AUDIO ZOOM FOR SMARTPHONES BASED ON MULTIPLE ADAPTIVE BEAMFORMERS

Ngoc Q. K. Duong[1], *Pierre Berthet*[2], *Sidkièta Zabre*[3], *Michel Kerdranvat*[1], *Alexey Ozerov*[1], *and Louis Chevallier*[1]

[1] Technicolor, 975 avenue des Champs Blancs, 35576 Cesson Sévigné, France
[2] 3D Sound Labs, 22 rue de la Rigourdière, 35510 Cesson Sévigné, France
[3] Altran Technologies, 3 Rue Louis Braille, 35136 Saint-Jacques-de-la-Lande, France

Email: [1]{quang-khanh-ngoc.duong, firstname.lastname}@technicolor.com,
[2]p.berthet@3dsoundlabs.com,[3]sidkieta.zabre@altran.com

ABSTRACT

Some recent smartphones have offered the so-called *audio zoom* feature which allows to focus sound capture in the front direction while attenuating progressively surrounding sounds along with video zoom. This paper proposes a complete implementation of such function involving two major steps. First, targeted sound source is extracted by a novel approach that combines multiple adaptive beamformers having different look directions with a post-processing algorithm. Second, spatial zooming effect is created by leveraging the microphone signals and the enhanced target source. Subjective test with real-world audio recordings using a mock-up simulating an usual shape of the smartphone confirms the rich user experience obtained by the proposed system.

Index Terms— Audio zoom on smartphone, sound capture, robust adaptive beamformer, post-processing.

1. INTRODUCTION

Mobile devices such as smartphones and tablets have become very popular nowadays for many users. Their hardware and processing power has also been improved day by day, that makes them able to offer more enhanced applications with richer user experience. This paper considers a so-called *audio zoom* application [1, 2] where mobile devices can focus the sound capture on a desired direction while attenuating progressively surrounding sounds[1]. Audio zoom has been commercialized in recent smartphones (*e.g.,* Samsung Galaxy S5 and LG G2), and it would be even more powerful in future products owning larger microphone array.

In order to perform audio zoom, the target sound source needs to be isolated from other surrounding sounds (*i.e.,* interferences originated from unwanted spatial directions) first. Thanks to the hardware improvement where most smartphone nowadays possesses two or more microphones (*e.g.,* Apple iPhone 5 showed up with not one or two, but even three microphones[2]), research in microphone array processing field can be well-applied to this considered problem. Specifically, beamforming [3, 4] and audio source separation (ASS) [5, 6] can be considered as the most appropriate approaches. As ASS generally requires higher computation cost than beamforming algorithms since it usually involves, in addition, the advanced spectral modelling of the audio sources [7]. Thus, by considering the critical constraint of limited processing power in mobile devices, we design our signal enhancement algorithm for the target source grounded on beamforming technique[3]. However, since beamforming usually requires a large microphone array in order to create a narrow beam capturing sound from a desired direction, we propose in this paper a novel approach that combines multiple robust adaptive beamformers [8, 9] with a derived post-processing algorithm taking into account outputs of the beamformers so as to greatly enhance the targeted sound source. Once the target sound source is extracted, we further propose the creation of zooming effect as second step of the audio zoom system. Note that in the considered beamforming implementation, one beamformer has directivity pattern that emphasizes the target source while, on the contrary, the other beamformers suppress the target source. Similar strategy has been presented in [10] with the use of two fixed null beamformers, instead of multiple adaptive beamformers as considered in this paper, and spectral substraction as post-processing algorithm.

The paper aims to design a complete audio zoom system which can be implemented in mobile devices as an emerging application with reasonable processing cost. Yet, to the best of our knowledge, non of the scientific publications has been described such a similar system. It is also worth noting that the proposed approach has been implemented as part of a MediaPlayer running real-time on Android smartphones[4].

The rest of the paper is organized as follows. In Section 2 we present the global workflow as well as the detail steps of the proposed audio zoom system. We conduct experiment with subjective test on real-world sound scene recordings to validate the effectiveness of the proposed approach in Section 3. Finally we conclude in Section 4.

2. PROPOSED AUDIO ZOOM SYSTEM

General workflow of the proposed audio zoom approach is shown in Fig. 1. It consists in two major steps: (1) target sound source extraction and (2) zooming effect creation. These steps will be described in detail in Section 2.1 and Section 2.2, respectively.

This work has been done while the second the third authors were with Technicolor.

[1]https://www.youtube.com/watch?v=7DEyuapmRCs

[2]http://www.idownloadblog.com/2012/09/12/iphone-5-three-mics/

[3]Note that, preliminary study in [7] did not show remarkable advantage of ASS compared to beamforming in some specific setups such as a single target source in noise field.

[4]The demostration will be presented at the Show and Tell session of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)

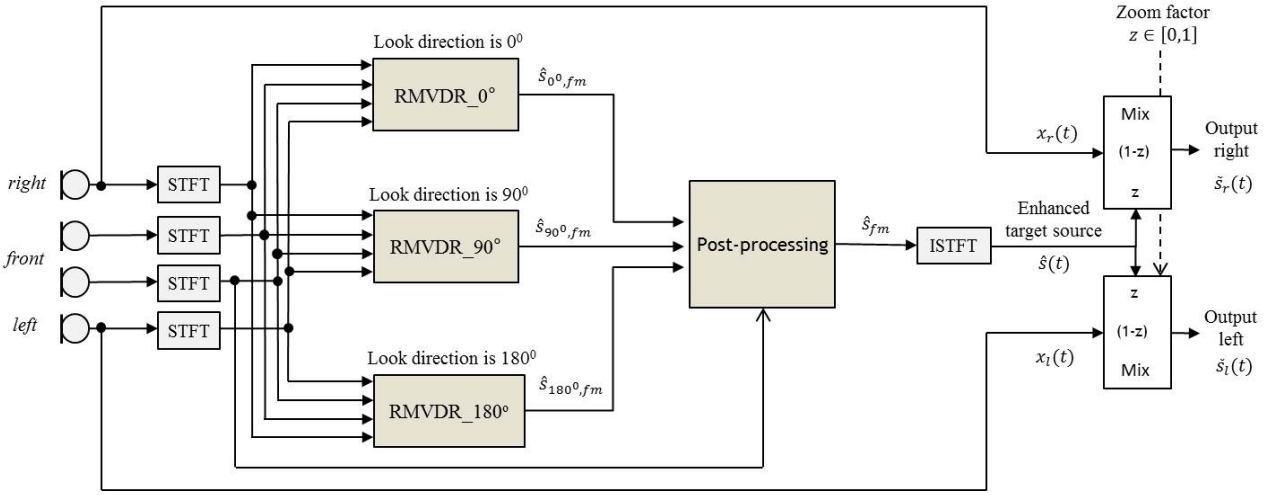

Fig. 1. General workflow of the proposed audio zoom implementation.

2.1. Target sound source enhancement

2.1.1. Robust adaptive beamforming

Let us denote by $\mathbf{x}_{fm} \in \mathbb{C}^{P\times 1}$ the complex-valued STFT coefficients in time frame m and frequency bin f of the mixture signal recorded by P microphones. Beamforming isolates sound coming from a target spatial direction θ_t by deriving a frequency dependent weight vector $\mathbf{w}_{\theta_t,f}$ such that its output is given by

$$\hat{s}_{\theta_t,fm} = \mathbf{w}_{\theta_t,f}^H \mathbf{x}_{fm}, \tag{1}$$

where $(.)^H$ denotes Hermitian transpose. As relevant to the considered audio zoom application, where users usually focus sound capture in the front direction that is perpendicular to the device's surface, in the rest of the paper we consider $\theta_t = 90^0$.

The optimal weight vector $\mathbf{w}_{\theta_t,f}$ can be obtained by minimizing the energy of the interfering sources and noise under the constraint to keep unit response in the target direction. In this derivation, the beamformer is known as Minimum Variance Distortionless Response (MVDR) [8] - a well-known one in the literature and the resulting weight vector is given by

$$\mathbf{w}_{\theta_t,f} = \frac{\mathbf{R}_{i+n,fm}^{-1}\mathbf{d}_{\theta_t,f}}{\mathbf{d}_{\theta_t,f}^H \mathbf{R}_{i+n,fm}^{-1}\mathbf{d}_{\theta_t,f}}, \tag{2}$$

where $\mathbf{R}_{i+n,fm}$ is the interference-plus-noise covariance matrix and $\mathbf{d}_{\theta_t,f}$ is the steering vector which accounts for the time differences of arrival from the target source to the microphones and is computed as

$$\mathbf{d}_{\theta_t,f} = [1, e^{i2\pi f(\tau_{\theta_t 2}-\tau_{\theta_t 1})}, ..., e^{i2\pi f(\tau_{\theta_t P}-\tau_{\theta_t 1})}], \tag{3}$$

where $\tau_{\theta_t p}$ is the the time it takes for the sound to travel from target source at direction θ_t to microphone p ($p = 1, ..., P$) on a direct path.

In practice since $\mathbf{R}_{i+n,fm}$ is unknown, it is often replaced by the sample covariance matrix $\widehat{\mathbf{R}}_{\mathbf{x},fm} = \mathbb{E}[\mathbf{x}_{fm}\mathbf{x}_{fm}^H]$ [8]. A more advanced approach [9], known as robust adaptive beamforming, proposes to estimate this interference-plus-noise covariance matrix by integrating the spatial spectrum distribution over all possible directions containing unwanted signals $\mathbf{\Theta}_{i+n}$ as

$$\widehat{\mathbf{R}}_{i+n,fm} = \int_{\mathbf{\Theta}_{i+n}} \frac{\mathbf{d}_{\theta,f}\mathbf{d}_{\theta,f}^H}{\mathbf{d}_{\theta,f}^H \widehat{\mathbf{R}}_{\mathbf{x},fm}^{-1}\mathbf{d}_{\theta,f}} d\theta, \tag{4}$$

where $\mathbf{d}_{\theta,f}$ is computed similarly to (3) for direction θ.

2.1.2. Proposed implementation of multiple beamformers

In our implementation, in order to reduce the computation cost for smartphone application we first replace the integration (4) by the sum over several unwanted directions (*e.g.,* $\widehat{\mathbf{\Theta}}_{i+n} = 0^0, 45^0, 135^0, 180^0$ when the target direction is $\theta_t = 90^0$). Additionally, we incorporate the diagonal loading technique investigated in [11] to enhance the directivity pattern design. The resulting weight of the proposed beamforming implementation, named robust MVDR (RMVDR), is estimated as

$$\widehat{\mathbf{w}}_{\theta_t,f} = \frac{(\widehat{\mathbf{R}}_{i+n,fm} + \gamma\mathbf{I})^{-1}\mathbf{d}_{\theta_t,f}}{\mathbf{d}_{\theta_t,f}^H(\widehat{\mathbf{R}}_{i+n,fm} + \gamma\mathbf{I})^{-1}\mathbf{d}_{\theta_t,f}}, \tag{5}$$

where $\mathbf{I}$ is the $P \times P$ identity matrix, γ is a loading factor preventing instability [11], and the interference-plus-noise covariance matrix is computed by

$$\widehat{\mathbf{R}}_{i+n,fm} = \sum_{\theta\in\widehat{\mathbf{\Theta}}_{i+n}} \frac{\mathbf{d}_{\theta,f}\mathbf{d}_{\theta,f}^H}{\mathbf{d}_{\theta,f}^H \widehat{\mathbf{R}}_{\mathbf{x},fm}^{-1}\mathbf{d}_{\theta,f}}. \tag{6}$$

Since equipping a large microphone array for a smartphone so that beamforming can isolate well the target source is not feasible in practice, we further propose to use multiple RMVDR where one of them enhances the target source (*i.e.,* RMVDR_90^0) and the others enhance unwanted sound coming from other directions (*e.g.,* RMVDR_0^0 and RMVDR_180^0, these beamformers have look directions perpendicular to the desired one). This implementation is depicted in Fig. 1 for the case when three RMVDRs are used. Note that the overall computational cost does not increase linearly with respect to the number of RMVDRs used since the sample covariance matrix $\widehat{\mathbf{R}}_{\mathbf{x},fm}$ needs to be computed once, and similarly steering vectors $\mathbf{d}_{\theta,f}$ needed in (5) and (6) can be shared between RMVDRs.

We will discuss the post-processing of the outputs of these beamformers so as to further isolate the target sound source, compared to the conventional case where only RMVDR_90^0 is used, in Section 2.1.3.

2.1.3. Proposed post-processing algorithm

Denoting by $\hat{s}_{\theta_t,fm}$ and $\hat{s}_{\theta,fm}$ the output of RMVDRs looking at the target direction θ_t and other directions $\theta \neq \theta_t$, respectively. As example in our setting shown in Fig. 1, $\theta_t = 90^0$ while $\theta = 0^0$ or 180^0. However, one can easily extend the algorithm with the use of more RMVDRs and any desired direction than the 90^0. We propose to compute the STFT coefficients of the post-processed output signal for audio zoom as

$$\hat{s}_{fm} = \begin{cases} x_{p,fm} & \text{if } |\hat{s}_{\theta_t,fm}| > \alpha \max\{|\hat{s}_{\theta,fm}|, \forall \theta \neq \theta_t\} \\ \beta_{fm}\hat{s}_{\theta_t,fm} & \text{otherwise} \end{cases} \tag{7}$$

where $|.|$ denotes the absolute value, p denotes a reference microphone signal such as $p = 2$ for a front microphone in our setting, $\alpha > 1$ is a tuning constant, and

$$\beta_{fm} = \frac{1}{\epsilon + \frac{\max\{|\hat{s}_{\theta,fm}|, \forall \theta \neq \theta_t\}}{|\hat{s}_{\theta_t,fm}|}} \tag{8}$$

where ϵ is a constant (*e.g.,* $\epsilon = 1$).

Our derivation to equations (7) and (8) is motivated by the well-known observation that the sound sources are usually non-overlapped in the time-frequency (T-F) domain. As can be seen from the first line of (7), for time-frequency (T-F) points where the estimated target source is really dominant than the others, we take signal from a front microphone $x_{p,fm}$ as the final output so as to maximize the sound quality[5]. In this case, a reference microphone signal is a good estimate of the target source since other sources are considered to be inactive. Otherwise, the estimated target STFT coefficients $\hat{s}_{\theta_t,fm}$ will be considered. The derivation to equation (8) can be explained by the fact that in T-F points where sound from non-desired directions is really dominant (*i.e.,* $\max\{|\hat{s}_{\theta,fm}|, \forall \theta \neq \theta_t\} >> \hat{s}_{\theta_t,fm}$), the target source $\hat{s}_{fm}$ should be considered as inactive. Thus its value should close to 0 as β_{fm} will be very small. In neutral case where none of the estimated sources is really dominant, the smaller $\hat{s}_{\theta_t,fm}$ compared to the other sources, the more amplification it should be, as β_{fm} increase, in order to further improve the designed zooming effect as presented in Section 2.2. Finally, the time domain signal $\hat{s}(t)$ of the enhanced target source is obtained by the inverse STFT of $\hat{s}_{fm}$.

2.2. Proposed audio zoom effect creation

Let us denote by $z \in [0, 1]$ the zooming factor where the higher value of z the more target sound source is focused, and $z = 1$ corresponds to the maximum zoom (*i.e.,* 100%). In order to maintain spatial effect of the perceived stereo output signal, we propose to mix the estimated target source after the post-processing $\hat{s}(t)$ with the original signals recorded by left and right microphones, denoted as $x_l(t)$ and $x_r(t)$, respectively. The final left and right channels of the output signal, denoted by $\tilde{s}_l(t)$, and $\tilde{s}_l(t)$, respectively, are computed as

$$\tilde{s}_l(t) = z * \hat{s}(t) + (1 - z) * x_l(t), \tag{9}$$

$$\tilde{s}_r(t) = z * \hat{s}(t) + (1 - z) * x_r(t). \tag{10}$$

It can be seen that there is no zooming effect when $z = 0$, and when z increases the estimated target source $\hat{s}(t)$ contributes more to the output signal as it should be more progressively focused. In case of maximum zoom with $z = 1$, both output channels take the same value (*i.e.,* $\tilde{s}_l(t) = \tilde{s}_r(t) = \hat{s}(t)$) so that the user can experience spatial effect of the isolated sound as if it comes from the front direction ($\theta_t = 90^0$) and the target sound source is most focused.

3. EXPERIMENTS

We fist describe the recording setup in Section 3.1. We then present the algorithm implementation and result of the subjective test where different users experienced audio zooming effect created by the proposed approach in Section 3.2.

3.1. Experiment setup

In order to make a test close to the real situation, we built a mock-up containing four microphones mimicking a smartphone as shown in Fig. 2. In this setting, two microphones are located at the top and bottom of the mock-up as usual with most available smartphones, two other microphones are located at the back side so as to ease sound capture during the video recording. The detail (x,y,z) coordinates of these microphones, measured in centimeter, are (6.5, 2, 0.5); (3.3, 0, 0); (-0.033, 0, 0); (-6.5, 2, 0.5), respectively.

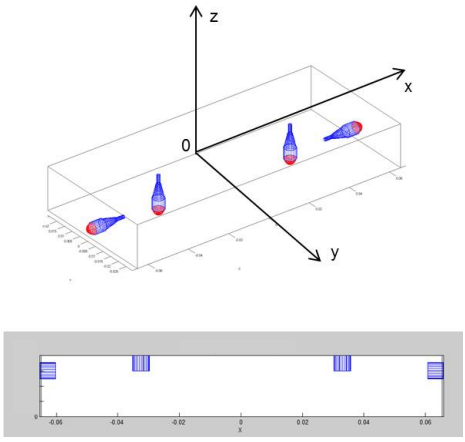

Fig. 2. Mock-up with 4 microphones for the experiment.

We performed two 40 second length indoor audio recordings without video capture. The setups are shown in Fig. 3(a) and Fig. 3(b), respectively, where M_1 and M_2 are two musical instruments while S_1 and S_2 are two speeches. In both cases, audio zoom algorithm aims to enhance two sound sources located near the center while progressively attenuating two other unwanted sources. For a more realistic evaluation of the user perception when audio zoom is performed together with video zoom, we made an additional outdoor recording in a park as shown in Fig. 3(c) where audio and video is captured together. The recording duration is 90 seconds and audio zoom algorithm aims to focus on the bird song while canceling surrounding sounds including human walking, speech, environmental wind, etc.,.

[5]Note that in the output of RMVDR there is usually some artifact due to the nonlinear processing, and the signal distortion is more severe at high frequencies where the array's geometry error has more impact.

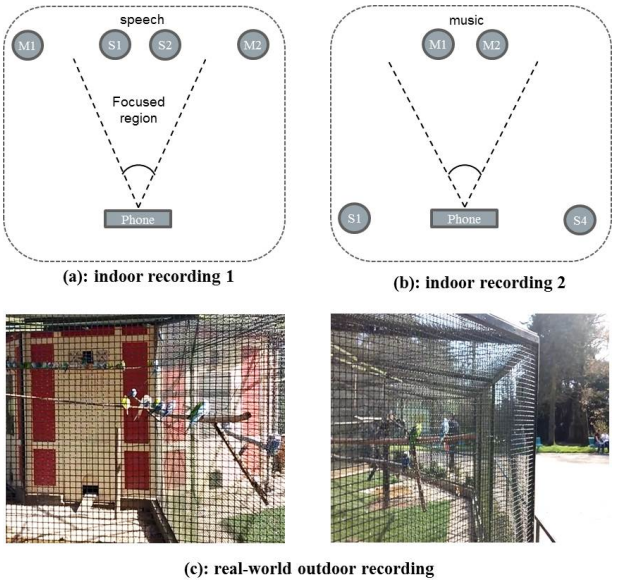

Fig. 3. Experiment setup for user test on audio zoom feature.

Setup	Preference	Indoor 1	Indoor 2	Outdoor
Test 1	Baseline (B)	8	6	5
	Similar quality	4	2	3
	B + OMLSA	1	5	5
Test 2	Baseline	2	1	2
	Similar quality	0	2	1
	Proposed	**11**	**10**	**10**
Test 3	Proposed (P)	6	5	7
	Similar quality	1	3	0
	P + Energy boost	6	5	6

Table 1. Results with subjective listening tests performed by 13 users.

3.2. Result with subjective test

We developed an application with a friendly graphical user interface (GUI) so as user can perform audio zoom and experience the audio quality obtained by different implementations via a headphone. It is worth noting that we prefer the real-world user test than the objective evaluation since the former case is more relevant to the target application. We invited 13 people at different ages to participate in three different listening tests where each of them was asked to indicate which algorithm yields better zooming experience, or they offer the similar quality in his/her opinion. The results for three test cases, performed in double-blind fashion, and for each recording condition are shown in Table 1 where the value means the number of rated users in each option. Note that in all tests, the second step for creating audio zooming effect is implemented similarly for all other approaches under comparison.

We first validate whether the state-of-the-art post filtering technique brings some benefit when it is implemented after beamforming as a standard way [7] in the "Test 1". For this purpose, we asked users' opinion when they experienced results obtained by the baseline robust adaptive beamformer (RMVDR_90^0) and that obtained by the RMVDR_90^0 followed by the well-known Optimal Modified Minimum Mean-Square Error Log-Spectral Amplitude (OMLSA)[6] post-filtering algorithm (named "B+OMLSA" in Table 1). Note that other post-filters (*e.g.,* Zelinskis [12] and McCowans [13]) can also be tested, but as observed in [7] that they did not bring benefit compared to OMLSA, we consider OMLSA as the state-of-the-art post filter in our implementation and test. As can be seen, for indoor recording more users prefer not to use OMLSA since it brings additional signal distortion. For outdoor recording, even though OMLSA really suppresses more background diffuse noise, it still does not bring benefit in the test. This listening test is actually coherent with the observation in [7] that MVDR+OMLSA adds further signal distortion compared to MVDR alone so as user perceives more artifact.

The "Test 2" aims to compare the proposed approach, *i.e.* three RMVDR having look directions of 0^0, 90^0, and 180^0, respectively, and post-processing as shown in Fig. 1, (named "Proposed"), with the state-of-the-art beamforming approach using one RMVDR_90^0. Note that we did not compare to the case where RMVDR_90^0 is followed by OMLSA here since it has been shown in the Test 1 that users prefer RMVDR_90^0 alone. We also implemented a method using null-beamformers and post-processing algorithm described in [10], but subjectively observed that it performs poorer than the two considered algorithms, so we did not formally perform user test with it in order to avoid too much listening for users. As can be seen in Table 1, most users prefer the audio zoom quality obtained by the proposed approach in all three recording conditions. As example, for the real-world outdoor recording where audio zoom was performed together with video zoom to maximize the user experience, 10 users prefer the result of the proposed approach while only 2 users prefer the result of the baseline. It is also worth noting that our informal listening test in case of using two microphones, instead of four, also shares the same experience that the proposed approach performs better than the others.

The final test was devoted to the zooming effect only where we want to validate if increasing the volume of the enhanced signal can improve overall user experience. Thus we compare the "proposed" with a case where the enhanced signal after beamforming and post processing $\hat{s}(t)$ is boosted by 6dB energy before mixing with the original microphone signals in the zooming creation step. The result is shown in "Test 3". Surprisingly, overall performance for three recording conditions shows that user experience is generally not improved as expected when increasing volume of the target sound. This can be explained by the fact that $\hat{s}(t)$ still contains noticeable distortion so that when its volume increases users also perceive more artifacts.

4. CONCLUSION

In this paper, we have presented a novel approach for performing audio zoom, an emerging application, in mobile devices with low computation cost. The proposed implementation combines several robust adaptive beamformers with a derived post-processing algorithm to further enhance the targeted sound source. We also describe the design of zooming effect so as to improve the user perceptual experience. Subjective tests with both real-world indoor and outdoor recordings confirm the effectiveness of the derived approach. Future research would be devoted to perform a formal objective evaluation where ground truth is available. Additionally, the investigation of audio source separation based approach where the target direction can be taken into account as prior information [14] would be potential.

[6]Matlab code is available at:
http://webee.technion.ac.il/Sites/People/IsraelCohen/Download/omlsa.m

5. REFERENCES

[1] C. Avendano and L. Solbach, "Audio zoom," US Patent Submitted 20 110 129 095A1, 2011. [Online]. Available: http://www.google.com/patents/US20110129095

[2] K. Lee, H. Song, Y. Lee, Y. Son, and J. Kim, "Mobile terminal and audio zooming method thereof," US Patent Submitted 20 130 342 730A1, 2013. [Online]. Available: http://www.google.com/patents/US20130342730

[3] B. V. Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988.

[4] J. Li and P. Stocia, *Robust adaptive beamforming*. Eds. New York: Wiley, 2005.

[5] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[6] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[7] J. Thiemann and E. Vincent, "An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement," in *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–5.

[8] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*. Springer Verlag, 2010, ch. 2, pp. 19–38.

[9] Y. Gu and A. Leshem, "Robust adaptive beamforming based on interference covariance matrix reconstruction and steering vector estimation," *IEEE Trans. on Signal Processing*, vol. 60, no. 7, pp. 3881–3885, 2012.

[10] S. Takada, S. Kanba, T. Ogawa, K. Akagiri, and T. Kobayashi, "Sound source separation using null-beamforming and spectral subtraction for mobile devices," in *Proc. IEEE Workshop on Aplications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 30 – 33.

[11] X. Mestre and M. Lagunas, "On diagonal loading for minimum variance beamformers," in *Proc. IEEE Int. Symp. on Signal Processing and Information Technology (ISSPIT)*, 2003, pp. 459–462.

[12] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, pp. 2578–2581.

[13] I. A. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 1905–1908.

[14] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, no. 1, pp. 1–11, 2013.