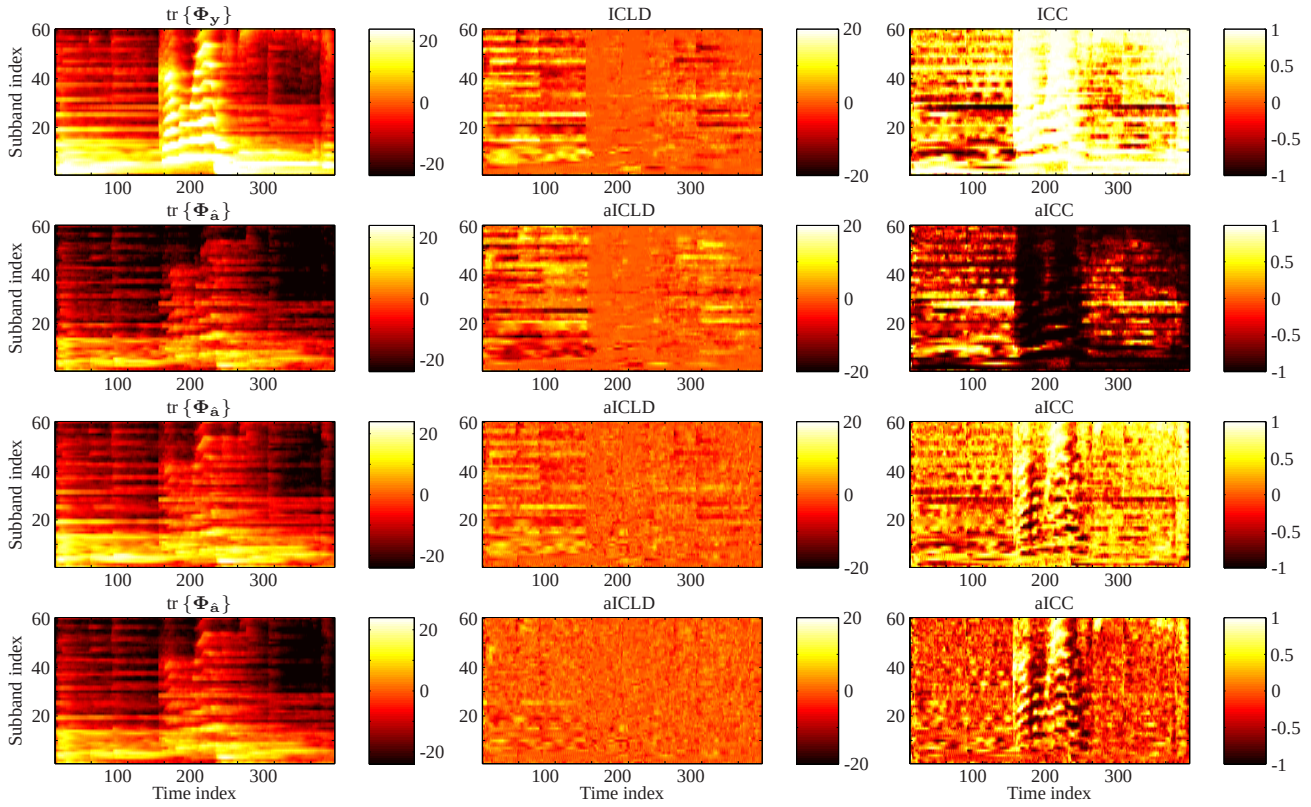

Fig. 5. Input signal (1st row) and ambient output signal obtained with MWF (2nd row), PMWF with $\beta = 10$ (3rd row), and using the proposed method (4th row). The spectrograms tr $\{\mathbf{\Phi_y}\}$ and tr $\{\mathbf{\Phi_{\hat{a}}}\}$ are shown in the range between -24 dB (black) and 24 dB (white) in the left column. ICLDs of $\mathbf{y}$ and $\hat{\mathbf{a}}$ are shown in the range between -20 dB (black) and 20 dB (white) in the middle column. ICCs of $\mathbf{y}$ and $\hat{\mathbf{a}}$ are shown in the range between -1 (black) and 1 (white) in the right column.

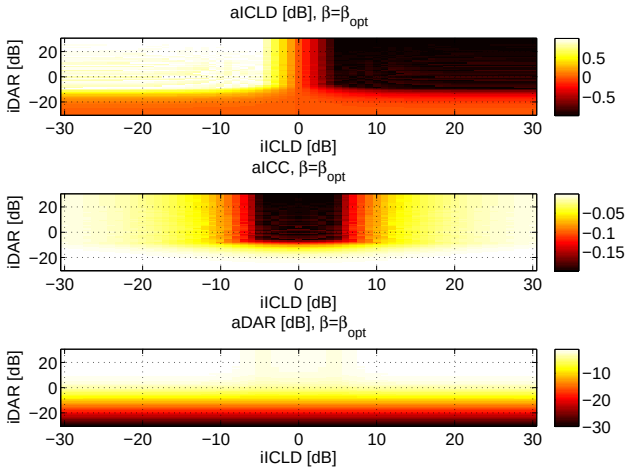

Fig. 4. aICLD, aICC and aDAR obtained with $\beta = \beta_{\text{opt}}$.

center) results in aICC values close to -1. The time-frequency bins corresponding to the direct signals that are panned off-center show aICLD values having opposite signs compared to the ICLD of the input. For the PMWF with $\beta = 10$, the spatial cues are less distorted, but the attenuation of the direct signal is the lowest and the correlation between the ambient output signals is the highest in comparison to the two other methods.

The proposed method yields a compromise, as expected, where the distortions of the spatial cues are small and the attenuation of the direct signal is only slightly worse in comparison to the MWF. It can also be observed that the distortions of the spatial cues exceed the thresholds Λ_{ICLD} and Λ_{ICC}. Experiments with synthetic signals showed that this phenomenon does not occur when processing stationary signals. We hypothesize that it is caused by the non-stationarity of the input signals and the resulting errors in the estimation of the input PSD matrix. However, the distortions of the spatial cues are much smaller for the proposed processing compared to the two other methods.

Informal listening reveals that the ambient output signals obtained with the proposed method have subjectively a good sound quality, and that the described trade-off is audible.

6. CONCLUSIONS

We proposed a method for DAD based on the PMWF that enables a trade-off between the attenuation of the interfering signal and the distortion of the desired signal. We have shown that the MWF introduces distortions of the spatial cues of the ambient output signals. To solve this problem, we control the trade-off parameter of the PMWF such that these distortions are below thresholds of hearing that we derive from the JNLV and the JNDIC.

7. REFERENCES

[1] J.B. Allen, D.A. Berkeley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, 1977.

[2] C. Avendano and J. M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio upmix," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP*, 2002.

[3] J. Merimaa, M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. Audio Eng. Soc. 123rd Conv.*, 2007.

[4] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 2141–2150, 2007.

[5] A. Walther and C. Faller, "Direct-ambient decomposition and upmix of surround sound signals," in *Proc. of IEEE WASPAA*, 2011.

[6] J. Thompson, B. Smith, A. Warner, and J.-M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pairwise correlations," in *Proc. of the AES 133rd Conv.*, 2012.

[7] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using Non-negative Matrix Factorization," in *Proc. Audio Eng. Soc. 30th Int. Conf.*, 2007.

[8] C. Uhle and C. Paul, "A supervised learning approach to ambience extraction from mono recordings for blind upmixing," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2008.

[9] G. Soulodre, "System for extracting and changing the reverberant content of an audio input signal," US Patent 8,036,767, Oct. 2011.

[10] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, 2006.

[11] R. Irwan and R. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, vol. 50, 2002.

[12] M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP*, 2007.

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

[14] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*, Springer, 3rd edition, 2007.

[15] A. Walther and C. Faller, "Interaural correlation discrimination from diffuse field reference correlation," *J. Acoust. Soc. Am.*, vol. 133, no. 3, 2013.

[16] M. Tohyama and A. Suzuki, "Interaural cross-correlation coefficients in stereo-reproduced sound fields," *J. Acoust. Soc Am.*, vol. 85, no. 2, 1989.