FOREGROUND SUPPRESSION FOR CAPTURING AND REPRODUCTION OF CROWDED ACOUSTIC ENVIRONMENTS

Nikolaos Stefanakis [1] *and Athanasios Mouchtaris* [1,2]

[1]Foundation of Research and Technology Hellas, Institute of Computer Science, 70013 Heraklion, Crete, Greece

[2]University of Crete, Department of Computer Science, 71409 Heraklion, Crete, Greece

ABSTRACT

Traditionally, sensor arrays and spatial filtering aim to enhance individual sources by suppressing ambient noise and reverberation. In this paper, the exactly opposite problem is examined, that of suppressing individual sources in favour of the ambient sound and of the whole acoustic scene in general. We consider a compact circular sensor array which is embedded in a crowded ambient acoustic environment and is at the same time prone to interference from directional speech originating from multiple nearby speakers. We propose a method for suppressing the undesired components and we compare its performance with two established approaches in spatial audio processing, namely, direct-to-diffuse decomposition and Primary-Ambient Extraction (PAE). Experimental results and a listening test which are presented illustrate the superiority of our method.

Index Terms— Diffuseness estimation, primary-ambient extraction (PAE), priniciple component analysis (PCA), foreground suppression

1. INTRODUCTION

Several applications related to audio processing benefit from decomposing the information in the audio channels into a directional and a diffuse component. Estimation of the diffuseness of the sound field is useful, for example, to manipulate and reproduce spatial sound [1], to enhance speech by suppressing ambient noise [2] and to extract and enhance reverberation [3]. In spatial audio, it is becoming a common practice to render point-like directional sources and ambient sound differently [4, 5, 1]. This allows for flexible parameterization of the spatial information, which in turn can be exploited for reducing data rate and for alleviating compatibility problems between different reproduction systems.

Commonly, techniques for decomposing the sound field rely on subspace methods or on the Magnitude Squared Coherence (MSC). Techniques belonging in the first family are usually exploited in Primary-Ambient Extraction (PAE) [6, 7, 8]. PAE is used for the analysis and extraction of the audio content in stereo recordings, usually with the purpose of delivering it to a playback system which employs a higher number of reproduction channels. While the mixing conditions in the stereo channels are generally unknown, the main assumption in PAE is that the directional (primary) components in the mix are dominant over the diffuse (ambient) components and that they are coherent within the audio channels. On the other hand, MSC can be used for estimating the diffuseness of the sound field [9, 10, 11] and efficient ways to measure MSC have been suggested by the authors in [12, 13, 14].

This research has been co-financed by the European Union and Greek national funds through the National Strategic Reference Framework (NSRF), Research Funding Program: "Cooperation-2011", Project "SeNSE".

In this paper, we provide a comparative study among different methods for decomposing the observed sound field, considering that the ambient or *background* sound is the only important information that needs to be captured and transmitted to the listener. We consider a compact circular sensor array which is embedded in a crowded acoustic environment and is at the same time subject to interference from multiple nearby speakers. This is a typical scenario which may occur in the capturing and broadcasting of the sound scene in the case of an athletic event. It would be desired, for example, to create a panoramic image of the spectators responses during the game without the inevitable masking that the ones in the foreground may cause to the overall acoustic scene. In what follows, we present the signal model and we illustrate how simple direct-to-diffuse decomposition can be exploited for the purposes of this task. We then propose a novel approach for improving foreground suppression, as opposed to a classical subspace method which is dictated by treating the problem as in PAE.

2. SIGNAL MODEL

We distinguish the sound scene into two basic components which are assumed to be jointly uncorrelated; the *foreground* scene, which constitutes of a small number of directional sources (the foreground sources) at discrete locations in the vicinity of the sensor array, and the *background* scene, which includes the ambience sound as well as the direct path from all the remaining sources which are further away. The analysis is implemented in the short-time frequency domain. Let $X_m(k,i)$ be the signal recorded at the mth sensor at time frame i and discrete frequency k. We can express it as a superposition of a foreground and a background component, $F_m(k,i)$ and $B_m(k,i)$ respectively. By omitting the time-frame index i from now on and by assuming low microphone noise we may write

$$X_m(k) = F_m(k) + B_m(k),\ m = 1, ...M, \tag{1}$$

where M is the number of sensors. We also consider an extension of the same signal model in the subband domain, which is based on grouping of the frequency bins into multiple partitions. In particular, we consider a non-uniform partitioning of $J = 20$ non-overlapping subband regions with corners defined by the frequency indexes $\{1, b_2, ..., b_J, N_{FFT}/2 + 1\}$, where N_{FFT} is the STFT length. The partitioning is based on the Equivalent Rectangular Bandwidth (ERB) and the width of each frequency-subband is approximately 2 ERB [4]. The jth subband region of the mth microphone signal may be then defined as $\mathbf{X}_{m,j} = [X_m(b_j), ..., X_m(b_{j+1} - 1)]^T$ and letting $\mathbf{F}_{m,j}$, $\mathbf{B}_{m,j}$ be accordingly constructed, the previous model may be also written as

$$\mathbf{X}_{m,j} = \mathbf{F}_{m,j} + \mathbf{B}_{m,j},\ m = 1, ..., M. \tag{2}$$

3. DIFFUSENESS ESTIMATION

The diffuseness of a sound field can be estimated with practical microphone setups based on the Magnitude Squared Coherence (MSC) between two microphone signals $X_m(k)$ and $X_n(k)$ as [15]

$$C_{mn}(k) = \frac{|E\{X_m(k)X_n(k)^*\}|^2}{E\{|X_m(k)|^2\}E\{|X_n(k)|^2\}}, \quad (3)$$

where $(\cdot)^*$ denotes complex conjugation and $E\{\cdot\}$ denotes the expectation operator. The minimum of this function is obtained for purely diffuse sound field (close to 0) and the maximum for only direct sound (close to 1). However, when using a compact sensor array, the correlation of the microphone signals at the low frequencies is high, leading to values of MSC close to 1, even if the sound field is purely diffuse. A way for avoiding such a biased estimation has been proposed by Thiergart et al., who define a diffuseness estimator by scaling the measured MSC with respect to a theoretical estimation of diffuse noise coherence [10, 11]. This estimation is nothing else than the theoretical value of the coherence, which, given a particular noise model, would be measured with the actual array geometry and microphone type. For example, assuming spherical isotropic noise and an array of M omnidirectional sensors, the MxM noise coherence matrix $\mathbf{\Gamma}(k)$ can be modelled as

$$\Gamma_{mn}(k) = \frac{\sin(2\pi f_k d_{mn}/c)}{2\pi f_k d_{mn}/c}, \quad (4)$$

where c is the speed of sound, f_k is the frequency in Hertz corresponding to the k-th frequency index and d_{mn} is the distance between sensors m and n. An estimation of diffuseness $\Psi(k)$ at frequency bin k may then be derived as [11]

$$\Psi(k) = \frac{1 - C_{mn}(k)}{1 - \Gamma^2_{mn}(k)}. \quad (5)$$

The diffuseness estimator in Eq. 5 represents a linear scaling of the measured MSC to the range [0,1] such that $\Psi(k) = 1$ in purely diffuse fields and $\Psi(k) = 0$ in non-diffuse fields. To be noticed that the estimated diffuseness value may exceed the theoretical maximum value of 1 when $C_{mn}(k)$ becomes smaller than the assumed minimum $\Gamma_{ij}(k)^2$. In this paper, values of $\Psi(k)$ greater than one are treated as 1.

Assuming that the directional and the diffuse component are mutually uncorrelated, the sound pressure $X_m(k)$ at any sensor can be decomposed into a directional and a diffuse component [11] as

$$X^{dir}_m(k) = \sqrt{1 - \Psi(k)}X_m(k) \quad (6)$$
$$X^{dif}_m(k) = \sqrt{\Psi(k)}X_m(k) \quad (7)$$

where superscripts dir and dif refer to the directional and diffuse part respectively. The presented signal decomposition approach is naturally linked to the problem of foreground suppression, since it is expected that the foreground sound sources will have a dominant direct path and therefore they will be present in $X^{dir}_m(k)$, but absent from $X^{dif}_m(k)$. The diffuse signal component $X^{dif}_m(k)$ may thus be seen as a first solution to the foreground-suppression problem.

4. IMPROVED FOREGROUND SUPPRESSION

Observe that in the last equation, the directional and the diffuse component have different amplitudes but equal phases. In practice, this results to $X^{dif}_m(k)$ being correlated to $X^{dir}_m(k)$, which in turn results to the foreground components being still audible in the diffuse channel. We present in what follows two alternative approaches for alleviating this problem. The basic concept is to derive an estimation of the foreground signal by exploiting the diffuse-to-direct decomposition and then to use this estimation in order remove the foreground components from each microphone signal independently. An important advantage of this approach is that ideally, it will leave the phase and amplitude of the background signal at each microphone unaffected. As a result, any type of spatial filtering technique (e.g. beamforming) may be used for spatial rendering of the background acoustic scene (see Fig. 1(d)).

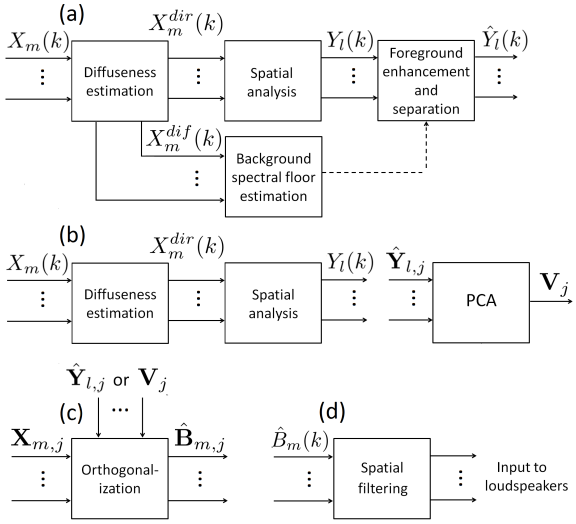

Fig. 1. Block diagram of the foreground estimation approach based on WDO in (a) and on PCA in (b). The orthogonalization process is shown in (c). The estimated background signal at the microphones may be subject to additional spatial rendering for reproduction with a multichannel loudspeaker system as shown in (d).

4.1. Spatial analysis

Following the direct-to-diffuse decomposition of the previous section, the spatial analysis stage considers a set of fixed filter-sum superdirective beamformers which filter the directional signals $X^{dir}_m(k)$ in order to capture the foreground scene. In each time frame i, the beamforming process employs L concurrent beamformers whose look directions are uniformly distributed over the azimuth plane at angles $\theta_l = 360(l-1)/l$ in degrees. Each beamformer steers its beam to one fixed direction yielding in total L signals $Y_l(k) = \sum_{m=1}^{M} w^*_m(k,\theta_l)X^{dir}_m(k)$, $l = 1, ..., L$ in the frequency domain. While a variety of approaches can be used for optimizing the beamformer weights $w_m(k,\theta_l)$, in this paper we have chosen beamformers which maximize the array gain under the assumption of spherically isotropic noise field as [16]

$$\mathbf{w}(k,\theta_l) = \frac{[\epsilon\mathbf{I} + \mathbf{\Gamma}(k)]^{-1}\mathbf{d}(k,\theta_s)}{\mathbf{d}(k,\theta_s)^H[\epsilon\mathbf{I} + \mathbf{\Gamma}(\mathbf{k})]^{-1}\mathbf{d}(k,\theta_s)}, \quad (8)$$

where ϵ is a positive scalar used to satisfy the white noise gain constrain, the MxM matrix $\mathbf{\Gamma}(k)$ is defined in Eq. (4), $\mathbf{w}(k,\theta_l) = [w_1(k,\theta_l), ..., w_M(k,\theta_l)]^T$ is the vector with the beamformer weights and $\mathbf{d}(k,\theta_l) = [e^{-j2\pi f_k \tau_1(\theta_l)}, ..., e^{-j2\pi f_k \tau_M(\theta_l)}]^T$ is

the row vector of phase shifts to align the sensor outputs for a signal from direction θ_l for the specific array geometry. These beamformers are characterized by unity signal response and zero phase shift [16].

4.2. Foreground suppression based on the WDO assumption

The beamformer outputs $Y_l(k)$ are subject to further processing which results to separation of the foreground sources according the their spatial locations. The approach here is based on the assumption of W-Disjoint Orthogonality (WDO), which is a valid assumption for signals with a sparse time-frequency representation such as speech [17, 18]. We basically assume that, at each time-frequency element, there is only one dominant foreground sound source and that it is unlikely that two or more foreground speakers will carry significant energy in the same time-frequency element. Similar as in [19, 20], the WDO assumption is imposed in the foreground channels $\tilde{Y}_l(k)$ through the process

$$\tilde{Y}_l(k) = \begin{cases} Y_l(k), & \text{if } |Y_l(k)| > |Y_{l'}(k)|, \forall l \neq l'; \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Eq. (9) implies that for each frequency element only the corresponding element from one of the beamformer signals is retained, that is, the one with the highest energy with respect to the other signals at that frequency bin. As a consequence, the resulting separated foreground channels have disjoint support and are therefore orthogonal to one another, i.e., $\tilde{Y}_l(k)\tilde{Y}_{l'}(k) = 0$, if $l \neq l'$.

The foreground channels are then subject to an enhancement approach which aims at discarding frequency components whose energy is lower than a specified threshold. This threshold is based on an estimation of the background spectral floor, which is in turn defined by using the diffuse part in the microphone signals $X_m^{dif}(k)$. Rather than calculating it separately at each frequency bin, the background spectral floor is averaged over all frequency bins in the same subband region, forming J spectral floor estimations

$$p_{k\in U_j} = p_j = \sqrt{\frac{1}{b_{j+1} - b_j} \sum_{k\in U_j} |X_1(k)|^2}, \; j = 1, ..., J. \quad (10)$$

Due to the small distance between the sensors, we may assume that the auto spectral densities of the sensors vary trivially from one sensor to the other and the index of the first sensor is here arbitrarily chosen. The separated foreground channels $\tilde{Y}_l(k)$ may thus be further modified as follows

$$\hat{Y}_l(k) = \begin{cases} \tilde{Y}_l(k), & \text{if } \left|\tilde{Y}_l(k)\right| > \mu p_j, \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where $\mu > 0$ is a free scaling parameter and j is the index of the subband region containing the kth frequency bin. Following this step, the resulting time-frequency foreground channels become sparser, in comparison to the previous stage, and the sparsity is depended on the value of μ.

Orthogonalization of the microphone signals with respect to the foreground channels $\hat{Y}_l(k)$ is the final step which completes the process (see Fig. 1(c)). First, the frequency-domain microphone signals and foreground channels are partitioned into the J subbands by grouping of the FFT bins. Due to the orthogonality in the foreground channels, the process may be accomplished independently for each channel, avoiding thus the use of matrix inversion. The procedure, which is repeated for all microphone signals, may be written for the mth microphone signal as

$$\hat{\mathbf{B}}_{m,j} = \mathbf{X}_{m,j} - \sum_{l=1}^{L} \hat{\mathbf{Y}}_{l,j} c_{mjl}, \quad (12)$$

with complex coefficient c_{mjl} resulting from simple orthogonal projection as

$$c_{mjl} = \frac{\hat{\mathbf{Y}}_{l,j}^H \mathbf{X}_{m,j}}{\left\|\hat{\mathbf{Y}}_{l,j}\right\|_2^2}. \quad (13)$$

The signals $\hat{\mathbf{B}}_{m,j}$ are the final output of the process, representing an estimation of the background components at each subband region and microphone. Observe that, in contrast to direct-to-diffuse decomposition, here $\hat{\mathbf{B}}_{m,j}$ is orthogonal to the estimated foreground component $\hat{\mathbf{F}}_{m,j} = \sum_{l=1}^{L} \hat{\mathbf{Y}}_{l,j} c_{mjl}$. The particular approach is named W-disjoint Orthogonality based Foreground Suppression (WDO-FS) and may be schematically represented by subfigures (a) and (c) in Fig. 1.

4.3. PCA based foreground suppression

In [6], it is proposed to extract the primary component from multichannel audio by using Principal Component Analysis (PCA). Use of PCA relies on the assumption that the primary components are dominant over the ambient components and furthermore, coherent within the audio channels. Therefore, these components will emerge by performing some sort of eigenanalysis and by looking into the principle eigenvectors. This concept is adopted in our foreground suppression problem by performing PCA on the output of the spatial analysis stage as follows.

Consider the beamformer outputs at the jth subband region after the spatial analysis step stacked in the single matrix

$$\mathbf{Y}_j = [\mathbf{Y}_{1,j} \; \mathbf{Y}_{2,j} \cdots \mathbf{Y}_{L,j}]. \quad (14)$$

Then the eigenvector $\mathbf{V}_j$ corresponding to the largest eigenvalue of the covariance matrix $\mathbf{R} = \mathbf{Y}_j \mathbf{Y}_j^H$ contains a unit norm version of the primary component. The microphone signals may then be orthogonalized with respect to $\mathbf{V}_j$ as

$$\hat{\mathbf{B}}_{m,j} = \mathbf{X}_{m,j} - \mathbf{V}_j c_{mj}, \; m = 1, ..., M, \quad (15)$$

where now $c_{mj} = \mathbf{V}_j^H \mathbf{X}_{m,j}$. The signals $\hat{\mathbf{B}}_{m,j}$, $m = 1, .., M$ are the final output of the process, representing an estimation of the background component at each microphone. The complete process is named PCA based Foreground Suppression (PCA-FS) and may be schematically represented by subfigures (b) and (c) in Fig. 1.

5. EXPERIMENTAL VALIDATION

Experimental results are presented for recordings produced with a uniform circular array of four omnidirectional microphones of radius $R = 0.02$ m. The recordings constituting the background scene took place in a large reverberant basketball court, during the graduation ceremony of the University of Crete. Both the number of spectators as well as the size of the enclosure were ideal in terms of what can be defined as an "ambient" sound field. A lot of people were talking, cheering and applauding simultaneously, while their distance from the sensors was above 7 meters in most of the cases. Additional microphone signals were simulated by convolving speech signals with the acoustic transfer function corresponding to four speakers located

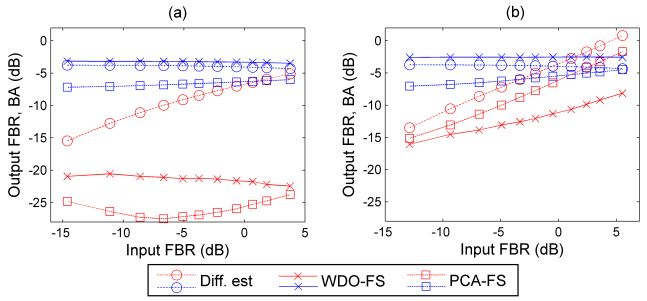

Fig. 2. Output FBR (red) and BA (blue) as a function of the input FBR for 2 foreground speakers in (a) and 4 speakers in (b).

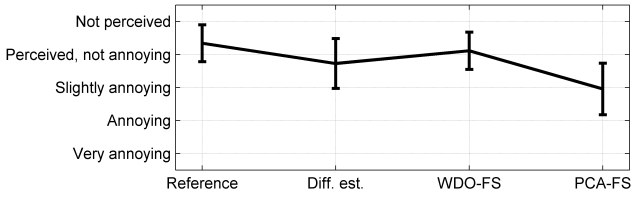

Fig. 3. Perceived quality of the extracted background signal.

at a small distance and at different angles from the array. The exact locations of the individual speakers with respect to the center of the sensor array were at (-0.34,0.94,0.20), (0.92,0.92,0.23), (0.63,-1.36,0.30) and (-0.15,-1.69,0.30) m.

We have designed a simple mixing procedure in order to superimpose the speech signals onto the signals recorded in the basketball court as follows,

$$x_m(n) = b_m(n) + a f_m(n), \quad (16)$$

$$\mathbf{x}_m = \mathbf{b}_m + a\mathbf{f}_m, \quad (17)$$

where $b_m(n)$ and $f_m(n)$ are the time-domain signals at the mth microphone for the real and the synthetic recordings respectively, $\mathbf{b}_m$ and $\mathbf{f}_m$ are the corresponding column vectors by aggregating all samples together and a is a scalar used for varying the balance in the mix. Our intention is to consider b_m and f_m as the background and the foreground component respectively. Although b_m may inevitably contain some directional components, the nature of these recordings is sufficiently different in order to support our investigation.

In order to quantify the performance of foreground suppression, we define the Foreground to Background Ratio (FBR) in the time domain,

$$\text{FBR} = 20\log_{10}(\|a\mathbf{f}_1\|_2 \,/\, \|\mathbf{b}_1\|_2), \quad (18)$$

which can be measured directly at the input stage since both $\mathbf{b}_m$ and $a\mathbf{f}_m$ are known. FBR is measured also in the time-domain output signal of each algorithm, using zero lag cross-correlations. In particular, the FBR is calculated as the ratio of the energy in the output signal which is parallel to $\mathbf{f}_1$ to the energy which is parallel to $\mathbf{b}_1$. An additional criterion examined is the Background Attenuation (BA), which expresses the amount of energy subtracted from the background signal and can be measured at the output signal of each algorithm in a similar manner.

The conditions for this experiment are as follows; we used the overlap-add method with an FFT size of 4096 samples, a frame overlap of 50% and a sampling frequency of 44.1 kHz. The spatial analysis stage consisted of L=8 beamformers uniformly distributed at the azimuth plane and the WDO-FS scaling paremeter μ was set to 1.5. For the computation of the cross-correlations required in Eq. (3) we used a casual recursive formula with a forgetting factor value of 0.35. At each time frame and frequency, the MSC values were calculated for both opposite microphone pairs and the greatest of these two values was used for the calculation of the diffuseness.

Results are presented for 15 seconds of audio duration by plotting the variation of output FBR and BA at the output of each method as a function of the input FBR in Fig. 2. Results are shown for only two speakers active in (a) and for all four speakers in (b). It can be seen that the suppression performance degrades for all techniques when the number of foreground speakers increases, whereas BA is more or less the same. In the case of four speakers, WDO-FS has by far the best suppression performance. Interestingly, the same method also exhibits the least BA values. As expected, simple use of the estimated diffuse component has the weakest performance in terms of suppression among all three techniques. Also, PCA-FS performs better than the latter in terms of suppression, but it produces high attenuation values, meaning that important information from the background is lost.

Listening tests were also conducted in order to evaluate the presented methods. Eleven subjects were asked to judge the sound quality of the output signal with the four speakers in comparison to a reference signal for values of input FBR of -6, -3, 0 and 3 dB. The reference signal, was simply the original signal recorded at the first microphone plus a weighted version of the foreground signal, so that the FBR in the reference signal matches the output FBR of each algorithm. This was in order to ensure that the listeners judge the sound quality of each audio file and not the content. The participants were given a 5 scale grading system, with 1 being "very annoying" difference compared to the reference and 5 being "not perceived" difference from the reference. The mean scores across all subjects and 95% confidence intervals are shown in Fig. 3. It can be seen that WDO-FS and PCA-FS have the best and worst scores respectively, which somehow follows their respective BA values depicted in Fig. 2. To the authors opinion, the rapid variation of the projection coefficients at each time frame in Eqs. (12) and (15) acts as a source of distortion for WDO-FS and PCA-FS, but as the results of the listening test indicate, at reasonable FBR values, this is not perceived at an annoying level for WDO-FS (sounds are available online at http://users.ics.forth.gr/mouchtar/icassp2015/).

6. CONCLUSION

When it comes to capturing and reproduction of crowded acoustic environments, it would be advantageous to suppress sources in the foreground in order to improve the end-user's experience of the overall acoustic event. Although not originally intended for this purpose, diffuseness estimation techniques and PAE may be seen as two existing approaches for addressing the problem. By slightly modifying PAE to operate on a compact sensor array, we showed that a better suppression performance may be achieved, but this performance deteriorates significantly as the number of foreground sources increases. On the other hand, the proposed WDO-FS is more robust to the number of foreground sources and also achieves the best subjective score in terms of sound quality. This research is motivated by the emerging demand in satellite or Web-based sports events broadcasting to deliver a surrounding and immersive experience to the home-user, potentially by giving him the right to select the audiovisual content of his preference from a given list of options.

7. REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, 2007.

[2] N. Ito, N. Ono, E. Vincent, and S. Sagayama, "Designing the wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra," in *Proc. of ICASSP*, 2010, pp. 2818–2821.

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

[4] C. Faller and F. Baumgarte, "Binaural cue coding-part ii: Schemes and applications," *IEEE Trans. on Speech and Audio Process.*, vol. 11, no. 6, pp. 520–531, 2003.

[5] M. Briand, D. Virette, and N. Martin, "Parametric representation of multichannel audio based on principal component analysis," in *AES 120th Conv.*, 2006, paper 6813.

[6] M. Goodwin and J. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. of ICASSP*, 2007, vol. 1, pp. 1–9.

[7] J. He, W. Gan, and E. Tan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. of ICASSP*, 2014, pp. 2892–2896.

[8] J. He, L. Tan, and Gan W., "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 22, pp. 505–517, 2014.

[9] C. Avendano and J. Jot, "A frequency domain approach to mutlichannel upmix," *J. Audio Eng. Soc*, vol. 52, no. 7/8, pp. 740–749, 2004.

[10] O. Thiergart, G. Del Galdo, and A. Habets, "Diffuseness estimation with high temporal resolution via spatial coherence between virtual first-order microphones," in *Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 217–220.

[11] O. Thiergart, M. Kallinger, G. Del Galdo, and F. Kuech, "Parametric spatial sound processing using linear microphone arrays," in *Microelectronic Systems*, pp. 321–329. Springer, 2011.

[12] G. Carter, C. Knapp, and A. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing," *IEEE Trans. on Audio and Electroacoustics*, vol. 21, no. 4, pp. 337–344, 1973.

[13] I. Santamaria and J. Vía, "Estimation of the magnitude squared coherence spectrum based on reduced-rank canonical coordinates," in *Proc. of ICASSP*, 2007, vol. 3, pp. 985–988.

[14] D. Ramirez, J. Vía, and I. Santamaria, "A generalization of the magnitude squared coherence spectrum for more than two signals: definition, properties and estimation," in *Proc. of ICASSP*, 2008, pp. 3769–3772.

[15] B. Cron and C. Sherman, "Spatial-correlation functions for various noise models," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1732–1736, 1962.

[16] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 35, pp. 1365–1376, 1987.

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

[18] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proc. of ICASSP*, 2002, vol. 1, pp. 529–532.

[19] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Capturing and reproducing spatial audio based on a circular microphone array," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013.

[20] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Directional coding of audio using a circular microphone array," in *Proc. of ICASSP*, 2013, pp. 296–300.