ON THE USE OF AUTOREGRESSIVE MODELING FOR LOCALIZATION OF SPEECH

Jacek Dmochowski, Jacob Benesty, and Sofiène Affes

Université du Québec, INRS-EMT
800 rue de la Gauchetière Ouest, Montréal, Québec, Canada, H5A 1K6
{dmochow, benesty, affes}@emt.inrs.ca

ABSTRACT

The localization of speech is essential for improving the quality of hands-free pick-up as well as for applications such as automatic camera steering. This paper proposes a source localization method tailored to the distinct nature of speech that is based on the linearly constrained minimum variance (LCMV) beamforming method. The LCMV steered beam *temporally* focuses the array onto the desired signal. By modeling the desired signal as an autoregressive (AR) process and embedding the AR coefficients in the linear constraints, the localization accuracy is significantly improved as compared to existing techniques.

1. INTRODUCTION

The localization of one or more speech sources is an important problem both as a prior to spatial filtering, and in applications such as automated video-camera steering. Source localization may be viewed as the spatial analogue of spectral estimation [1]: spatial spectral estimation. The multidimensional Fourier transform of a space-time signal is known as the *wavenumber-frequency* transform [2]. The resulting transform is a function of both spatial and temporal frequencies. As a result, the temporal nature of the involved signals certainly impacts the source localization process.

Source localization algorithms may roughly be divided into two categories. The first category involves a two-step approach comprised of time delay estimation in the first step, followed by a mapping of the relative delays to the source location in the second step [3]. The second category is based on parameterized spatial correlation [4], and includes the well known steered-response power (SRP) algorithm [5], [6].

Human speech has many distinct features: for example, voiced segments are quasi-periodic with dominant formant regions. In general, speech is quite colored, and thus may be represented quite accurately by an autoregressive (AR) process. It is surprising that the majority of acoustic source localization algorithms proposed to date have not exploited the predictable nature of speech.

This paper presents a source localization method based on the well-known linearly constrained minimum variance (LCMV) technique proposed by Frost [7]. Instead of viewing the LCMV filter as a beamformer, however, we present the structure as a spatial spectral estimator which takes into account the temporal characteristics of the desired signal via the LCMV constraints.

2. SIGNAL MODEL

Assume that an array of N microphones samples the sound field in an anechoic environment. The output of microphone n at time sample k is then modeled as:

$$x_n(k) = \alpha_n(\mathbf{r}_\mathrm{s}) s\left[k - \tau - \mathcal{F}_{1n}(\mathbf{r}_\mathrm{s})\right] + v_n(k), \qquad (1)$$

where $\alpha_n(\mathbf{r}_\mathrm{s}), n = 1, 2, \ldots, N$, models the attenuation of the source signal at microphone n as a function of the source location $\mathbf{r}_\mathrm{s} = (r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})$, where r_s, ϕ_s, and θ_s denote the range, elevation, and azimuth, respectively, s is the source signal, τ is the propagation time (in samples) from the source to sensor 1, $\mathcal{F}_{ij}(\mathbf{r}_\mathrm{s})$ is a function that relates the source position to the relative delay between microphones i and j, and v_n is the additive noise at microphone n. In the far-field case, it is appropriate to assume $\alpha_n(\mathbf{r}_\mathrm{s}) = 1, \forall n, \mathbf{r}_\mathrm{s}$.

The function $\mathcal{F}_{ij}$ is related to the distances between the source and the sensors i and j:

$$\mathcal{F}_{ij}(\mathbf{r}_\mathrm{s}) = \frac{d_{j,\mathrm{s}}(\mathbf{r}_\mathrm{s}) - d_{i,\mathrm{s}}(\mathbf{r}_\mathrm{s})}{c}, \qquad (2)$$

where c is the speed of propagation. When the distance from the source to the array is large in comparison to the extent of the spatial aperture, the source is said to be in the "far-field" and the incoming wave front may be assumed to be planar. In that case, $\mathcal{F}_{ij}$ becomes independent of the source range:

$$\mathcal{F}_{ij}(\phi_\mathrm{s}, \theta_\mathrm{s}) = \frac{\boldsymbol{\zeta}^T(\phi_\mathrm{s}, \theta_\mathrm{s})\left(\mathbf{z}_j - \mathbf{z}_i\right)}{c}, \qquad (3)$$

where

$$\boldsymbol{\zeta}(\phi_\mathrm{s}, \theta_\mathrm{s}) = \begin{bmatrix} \sin\phi_\mathrm{s}\cos\theta_\mathrm{s} & \sin\phi_\mathrm{s}\sin\theta_\mathrm{s} & \cos\phi_\mathrm{s} \end{bmatrix}^T \qquad (4)$$

is a unit vector which points in the direction of propagation of the source, and $\mathbf{z}_i = \begin{bmatrix} z_{i,x} & z_{i,y} & z_{i,z} \end{bmatrix}^T$ is the position vector of the ith sensor. Furthermore, if the source lies in the same plane as the array of sensors, $\boldsymbol{\zeta}$ becomes independent of the elevation angle, and as a result, $\mathcal{F}_{ij}$ loses its dependence on ϕ_s. In that case $\mathbf{z}_i = \begin{bmatrix} z_{i,x} & z_{i,y} \end{bmatrix}^T$ and $\boldsymbol{\zeta}$ effectively become two-dimensional, with

$$\boldsymbol{\zeta}(\phi_\mathrm{s}, \theta_\mathrm{s})\,|_{2-\mathrm{D}} = \begin{bmatrix} \cos\theta_\mathrm{s} & \sin\theta_\mathrm{s} \end{bmatrix}^T. \qquad (5)$$

3. LCMV SPATIAL SPECTRAL ESTIMATION

Beamforming and source localization are very inter-related. For example, notice that the conventional delay-and-sum beamformer (DSB) may be viewed as a spatial filter when applying a single set of equalizing delays to the microphones and then summing, but also as a source localizer: by steering the DSB to all candidate locations and determining the location which radiates the most energy, the source may be localized. Since the DSB only processes

Thanks to NSERC for supporting this work.

one temporal sample at a time, it is unable to perform any temporal signal discrimination. Though the LCMV filter is celebrated in the context of signal enhancement, it has never been viewed as a spectral estimation tool. In this section, it is shown that by using the LCMV framework, the temporal properties of the desired signal may be exploited to generate enhanced estimates of the spatial properties of the signal.

The proposed technique performs both spatial and temporal discrimination. A delay is first applied to each microphone such that the propagation delays $\mathcal{F}_{1n}(\mathbf{r}_\mathrm{s})$ are equalized. This processing is done for *all* possible source locations, leading to the parameterized output:

$$x_{n,\mathrm{p}}(k,\mathbf{r}) = x_n\left[k + \mathcal{F}_{1n}(\mathbf{r})\right], \tag{6}$$

where $\mathbf{r}$ is the steered location (i.e., the parameter). When the steered location $\mathbf{r}$ matches the actual location $\mathbf{r}_\mathrm{s}$, the desired signal is time-aligned:

$$x_{n,\mathrm{p}}(k,\mathbf{r}_\mathrm{s}) = \alpha_n(\mathbf{r}_\mathrm{s})s\left[k-\tau\right] + v_n\left[k + \mathcal{F}_{1n}\left(\mathbf{r}_\mathrm{s}\right)\right]. \tag{7}$$

In vector notation, the received and time-aligned signals are written as:

$$\mathbf{x}_\mathrm{p}(k,\mathbf{r}) = \mathbf{A}(\mathbf{r}_\mathrm{s})\mathbf{s}_\mathrm{p}\left(k-\tau,\mathbf{r}\right) + \mathbf{v}_\mathrm{p}(k,\mathbf{r}), \tag{8}$$

where

$$\begin{aligned}
\mathbf{x}_\mathrm{p}(k,\mathbf{r}) &= \left[\begin{array}{cccc} x_{1,\mathrm{p}}(k,\mathbf{r}) & \cdots & x_{N,\mathrm{p}}(k,\mathbf{r}) \end{array}\right]^T, \\
\mathbf{A}(\mathbf{r}_\mathrm{s}) &= \mathrm{diag}\left[\alpha_1(\mathbf{r}_\mathrm{s}),\ldots,\alpha_N(\mathbf{r}_\mathrm{s})\right], \\
\mathbf{s}_\mathrm{p}\left(k-\tau,\mathbf{r}\right) &= \left[\begin{array}{cc} s\left[k-\tau\right] & \cdots \\ \cdots & s\left[k-\tau-\mathcal{F}_{1N}\left(\mathbf{r}_\mathrm{s}\right)+\mathcal{F}_{1N}\left(\mathbf{r}\right)\right] \end{array}\right]^T, \\
\mathbf{v}_\mathrm{p}(k,\mathbf{r}) &= \left[\begin{array}{cc} v_1\left(k\right) \cdots & v_N\left[k+\mathcal{F}_{1N}\left(\mathbf{r}\right)\right] \end{array}\right]^T,
\end{aligned}$$

and $\mathrm{diag}(\cdot)$ is a diagonal matrix whose nonzero entries are indicated by the arguments.

To allow for temporal processing, we append the previous L_h-1 samples of each microphone to form a spatiotemporal aperture:

$$\overline{\mathbf{x}}_\mathrm{p}(k,\mathbf{r},L_h) = \overline{\mathbf{A}}(\mathbf{r}_\mathrm{s})\overline{\mathbf{s}}_\mathrm{p}(k-\tau,\mathbf{r},L_h) + \overline{\mathbf{v}}_\mathrm{p}(k,\mathbf{r},L_h), \tag{9}$$

where

$$\begin{aligned}
\overline{\mathbf{x}}_\mathrm{p}(k,\mathbf{r},L_h) &= \left[\begin{array}{ccc} \mathbf{x}_\mathrm{p}^T(k,\mathbf{r}) & \cdots & \mathbf{x}_\mathrm{p}^T(k-L_h+1,\mathbf{r}) \end{array}\right]^T, \\
\overline{\mathbf{A}}(\mathbf{r}_\mathrm{s}) &= \left[\begin{array}{cccc} \mathbf{A}(\mathbf{r}_\mathrm{s}) & \mathbf{0}_{N\times N} & \cdots & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N} & \mathbf{A}(\mathbf{r}_\mathrm{s}) & \cdots & \mathbf{0}_{N\times N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{N\times N} & \mathbf{0}_{N\times N} & \cdots & \mathbf{A}(\mathbf{r}_\mathrm{s}) \end{array}\right], \\
\overline{\mathbf{s}}_\mathrm{p}(k-\tau,\mathbf{r},L_h) &= \left[\begin{array}{cc} \mathbf{s}_\mathrm{p}^T\left(k-\tau,\mathbf{r}\right) & \cdots \\ \cdots & \mathbf{s}_\mathrm{p}^T\left(k-\tau-L_h+1,\mathbf{r}\right) \end{array}\right]^T, \\
\overline{\mathbf{v}}_\mathrm{p}(k,\mathbf{r},L_h) &= \left[\begin{array}{ccc} \mathbf{v}_\mathrm{p}^T(k,\mathbf{r}) & \cdots & \mathbf{v}_\mathrm{p}^T(k-L_h+1,\mathbf{r}) \end{array}\right]^T,
\end{aligned}$$

where $\mathbf{0}_{N\times N}$ is an N-by-N matrix of zeros and $\overline{\mathbf{A}}(\mathbf{r}_\mathrm{s})$ has size NL_h-by-NL_h. We apply a multichannel finite impulse response (FIR) filter to the spatiotemporal aperture to yield an array output that is temporally constrained:

$$\mathbf{h}^T\left(\mathbf{r}\right)\overline{\mathbf{A}}(\mathbf{r}_\mathrm{s})\overline{\mathbf{s}}_\mathrm{p}\left(k-\tau,\mathbf{r},L_h\right) = \sum_{l=0}^{L_h-1} f_l s\left(k-\tau-l\right), \tag{10}$$

where

$$\mathbf{h}\left(\mathbf{r}\right) = \left[\begin{array}{cccc} \mathbf{h}_{:0}^T\left(\mathbf{r}\right) & \mathbf{h}_{:1}^T\left(\mathbf{r}\right) & \cdots & \mathbf{h}_{:L_h-1}^T\left(\mathbf{r}\right) \end{array}\right]^T \tag{11}$$

denotes the multichannel filter with $\mathbf{h}_{:i}\left(\mathbf{r}\right) = \left[\begin{array}{cccc} h_{1i}\left(\mathbf{r}\right) & h_{2,i}\left(\mathbf{r}\right) & \cdots & h_{Ni}\left(\mathbf{r}\right) \end{array}\right]^T$ denoting the subfilter applied to the set of samples at time sample $k-\tau-i$. The spatiotemporal filter coherently sums a signal propagating from location $\mathbf{r}$ followed by a temporal filtering which is specified by the coefficients $f_l, l=0,1,\ldots,L_h-1$.

Assuming a source propagating from the steered location $\mathbf{r}$, the constraints follow from (10) as:

$$\mathbf{c}_{\alpha,l}^T(\mathbf{r})\mathbf{h}\left(\mathbf{r}\right) = f_l, l=0,1,\ldots,L_h-1, \tag{12}$$

where

$$\mathbf{c}_{\boldsymbol{\alpha},l}(\mathbf{r}) = \left[\begin{array}{ccccc} \mathbf{0}_{N\times 1}^T & \cdots & \underbrace{\boldsymbol{\alpha}^T(\mathbf{r})}_{l\text{th group}} & \cdots & \mathbf{0}_{N\times 1}^T \end{array}\right]^T$$

is the lth constraint vector of length NL_h, and

$$\boldsymbol{\alpha}(\mathbf{r}) = \left[\begin{array}{cccc} \alpha_1(\mathbf{r}) & \alpha_2(\mathbf{r}) & \ldots & \alpha_N(\mathbf{r}) \end{array}\right]^T.$$

The L_h constraints of (12) may be neatly expressed in matrix notation as:

$$\mathbf{C}_{\boldsymbol{\alpha}}^T(\mathbf{r})\mathbf{h}\left(\mathbf{r}\right) = \mathbf{f}, \tag{13}$$

where

$$\begin{aligned}
\mathbf{C}_{\boldsymbol{\alpha}}(\mathbf{r}) &= \left[\begin{array}{cccc} \mathbf{c}_{\boldsymbol{\alpha},0}(\mathbf{r}) & \mathbf{c}_{\boldsymbol{\alpha},1}(\mathbf{r}) & \cdots & \mathbf{c}_{\boldsymbol{\alpha},L_h-1}(\mathbf{r}) \end{array}\right], \\
\mathbf{f} &= \left[\begin{array}{cccc} f_0 & f_1 & \cdots & f_{L_h-1} \end{array}\right]^T.
\end{aligned}$$

After forming the desired constraints, the remaining degrees of freedom are utilized to minimize the average output power:

$$E\left\{y^2\left(k,\mathbf{r}\right)\right\} = E\left\{\left[\mathbf{h}^T(\mathbf{r})\overline{\mathbf{x}}_\mathrm{p}(k,\mathbf{r},L_h)\right]^2\right\}, \tag{14}$$

where

$$y\left(k,\mathbf{r}\right) = \mathbf{h}^T(\mathbf{r})\overline{\mathbf{x}}_\mathrm{p}(k,\mathbf{r},L_h),$$

which corresponds to minimizing the contribution of noise and interference to the spectral estimate. The minimization problem is thus, for every steered location $\mathbf{r}$:

$$\hat{\mathbf{h}}\left(\mathbf{r}\right) = \arg\min_{\mathbf{h}} \mathbf{h}^T\mathbf{R}_{\overline{\mathbf{x}}_\mathrm{p}\overline{\mathbf{x}}_\mathrm{p}}(k,\mathbf{r},L_h)\mathbf{h} \quad \text{subject to} \quad \mathbf{C}_\alpha^T\left(\mathbf{r}\right)\mathbf{h} = \mathbf{f}, \tag{15}$$

where

$$\mathbf{R}_{\overline{\mathbf{x}}_\mathrm{p}\overline{\mathbf{x}}_\mathrm{p}}(k,\mathbf{r},L_h) = E\left\{\overline{\mathbf{x}}_\mathrm{p}(k,\mathbf{r},L_h)\overline{\mathbf{x}}_\mathrm{p}^T(k,\mathbf{r},L_h)\right\}.$$

The solution to the constrained optimization problem is well-known; using the method of Lagrange multipliers,

$$\hat{\mathbf{h}}\left(\mathbf{r}\right) = \mathbf{R}_{\overline{\mathbf{x}}_\mathrm{p}\overline{\mathbf{x}}_\mathrm{p}}^{-1}(k,\mathbf{r},L_h)\mathbf{C}_\alpha\left(\mathbf{r}\right) \times \left[\mathbf{C}_\alpha^T\left(\mathbf{r}\right)\mathbf{R}_{\overline{\mathbf{x}}_\mathrm{p}\overline{\mathbf{x}}_\mathrm{p}}^{-1}(k,\mathbf{r},L_h)\mathbf{C}_\alpha\left(\mathbf{r}\right)\right]^{-1}\mathbf{f}. \tag{16}$$

The source location estimate follows as:

$$\begin{aligned}
\hat{\mathbf{r}}_\mathrm{s} &= \arg\max_{\mathbf{r}} S_{\mathrm{LCMV}}(\mathbf{r}) \\
&= \arg\max_{\mathbf{r}} \hat{\mathbf{h}}^T\left(\mathbf{r}\right)\mathbf{R}_{\overline{\mathbf{x}}_\mathrm{p}\overline{\mathbf{x}}_\mathrm{p}}(k,\mathbf{r},L_h)\hat{\mathbf{h}}\left(\mathbf{r}\right),
\end{aligned} \tag{17}$$

where $S_{\mathrm{LCMV}}(\mathbf{r}) = \hat{\mathbf{h}}^T\left(\mathbf{r}\right)\mathbf{R}_{\overline{\mathbf{x}}_\mathrm{p}\overline{\mathbf{x}}_\mathrm{p}}(k,\mathbf{r},L_h)\hat{\mathbf{h}}\left(\mathbf{r}\right)$ is the estimate of the spatial spectrum at the spatial frequency corresponding to location $\mathbf{r}$.

4. AUTOREGRESSIVE MODELING

The minimum variance distortionless response (MVDR) technique, proposed for narrowband signals by Capon [8], and later for the broadband case by Krolik and Swingler [9], is a particular case of the LCMV technique which employs $L_h = 1$ and $\mathbf{f} = f_0 = 1$. The lone MVDR constraint simply passes $s(k-\tau)$ through to the output with unity gain. Since the method uses $L_h = 1$, it is not able to provide any temporal signal discrimination. Moreover, due to the short aperture length, the noise reducing minimization procedure is limited in the degrees of freedom.

The proposed method alleviates the limitations noted above. In LCMV spectral estimation, the idea is to estimate the present sample as a linear combination of the past samples. This naturally calls for the modeling of the desired signal as an autoregressive (AR) process:

$$s(k) = \sum_{l=1}^{p} a_l s\left(k-l\right) + w(k), \tag{18}$$

where a_l are the predictive coefficients, p is the order of the AR model, and $w(k)$ is the prediction error.

In order to determine the constraint vector $\mathbf{f}$, consider (10) – the goal of the constraint is to estimate $s(k-\tau)$ using a linear combination of $\{s(k-\tau), s(k-\tau-1), \ldots, s(k-\tau-L_h+1)\}$:

$$\begin{aligned} \hat{s}(k-\tau) &= \mathbf{h}^T\left(\mathbf{r}\right)\overline{\mathbf{A}}(\mathbf{r}_{\mathrm{s}})\bar{\mathbf{s}}_{\mathrm{p}}\left(k-\tau, \mathbf{r}, L_h\right) \\ &= \sum_{l=0}^{L_h-1} f_l s\left(k-\tau-l\right), \end{aligned} \tag{19}$$

where $\hat{s}(k-\tau)$ denotes the estimate of the desired present sample. The MVDR method chooses $f_0 = 1, f_l = 0,\ l = 1, \ldots, L_h - 1$, yielding an errorless estimate but also meaning that temporal dependence is neglected. In the proposed method, the desired signal's temporal properties are taken into account via AR modeling. The AR parameters of the desired signal are embedded in the constraint vector $\mathbf{f}$ which in turn shapes the multichannel filter $\mathbf{h}(\mathbf{r})$. Connecting (18) to (19), the LCMV method chooses

$$f_0 = 0, \tag{20}$$

$$f_l = a_l,\ l = 2, \ldots, p, \tag{21}$$

resulting in an estimation error given by $s(k-\tau) - \hat{s}(k-\tau) = w(k)$.

A zero-mean estimation error is incurred by modeling the signal as an AR process. However, it is expected that the additional degrees of freedom in the multichannel filter $\mathbf{h}\left(\mathbf{r}\right)$ will lead to a greater level of noise reduction. Note that the filter $\mathbf{h}(\mathbf{r})$ *temporally* focuses the steered beam to pick up a signal with the temporal structure contained in $\mathbf{f}$. Any noise or interfering signal with a different temporal structure should be attenuated by this temporally focused filter. In practice, the AR parameters need to be estimated from the microphone signals using either a classical single-channel method such as solving of the Yule-Walker equations [10], or a multichannel method that somehow takes into account the data from all microphones [11].

5. SIMULATION EVALUATION

The proposed localization technique is evaluated in a computer simulation using the image method model of [12]. A 6-microphone uniform circular array with a 4.25 cm radius is simulated. The simulated room is rectangular with plane reflective boundaries and frequency-independent reflection coefficients. The room dimensions in centimeters are $(304.8, 457.2, 381)$. The center of the array sits at $(152.4, 228.6, 101.6)$. The speaker is located at $(152.4, 406.4, 101.6)$. The reverberation times are measured using the method of [13] and range from $T_{60} = 300$ ms to $T_{60} = 900$ ms, where T_{60} is the time for the impulse response's energy to decay by 60 dB. After convolving the source signal with the synthetic impulse responses, appropriately-scaled white Gaussian noise is added at the microphones to achieve the required SNR. SNRs of 10, 20, and 30 dB are simulated. The source signal is female English speech. The sampling rate is 48 kHz. Due to the planar array geometry and far-field source, the location space is one-dimensional and comprised of the set of azimuth angles in the range $0 - 359$ degrees, with a resolution of 1 degree. The azimuth angle estimates are computed once per 64 ms frame over a one-minute signal. The algorithms are evaluated in terms of the percentage of anomalous estimates – those that vary from the true azimuth by more than 5 degrees, and by the root-mean-square (rms) error for the nonanomalous estimates:

$$\mathrm{e}_{\mathrm{rms}} = \sqrt{\frac{1}{L_{\mathrm{na}}} \sum_{l \in \chi_{\mathrm{na}}} \left(\hat{\theta}_l - \theta_l\right)^2}, \tag{22}$$

where χ_{na} is the set of all nonanomalous estimates, L_{na} is the number of elements in χ_{na}, and $\hat{\theta}_l$ and θ_l are the estimated and actual azimuth angles of the source for frame l.

For comparison, the proposed estimators are compared to the SRP [5], [6] and MVDR [9] methods. The generalized cross-correlation (GCC) phase transform (PHAT) method [14] is employed to whiten the observed cross-correlations for all three methods. To estimate the AR coefficients of the desired signal, the Yule-Walker or "autocorrelation" method is employed using data collected from the first microphone. The parameterized spatiotemporal correlation matrix $\mathbf{R}_{\bar{\mathbf{x}}_{\mathrm{p}}\bar{\mathbf{x}}_{\mathrm{p}}}(k, \mathbf{r}, L_h)$ has size NL_h-by-NL_h and requires inversion – numerical stability problems may occur in practice. As a result, prior to inversion, the matrix is regularized using the Tikhonov method [15]:

$$\mathbf{R}^{-1}_{\bar{\mathbf{x}}_{\mathrm{p}}\bar{\mathbf{x}}_{\mathrm{p}}}(k, \mathbf{r}, L_h) \leftarrow \left[\mathbf{R}_{\bar{\mathbf{x}}_{\mathrm{p}}\bar{\mathbf{x}}_{\mathrm{p}}}(k, \mathbf{r}, L_h) + \delta \mathbf{I}_{NLh \times NL_h}\right]^{-1}, \tag{23}$$

where $\leftarrow$ denotes assignment, $\mathbf{I}_{NLh \times NL_h}$ is the NL_h-by-NL_h identity matrix, and δ is the regularization parameter, which is taken in the simulations as:

$$\delta = \frac{1}{NL_h} \mathrm{trace}\left[\mathbf{R}_{\bar{\mathbf{x}}_{\mathrm{p}}\bar{\mathbf{x}}_{\mathrm{p}}}(k, \mathbf{r}, L_h)\right] \Delta, \tag{24}$$

where Δ is the normalized regularization constant, with $\Delta = 0.1$ used in the simulations.

A concluding experiment with real impulse responses measured in the Bell Labs varechoic chamber [16] is also performed. The experiment utilizes a 6-element uniform linear array with an inter-microphone spacing of 10 cm. The reverberation time employed in the evaluation is 280 ms. SNRs of 0, 10, and 20 dB are simulated. The array stands near a wall, with the source located 4.43 m from the array at broadside. All other parameters remain the same as in the image method based simulations.

Table 1 displays the image method simulation results. It is evident that the proposed method provides increased robustness against reverberation and especially noise, with the reduction in anomalies reaching 23 % in the SNR $= 10$ dB, $T_{60} = 900$ ms case. As the SNR is increased, the performance benefit of

Table 1. Source localization performance of conventional and proposed estimators with synthetic impulse responses.

		SRP		MVDR		LCMV	
SNR (dB)	T_{60} (ms)	$\%_{\text{anomalies}}$ (%)	rms (degrees)	$\%_{\text{anomalies}}$ (%)	rms (degrees)	$\%_{\text{anomalies}}$ (%)	rms (degrees)
10	300	43.12	1.82	42.69	1.82	19.96*	0.85*
	600	48.45	1.78	48.67	1.78	31.91*	1.01*
	900	53.36	1.81	53.68	1.80	40.88*	1.10*
20	300	18.78	1.48	18.57	1.49	11.21*	0.85*
	600	27.32	1.45	27.21	1.44	19.74*	1.02*
	900	37.46	1.39	37.46	1.38	29.24*	1.10*
30	300	10.14	1.03	9.82	1.03	10.46*	0.81*
	600	21.34	0.99	21.34	0.99	21.99	0.97*
	900	30.31	1.03	30.10	1.04	30.42*	1.10*

*regularized with $\Delta = 0.1$

Table 2. Source localization performance of conventional and proposed estimators with real impulse responses.

	SRP		MVDR		LCMV	
SNR (dB)	$\%_{\text{anomalies}}$ (%)	rms (degrees)	$\%_{\text{anomalies}}$ (%)	rms (degrees)	$\%_{\text{anomalies}}$ (%)	rms (degrees)
0	18.59	1.67	19.23	1.68	13.46*	1.33*
10	12.18	1.33	12.82	1.31	5.77*	1.07*
20	10.90	1.16	10.26	1.13	7.69*	1.13*

*regularized with $\Delta = 0.1$

the LCMV scheme is somewhat reduced; at $\text{SNR} = 30$ dB, the LCMV method yields a performance comparable to the conventional methods. The results of the evaluation using real impulse responses is shown in Table 2. The LCMV technique leads to superior performance for all values of the SNR. Note that while the proposed method offers increased robustness to adverse conditions, it is also more computationally complex.

6. CONCLUSIONS

This paper has presented a novel source localization technique based on the LCMV beamformer proposed by Frost. It was shown that by accounting for the temporal properties of the desired signal in the linear constraints via AR modeling, source localization performance is significantly improved. The presented algorithm provides one way of accounting for the nature of speech in localization applications.

7. REFERENCES

[1] S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Upper Saddle River, NJ: Prentice-Hall, 1999.

[2] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Upper Saddle River, NJ: Prentice-Hall, 1993.

[3] Y. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, (eds.), Springer-Verlag, Berlin, Germany, 2007.

[4] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 1327–1339, May 2007.

[5] M. Omologo and P. G. Svaizer, "Use of the cross-power-spectrum phase in acoustic event localization," ITC-IRST Tech. Rep. 9303-13, Mar. 1993.

[6] J. Dibiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, eds., pp. 157–180, Springer-Verlag, Berlin, 2001.

[7] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

[8] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[9] J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1481–1494, Oct. 1989.

[10] S. L. Marple Jr., *Digital Spectral Analysis with Applications*, Englewood Cliffs, New Jersey: Prentice Hall, 1987.

[11] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. America*, vol. 120, pp. 4031-4039, Dec. 2006.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[13] M. R. Schroeder, "New method for measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

[14] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.

[15] A. N. Tikhonov, "On the stability of inverse problems," *Dokl. Akad. Nauk SSSR*, vol. 39, pp. 195–198, 1943.

[16] A. Härmä, "Acoustic measurement data from the varechoic chamber," Tech. Memo., Agere Systems, Nov. 2001.