Robust Source Localization in Reverberant Environments Based on Weighted Fuzzy Clustering

Marco Kühne, Roberto Togneri, *Senior Member, IEEE*, and Sven Nordholm, *Senior Member, IEEE*

***Abstract*—Successful localization of sound sources in reverberant enclosures is an important prerequisite for many spatial signal processing algorithms. We investigate the use of a weighted fuzzy c-means cluster algorithm for robust source localization using location cues extracted from a microphone array. In order to increase the algorithm's robustness against sound reflections, we incorporate observation weights to emphasize reliable cues over unreliable ones. The weights are computed from local feature statistics around sound onsets because it is known that these regions are least affected by reverberation. Experimental results illustrate the superiority of the method when compared with standard fuzzy clustering. The proposed algorithm successfully located two speech sources for a range of angular separations in room environments with reverberation times of up to 600 ms.**

***Index Terms*—Fuzzy clustering, microphone array, reverberation, source localization.**

I. Introduction

ACOUSTIC source localization by means of a microphone array is still an active field of research. It is the task of extracting the localization information of one or several sound sources by sampling the sound field through a number of spatially distinct microphones. One important application for source localization algorithms can be found in the field of blind source separation (BSS). Specifically, under-determined BSS strategies that rely on source sparseness frequently exploit localization information, such as directions-of-arrival (DOA), for segmenting the time-frequency (T-F) plane into disjoint regions each assigned to a particular source [1], [2]. For example, Araki *et al.* [2] use the k-means algorithm to perform the partitioning of the T-F plane into a number of clusters, which is assumed to be known *a priori*. For stationary sources, each of the cluster centers corresponds to the location parameter of a particular source. Each cluster membership matrix indicates to which degree a T-F point belongs to a particular source and can be interpreted as a T-F mask. Individual sources are separated from the mixture by selecting all T-F points belonging to the source's cluster as indicated by its membership mask. Location cues are most reliable in anechoic environments, where almost all T-F points contribute observations that are effective for clustering. However, for reverberant mixtures, the location cues become increasingly corrupted. This often leads to incorrect localization and partitioning results, mostly due to falsely detected cluster centers, e.g., source locations.

In this letter, we address this issue by presenting a robust source localization technique based on a weighted fuzzy c-means algorithm. Contrary to [2], where every observation is treated as equally important, we deal with imperfect location cues by indicating their usability for the clustering process. The observation weights are obtained prior to the clustering by scanning the T-F plane around sound onsets for regions with low DOA fluctuations. This was motivated by a number of previous studies which have shown that in echoic enclosures, only a small fraction of the location cues correspond to the correct source locations. For example, Faller and Merimaa [3] showed that binaural source localization remains feasible, even in highly reverberant conditions by selecting cues consistent with an interaural coherence measure. Huang *et al.* [4] emulated the precedence effect by concentrating on location cues extracted from sound onsets, a concept that is known to be exploited by the human auditory system [5]. Through computer simulations, we show that the proposed algorithm is successful in locating two speech sources for a range of angular separations in room environments with reverberation times (RT_{60}) of up to 600 ms.

The remainder of this letter is organized as follows. Section II describes the proposed localization algorithm in more detail and illustrates the importance of observation weighting for reverberant mixtures. Section III describes the experimental protocol and presents the results for a number of simulated source localization experiments. Finally, the letter concludes with a short summary in Section IV.

II. Robust Source Localization Algorithm

A. Mixing Model and Sparseness Assumption

Consider N sources in a reverberant enclosure impinging on a uniform linear microphone array (ULA) made up of M identical, omnidirectional sensors with inter-element spacing d (Fig. 1). The source positions are assumed to be stationary in the median plane at azimuth angles $\theta_1, \ldots, \theta_N$. It is further assumed that each microphone observation can be represented in the frequency domain as an instantaneous mixing model

$$X_m(k,l) \approx \sum_{n=1}^{N} H_{mn}(l)\, S_n(k,l), \quad m = 1, \ldots, M \qquad (1)$$

where k represents a time index, l is a frequency index, and $H_{mn}(l)$ is the room impulse response from source n to sensor m. $X_m(k,l)$ and $S_n(k,l)$ are the short-time Fourier transforms (STFTs) of the mth microphone observation and nth source defined on a T-F grid by the lattice spacing parameters (τ_0, ω_0). A common assumption for speech signals [1], [2] is that for each

Manuscript received August 02, 2008; revised October 09, 2008. Current version published January 08, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jen-Tzung Chien.

M. Kühne and R. Togneri are with the School of Electrical, Electronic, and Computer Engineering, The University of Western Australia, Crawley 6009, Australia (e-mail: marco@ee.uwa.edu.au; roberto@ee.uwa.edu.au).

S. Nordholm is with the Western Australian Telecommunications Research Institute, a Joint Venture Between The University of Western Australia and Curtin University of Technology, Crawley 6009, Australia (e-mail: sven@watri.org.au).

Digital Object Identifier 10.1109/LSP.2008.2009833

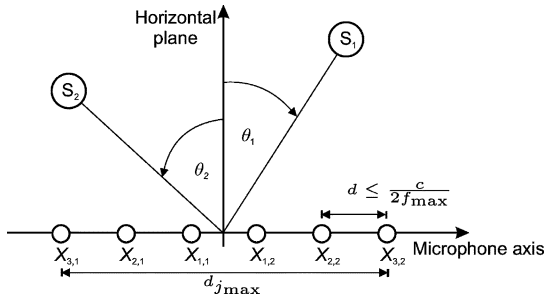

Fig. 1. Uniform linear microphone array with $j \in \{1, 2, 3\}$ sensor pairs $(X_{j,1}, X_{j,2})$ and two sources S_1, S_2 located at azimuth angles θ_1 and θ_2.

T-F point, only one arbitrary source S_n will be active, such that (1) simplifies to

$$X_m(k,l) \approx H_{mn}(l)\, S_n(k,l). \tag{2}$$

Note that assumptions (1) and (2) become increasingly unrealistic for short STFT window lengths and long reverberation times due to strong reflections from preceding sound events.

B. Spatial Feature Extraction

The most commonly adopted location feature is based on the estimation of the time delay between two microphones using the generalized cross-correlation or cross-power spectrum phase [6]. However, within the BSS framework of T-F masking, it is more common to use level ratios and/or phase differences which produce instantaneous location estimates for each T-F cell [1], [2]. According to [7], for sparse sources in echoic environments, the longer the distance d_j is between a sensor pair $(X_{j,1}, X_{j,2})$ (Fig. 1), the better the DOA localization performance will be. Hence, the instantaneous DOA at T-F point (k,l) is computed as the normalized phase difference

$$\psi(k,l) = -\frac{1}{l\omega_0 d_{j_{\max}} c^{-1}} \arg\left[\frac{X_{j_{\max},1}(k,l)}{X_{j_{\max},2}(k,l)}\right] \tag{3}$$

where $j_{\max}$ denotes the index of the sensor pair with the longest distance $d_{j_{\max}}$ and c is the propagation velocity of sound [2], [7]. However, when $d_{j_{\max}} > c/2f_{\max}$, with $f_{\max}$ being the signal's maximum frequency, the sensor pair violates the spatial aliasing theorem and the phase values in (3) become ambiguous. To avoid this problem, we employ the SPIRE algorithm [7] which utilizes the smaller non-aliased distance pairs to restore the aliased values of the longer distance pairs. SPIRE is applicable when multiple microphone pairs are available and at least one sensor-pair distance is shorter than the aliasing distance. Note that ψ can be converted to its equivalent azimuth angle via $\theta = \arcsin(\psi)$. The frequency normalization in (3) avoids the permutation problem usually encountered in frequency domain BSS and ensures that for short data, enough DOA measurements are available for clustering [2].

C. Fuzzy Clustering of DOA Values

A weighted fuzzy c-means (wFCM) algorithm [8] is then used for grouping the extracted features ψ into N clusters. In wFCM, clustering is achieved by minimizing the cost function

$$J = \sum_{\forall(k,l)} \sum_{n=1}^{N} u_n^q(k,l) w(k,l) \parallel \psi(k,l) - \hat{\psi}_n \parallel^2 \tag{4}$$

where $u_n(k,l) \in [0,1]$ represents the membership of $\psi(k,l)$ in the nth cluster, $w(k,l)$ is the observation weight for $\psi(k,l)$, $\hat{\psi}_n$ is the nth cluster center, and $\|\cdot\|$ is a distance metric, such as the L_2 norm. The parameter $q > 1$ controls the softness of the clustering and is fixed here to $q = 2$. Starting from a random partitioning, the cost function (4) is iteratively minimized by alternating the updates for centers and memberships

$$\hat{\psi}_n = \frac{\sum_{\forall(k,l)} u_n^q(k,l) w(k,l) \psi(k,l)}{\sum_{\forall(k,l)} u_n^q(k,l) w(k,l)} \tag{5}$$

$$u_n(k,l) = \left[\sum_{j=1}^{N} \left(\frac{\parallel \psi(k,l) - \hat{\psi}_n \parallel}{\parallel \psi(k,l) - \hat{\psi}_j \parallel}\right)^{2/(q-1)}\right]^{-1} \tag{6}$$

until an appropriate termination criterion is met. While the final centroids correspond to the DOA estimates, the membership matrices can be used as soft T-F masks in BSS applications [1], [2]. Note that wFCM defaults to the standard FCM clustering if the weights are chosen to be unity.

D. Observation Weights

The observation weights w are used to emphasize sound onsets and regions with low DOA fluctuations. The steeper an onset and the lower the local variance for a DOA measurement, the more weight should be given to this observation during clustering. In particular, the weights are computed as

$$w(k,l) = (l\omega_0)^2 w_{\text{ons}}(k,l) w_{\text{var}}(k,l) \tag{7}$$

where w_{ons} denotes the onset component and w_{var} are the weights based on the local DOA variance around $\psi(k,l)$. The motivation for (7) is that reliable DOA measurements are often found at echo-free sound onsets [4] and single source areas with low local DOA variances [9]. High variances, however, indicate regions where sources overlap or where reflections contaminate the DOA measurements. The additional term $(l\omega_0)^2$ gives more weight to high frequencies because the localization accuracy in low frequencies is often severely degraded by reverberation. The weighting mechanism is particularly helpful in reverberant conditions because it favors T-F points that better satisfy the sparseness assumption (2).

To determine w_{ons}, a simple onset weighting scheme is implemented. Let the instantaneous power be defined as in [1]

$$E(k,l) = |X_{j_{\max},1}(k,l) X_{j_{\max},2}(k,l)| \tag{8}$$

and let its first-order time difference on the log-scale be

$$o(k,l) = \log\left[\frac{E(k,l)}{E(k-1,l)}\right] = \log[E(k,l)] - \log[E(k-1,l)]. \tag{9}$$

After smoothing o with a 5×5 median filter, the onset weights are determined via the sigmoid compression

$$w_{\text{ons}}(k,l) = \frac{1}{1 + \exp\{\alpha_1[o(k,l) - \beta_1]\}} \tag{10}$$

where α_1 is the sigmoid slope and β_1 is the sigmoid center. Both parameters can be tuned, such that sound offsets with small or negative $o(k,l)$ values are suppressed and onsets with large positive $o(k,l)$ are emphasized. The second weight component is

derived from local DOA statistics gathered in a small neighborhood $N_{(k,l)}$ around each DOA measurement. Let the local DOA mean $\mu_\psi(k,l)$ around $\psi(k,l)$ be

$$\mu_\psi(k,l) = \frac{1}{|N_{(k,l)}|} \sum_{\forall(k',l')\in N_{(k,l)}} \psi(k',l') \tag{11}$$

and the local DOA variance $\sigma^2_\psi(k,l)$ around $\psi(k,l)$ be

$$\sigma^2_\psi(k,l) = \frac{1}{|N_{(k,l)}|-1} \sum_{\forall(k',l')\in N_{(k,l)}} |\psi(k',l') - \mu_\psi(k,l)|^2 \tag{12}$$

where the neighborhood $N_{(k,l)} := \{(k',l') : k' = k \wedge |l'-l| \leq P\}$ is chosen in this study as a nine-point window of adjacent frequency bins, e.g., $P = 4$. The variances are then mapped to the [0, 1] interval using the sigmoid function

$$w_{\text{var}}(k,l) = \frac{1}{1+\exp\{\alpha_2(\log[\sigma^2_\psi(k,l)] - \beta_2)\}} \tag{13}$$

where α_2 and β_2 are the sigmoid slope and center parameter, respectively. Again, both parameters can be tuned, such that T-F points with large DOA variances are suppressed and areas with low DOA fluctuations are emphasized. Optimal selection of the sigmoid parameters is beyond the scope of this letter, and therefore, all parameters are empirically derived through a series of tuning experiments (see Section III).

Fig. 2 illustrates the importance of the proposed observation weighting for a two-source configuration. For anechoic conditions, as shown in Fig. 2(a), almost all of the observations $\psi(k,l)$ in the T-F plane are reliable for clustering. The corresponding azimuth histogram with unity weights in Fig. 2(b) clearly shows two distinctive peaks close to the true DOA angles. However, as evident from Fig. 2(c), in reverberant conditions, only a small fraction of the DOA observations remain within a localization error of 5°. Consequently, the azimuth histogram with unity weights in Fig. 2(d) fails to identify the two sources. Only when the observations are weighted according to their reliability $w(k,l)$ do the two sources become visible again [Fig. 2(e) and (f)].

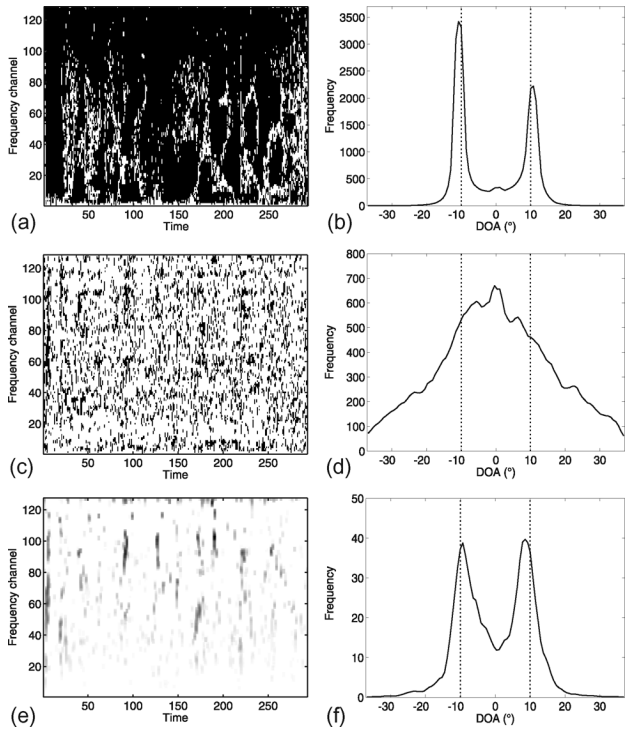

Fig. 2. Example of reliable localization information for two sources located at $-10°$ and $10°$ under (a)-(b) anechoic and (c)-(f) reverberant conditions with $\text{RT}_{60} = 600$ ms. (a) Anechoic DOA observations $\psi(k,l)$ with max. of 5° localization error (black points). (b) Anechoic azimuth histogram with unity weights. (c) Reverberant DOA observations $\psi(k,l)$ with max. of 5° localization error (black points). (d) Reverberant azimuth histogram with unity weights. (e) Estimated DOA weights $w(k,l)$. (f) Reverberant azimuth histogram with weights from (e).

III. Experimental Evaluation

A. Setup

Multipath sound propagation was simulated for a small rectangular room of dimensions 6 m × 4 m × 3 m (length × width × height). Wall reflections were estimated using the image model method for simulating small-room acoustics [10]. Room impulse responses for different reverberation times were generated for each sensor of a six-channel ULA with inter-element spacing of $d = 4.28$ cm at a sampling frequency of 8 kHz. The array was positioned in the middle of the room at a height of 2 m. Two speech sources with equal gain were placed at different horizontal angles facing array broadside and a distance of 1.5 m from the array center. A total of 240 different speech mixtures were constructed for testing with utterances from the TIMIT and TIDIGIT databases. Simulations were run for three different DOA configurations with azimuths of $(\theta_1,\theta_2) \in \{(-20°,20°),(-10°,10°),(-5°,5°)\}$ and three room reverberation times $\text{RT}_{60} \in \{0\text{ ms}, 300\text{ ms}, 600\text{ ms}\}$. The STFT frame size was 25 ms with a shift of 10 ms. Following a range of tuning experiments on a cross-validation set, the sigmoid slope parameters were fixed to $\alpha_1 = -3$ and $\alpha_2 = 8$. For each utterance, the sigmoid center β_1 was fixed to be the 98th percentile of the set $\bigcup_{(k,l)}\{o(k,l)\}$. Similarly, the center parameter β_2 was set to be the pth percentile of $\bigcup_{(k,l)}\{\log[\sigma^2_\psi(k,l)]\}$. The value of p was tuned for each configuration. The stronger the reverberation and the smaller the azimuth separation between sources, the lower the percentile was chosen. The localization performance of the fuzzy c-means algorithm was then measured with and without the proposed observation weighting.

B. Results and Discussion

As expected, for anechoic conditions (Fig. 3, left column) the localization performance using the standard fuzzy clustering was sufficient for all azimuth separation angles. The observation weighting had little effect in these cases as almost all T-F observations produced DOA values close to the true azimuth angles. However, for 300-ms and 600-ms reverberation times (Fig. 3, middle and right column) the standard clustering failed to locate the two sources correctly. Too many observations have become unreliable and have consequently started to bias the clustering towards incorrect solutions. On the other hand, the proposed weighting scheme successfully located both sources for most tested configurations, even for the challenging case of only 10° angular separation and 600-ms reverberation time (Fig. 3, right column, bottom row).

In terms of limitations, our current implementation is based on a linear array which is restricted to azimuth angle estimation and is subject to front-back confusions. However, for full three-dimensional localization, the outlined approach can easily

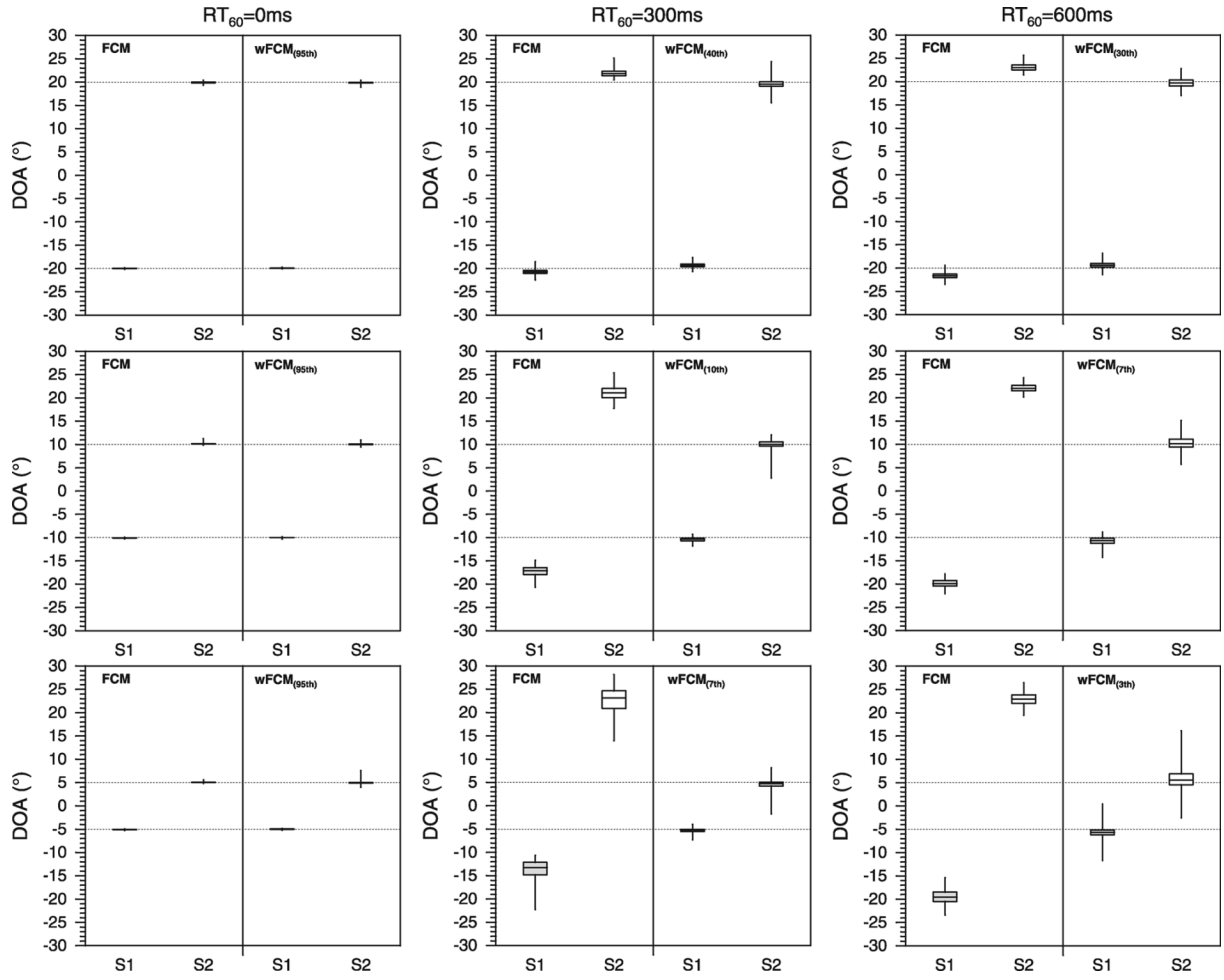

Fig. 3. DOA localization performance for a two-source configuration with different reverberation times (columns) and azimuth separations (rows). Results are presented for the standard FCM method and the proposed $\mathrm{wFCM}_{(\beta_2\text{-percentile})}$ algorithm with observation weighting. The true DOA angles for speaker $S1$ and $S2$ are indicated by horizontal dashed lines.

be extended to nonlinear array geometries [2], [7]. The major drawback of the current approach is the rather *ad-hoc* determination of the weighting factors which requires prior knowledge about the environment. Ideally, the sigmoid parameters should be automatically adapted by the algorithm itself. Further research is needed to address these issues and to extend the method to cases with moving speakers and sources with smooth or no onsets.

IV. Conclusion

We have presented a weighted fuzzy clustering algorithm for tackling multisource localization in reverberant environments. In order to increase the algorithm's robustness, observation weights were incorporated to emphasize reliable over unreliable DOA cues. The weights were derived from local DOA statistics around sound onsets because it is known that these regions are least affected by reverberation. Our experimental evaluation showed that the proposed method produces superior localization results when compared with standard fuzzy clustering, particularly in reverberant conditions. As a consequence, it is expected that the resulting cluster partitions will also lead to better T-F separation masks. In subsequent work, we therefore intend to investigate reverberant BSS problems by integrating our robust source localization scheme into T-F masking and spatial filtering techniques.

References

[1] Ö Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[3] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[4] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.

[5] R. Litovsky, H. Colburn, W. Yost, and S. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1633–1654, 1999.

[6] C. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[7] M. Togami, T. Sumiyoshi, and A. Amano, "Stepwise phase difference restoration method for sound source localization using multiple microphone pairs," in *Proc. IEEE ICASSP*, Honolulu, HI, 2007.

[8] S. Miyamoto, R. Inokuchi, and Y. Kuroda, "Possibilistic and fuzzy c-means clustering with weighted objects," in *IEEE Fuzzy Syst. Conf.*, Vancouver, BC, Canada, 2006.

[9] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Process.*, vol. 85, pp. 1389–1403, 2005.

[10] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.