Frequency Map Selection Using a RBFN-based Classifier in the MVDR Beamformer for Speaker Localization in Reverberant Rooms

Daniele Salvati, Carlo Drioli, Gian Luca Foresti

Department of Mathematics and Computer Science, University of Udine
daniele.salvati@uniud.it, carlo.drioli@uniud.it, gianluca.foresti@uniud.it

10.21437/Interspeech.2015-664

Abstract

We present the weighted minimum variance distortionless response (WMVDR), which is a steered response power (SRP) algorithm, for near-field speaker localization in a reverberant environment. The proposed WMVDR is based on a machine learning approach for computing the incoherent frequency fusion of narrowband power maps. We adopt a radial basis function network (RBFN) classifier for the estimation of the weighting coefficients, and a marginal distribution of narrowband power map as feature for the supervised training operation. Simulations demonstrate the effectiveness of the proposed approach in different conditions.

Index Terms: speaker spatial localization, near-field reverberant environment, broadband MVDR, machine learning, RBFN.

1. Introduction

Speaker spatial localization using microphone arrays is of considerable interest in applications of teleconferencing systems, hands-free speaker acquisition, human-machine interaction, sound recognition, and audio surveillance, both in indoor and outdoor space [1, 2, 3, 4, 5]. The steered response power (SRP) algorithms are a class of direct methods used to estimate the sound source position in space. The SRP is based on maximizing the power output of a beamformer. Typically, broadband SRP is computed in the frequency-domain by applying a discrete Fourier transform on a portion of the signal and by calculating the response power on each frequency bin. Subsequently, a fusion of these estimates is computed and the estimation of the speaker position is obtained by searching the maximum on the response power map of a target search area. The fusion of narrowband SRP can be obtained by incoherent [6, 7, 8] or coherent [9, 10, 11] averaging with respect to frequency.

In this paper, we present an incoherent frequency fusion method based on a machine learning approach. We consider the broadband minimum variance distortionless response (MVDR) beamformer for speaker localization in a near-field reverberant environment. The MVDR is a SRP algorithm based on the narrowband adaptive Capon beamformer [12]. Due to the nonstationarity property of speech signals, incoherent averaging effectiveness decreases when the signal-to-noise ratio (SNR) varies at each frequency bin, since the acoustic power map estimate at some frequencies may be affected by large errors, and the final frequency data combination may be inaccurate. In [8], it is shown that a post-filter normalization of each frequency response map substantially improves the spatial resolution of the MVDR beamformer, which is more robust against noise if compared to other algorithms. Unfortunately, the normalization has the disadvantage of emphasizing the noise in those frequencies in which the SNR is low, due to the quasi-periodic of the speech signal in certain voiced fragments. To mitigate this problem, we propose a weighted MVDR (WMVDR) beamformer, which is based on a radial basis function network (RBFN) [13] classifier for selecting only the frequency response maps that give a correct contribution to the final fusion of narrowband beamforming. By using the marginal distribution of acoustic maps as input vector, the RBFN is trained to classify the frequency maps in two classes (positively contributing maps vs maps providing a wrong contribution). If compared to other supervised learning approaches in the literature, in which classifiers are used to directly map the acoustic cues onto a position in the search space [14, 15, 16], in the proposed scheme the machine learning component can be paired to SRP methods, thus providing an incremental contribution to the localization performance. Simulations are shown to verify the effectiveness of the proposed machine learning approach in near-field reverberant rooms.

2. Near-field localization using MVDR beamformer

We consider a speech source that is active at time k in a reverberant room $G = G_x \times G_y \times G_z$, and we assume the source to be in the near-field. We can write the unknown coordinate position of the source as

$$\mathbf{r}_s(k) = [x_s \quad y_s \quad z_s]^T \tag{1}$$

and the positions of M microphones as

$$\mathbf{r}_m = [x_m \quad y_m \quad z_m]^T \quad m = 1, 2, \ldots, M. \tag{2}$$

Consider a time-domain signal block of L samples at time k (time index is omitted in the frequency-domain for simplicity), the reverberant model in frequency-domain can be expressed as

$$X_m(f) = H_m(f)S(f) + V_m(f) \tag{3}$$

where $m = 1, 2, \ldots, M$, f is the frequency bin index, $S(f)$ is the speech signal, $V_m(f)$ is the uncorrelated noise signal, and $H_m(f)$ is the acoustic transfer function from the speaker to the microphone m.

The MVDR beamformer [12] is one of the well-known adaptive beamforming techniques. Beamforming can be seen as a filtered combination of the delayed signals, and the frequency-domain output in matrix notation for frequency f can be written as

$$Y(f) = \mathbf{W}^H(f)\mathbf{X}(f) \tag{4}$$

where $\mathbf{X} = [X_1(f)\ X_2(f) \ldots X_M(f)]^T$, $\mathbf{W}(f) = [W_1(f)\ W_2(f) \ldots W_M(f)]^T$ is the beamformer weights for steering and filtering the data, and the superscript H represents

the Hermitian (complex conjugate) transpose. The power spectral density of the beamformer output is given by

$$P(f) = E[|Y(f)|^2] = \mathbf{W}^H(f)\mathbf{\Phi}(f)\mathbf{W}(f) \quad (5)$$

where $\mathbf{\Phi}(f) = E[\mathbf{X}(f)\mathbf{X}^H(f)]$ is the cross-spectral density matrix and $E[\cdot]$ denotes mathematical expectation.

Consider a generic space position $\mathbf{r}_g = [x_g \quad y_g \quad z_g]^T$ in the target area, the MVDR filter relies on the solution of the minimization problem

$$\underset{\mathbf{W}(f)}{\operatorname{argmin}} \mathbf{W}^H(f)\mathbf{\Phi}(f)\mathbf{W}(f) \text{ s.t. } \mathbf{W}^H(f)\mathbf{A}(f,\mathbf{r}_g) = 1 \quad (6)$$

where $\mathbf{A}(f,\mathbf{r}_g)$ is the steering vector corresponding to a given position $\mathbf{r}_g$, and it depends on the time difference of arrival (TDOA) of the spherical wavefront between microphones. We can write the TDOA between microphone i and j as

$$\tau_{i,j} = \frac{||\mathbf{r}_i - \mathbf{r}_g|| - ||\mathbf{r}_j - \mathbf{r}_g||}{c} \quad (7)$$

where $||\cdot||$ denotes Euclidean norm and c is the speed of sound. In the near-field, the steering vector takes the form

$$\mathbf{A}(f,\mathbf{r}_g) = [1, e^{\frac{j2\pi(f-1)\tau_{1,2}}{L}}, \ldots, e^{\frac{j2\pi(f-1)\tau_{1,M}}{L}}]^T. \quad (8)$$

The aim of the MVDR filter is to minimize the energy of noise and sources coming from different directions, while keeping a fixed gain on the desired position. Solving (6) using the method of Lagrange multipliers, we obtain

$$\mathbf{W}(f) = \frac{\mathbf{\Phi}^{-1}(f)\mathbf{A}(f,\mathbf{r}_g)}{\mathbf{A}^H(f,\mathbf{r}_g)\mathbf{\Phi}^{-1}(f)\mathbf{A}(f,\mathbf{r}_g)}. \quad (9)$$

In real applications, the inverse of the cross-spectral density matrix can be calculated using the Moore-Penrose pseudoinverse, defined as $\mathbf{\Phi}^+ = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^H$, where $\mathbf{\Phi} = \mathbf{U}\mathbf{S}\mathbf{V}^H$ is the singular value decomposition of the matrix $\mathbf{\Phi}$. Moreover, if $\mathbf{\Phi}$ is ill-conditioned, the spatial spectrum could be deteriorated by steering vector errors and finite sample effect. Therefore, a diagonal loading (DL) [17] method is adopted to calculate the inverse matrix in a stable way. The power spectrum of the beamformer output with MVDR filter and DL becomes

$$P(f,\mathbf{r}_g) = \frac{1}{\mathbf{A}^H(f,\mathbf{r}_g)(\mathbf{\Phi}(f) + \mu\mathbf{I})^+\mathbf{A}(f,\mathbf{r}_g)} \quad (10)$$

where $\mathbf{I}$ is the identity matrix and the loading level is $\mu = \frac{1}{L}\text{trace}[\mathbf{\Phi}(f)]\Delta$, where Δ is the normalized loading constant and $\text{trace}[\cdot]$ denotes the sum of the elements on the main diagonal of the cross-spectral density matrix.

The broadband MVDR using an incoherent arithmetic mean is given by

$$P_{\text{MVDR}}(\mathbf{r}_g) = \int_f P(f,\mathbf{r}_g)df \quad (11)$$

and the normalized MVDR (NMVDR) [8] can be written as

$$P_{\text{NMVDR}}(\mathbf{r}_g) = \int_f \frac{P(f,\mathbf{r}_g)}{\max_{\mathbf{r}'_g}[\mathbf{P}_{\mathbf{r}'_g}(f)]}df \quad (12)$$

where $\mathbf{P}_{\mathbf{r}'_g}(f) = [P(f,\mathbf{r}'_1), P(f,\mathbf{r}'_2), \ldots, P(f,\mathbf{r}'_g), \ldots]$ is the frequency power map for all the desired positions $\mathbf{r}'_g \in G$ and $\max[\cdot]$ denotes the maximum value. The normalization has the beneficial effect of increasing the spatial resolution [8], and thus it allows a better identification of direct path and reflections. Finally, the speaker spatial localization is estimated by picking the maximum value on the fusion map

$$\widehat{\mathbf{r}}_s(k) = \underset{\mathbf{r}'_g}{\operatorname{argmax}}[\mathbf{P}_{\mathbf{r}'_g}] \quad (13)$$

where $\mathbf{P}_{\mathbf{r}'_g} = [P(\mathbf{r}'_1), P(\mathbf{r}'_2), \ldots, P(\mathbf{r}'_g), \ldots]$ is the acoustic power map for all the desired positions.

3. Acoustic map selection using RBFN

Both MVDR and NMVDR have the disadvantage that, in noisy or reverberant conditions, some of the frequency maps in the fusion may exhibit an energy peak corresponding to a wrong position in the search space, thus providing a wrong contribution to the fusion map. To avoid to use this wrong information, we introduce the following weighted MVDR (WMVDR):

$$P_{\text{WMVDR}}(\mathbf{r}_g) = \int_f \gamma_f \frac{P(f,\mathbf{r}_g)}{\max_{\mathbf{r}'_g}[\mathbf{P}_{\mathbf{r}'_g}(f)]}df \quad (14)$$

where γ_f are weighting factors which attenuate those components that do not contribute positively to the correct localization of the acoustic source. Given a reference source position $\mathbf{r}_s$, and being

$$\widehat{\mathbf{r}}_s(k,f) = \underset{\mathbf{r}'_g}{\operatorname{argmax}}[\mathbf{P}_{\mathbf{r}'_g}(f)] \quad (15)$$

the estimate of the source position computed using only the information related to frequency f, the contribution at f to the localization error is defined as

$$E(f,\mathbf{r}_s) = ||\mathbf{r}_s - \widehat{\mathbf{r}}_s(f)||. \quad (16)$$

The weighting factors γ_f are modeled by a RBFN classifier, which is trained to attenuate those components which contribute negatively to the localization. Namely, the training set output of the RBFN is set as

$$\gamma_f = \begin{cases} 0 & \text{if } E(f,\mathbf{r}_s) > Th \\ 1 & \text{if } E(f,\mathbf{r}_s) < Th \end{cases} \quad (17)$$

where Th is a given threshold. The training set input is defined as the marginal distribution of the acoustic maps along x, y, and z axes:

$$I_f(x) = \int_y \int_z P(f,\mathbf{r}_g)dydz, \quad \forall x \in G_x \quad (18)$$

$$I_f(y) = \int_x \int_z P(f,\mathbf{r}_g)dxdz, \quad \forall y \in G_y \quad (19)$$

$$I_f(z) = \int_x \int_y P(f,\mathbf{r}_g)dxdy, \quad \forall z \in G_z \quad (20)$$

The input vector $\mathbf{I}_f = [I_f(x) \ldots f(y) \ldots I_f(z)]^T$ is a combination of the marginal distributions.

The RBFN supervised model, which is responsible to compute the weighting coefficients γ_f from the marginal distribution of acoustic maps, is then defined as

$$\gamma_f = \sum_{i=0}^{Q} w_{i,f}\psi_i(\mathbf{I}_f;\mathbf{d}_i) \quad (21)$$

with $\psi_i(\mathbf{I}_f;\mathbf{d}_i)$ being the Q radial kernels of the expansion, $\mathbf{d}_i$ being the set of parameters of the ith kernel, and $w_{i,f}$ being the

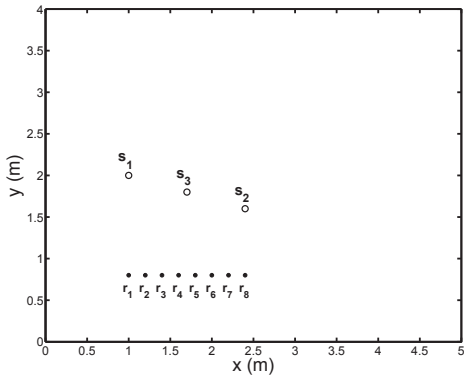

Figure 1: The simulated room setup with the positions of the array and sources.

expansion coefficients. If compared to commonly used Gaussian functions, compactly supported kernels lead to sparse and better conditioned kernel matrices, with computational advantages both during training of the models and during run of the trained models [18]. Thus we decided to use a Wendland kernel of the form $\psi(\mathbf{I}_f; \mathbf{m}, \sigma) = (1-r)_+^8(32r^3 + 25r^2 + 8r + 1)$, with $r = ||\mathbf{I}_f - \mathbf{m}||/\sigma$ being the distance of the input data from the center $\mathbf{m}$ of the kernel, weighted by the width parameter σ. The expansion coefficients $w_{i,f}$ can be computed by choosing among a number of training algorithm available to date. In this paper, we used a greedy OLS training algorithm, that iteratively adds a new kernel to the expansion in the position that maximally contribute to reduce the training data reconstruction error [19].

4. Simulations

The localization performance of the proposed machine learning approach is illustrated through a set of simulated experiments. A uniform linear array of 8 microphones was used. The distance between microphone was 0.2 m. The image-source method (ISM) was used to simulate reverberant audio data in room acoustics [20]. The ISM assumes that source and microphones are omnidirectional. A room of $(5 \times 4 \times 3)$ m was used. The localization in a two-dimensional plane was considered, and therefore both microphones and the source were positioned at a distance from the floor of 1.3 m. The room setup is depicted in Figure 1. We consider three source positions: s_1, s_2, and s_3. The acoustic map is computed on a grid with spatial resolution of 0.1 m. Note that the source were positioned on points of the grid. The reverberant condition has been set to 0.3 s reverberation time (RT_{60}). A 21 s duration adult female speech and a 24 s duration adult male speech were used as source signals. The tests were conducted by setting a SNR of 30 dB, which was obtained by adding mutually independent white Gaussian noise to each channel. The sampling frequency was 44.1 kHz, the block size L was 2048 samples. A frequency range between 80 Hz and 8000 Hz, since it is a typical spectrum range of speech signals, was used for computing the MVDR response power. We compare the performance of MVDR, NMVDR, and WMVDR. The localization performance has been evaluated with several Monte Carlo simulations, using 30 run-trials for each condition test. The RBFN parameters were set to $\sigma = 1$ and $Q = 100$. The threshold parameter was set to $Th = 0.5$ m. These values was determined empirically.

Table 1: RMS (m) error of localization performance.

	MVDR	NMVDR	WMVDR
Test 1	1.42	0.73	0.55
Test 2	1.32	0.68	0.54
Test 3	1.11	0.74	0.63

Table 2: RBFN statistic (%) for three frequency ranges.

	80-400 Hz	400-2000 Hz	2000-8000 Hz
Reject Map	91.1	79.5	86.1
RBFN Error	31.8	36.7	21.9

Three experiments have been conducted: Test 1 - Same speaker and same position: the female speech signal was positioned in s_1 and the RBFN was trained on the first 5 % of the signal. The localization performance was evaluated on the second 95 % of the signal. Test 2 - Different speakers and same position: the female speech signal was positioned in s_1 and the RBFN was trained on the first 5 % of the signal. The localization performance was evaluated on the speech male signal in position s_1. Test 3 - Same speaker and different positions: the female speech signal was positioned in s_1 and the the first 2.5 % of the signal was used to collect the training set vector. Then, the female speech signal was positioned in s_2 and the the first 2.5 % of the signal was used to collect more data training set. The localization performance was evaluated on the female speech signal in position s_3. The performance was evaluated with root mean square (RMS) error for all estimates. The results are shown in Table 1. We can observe the best performance of the WMVDR, and the capability of the RBFN classifier to select frequency maps for a different speaker and for a different position. Moreover, the results prove that the normalized post-filter is effective in a reverberant environment, since only free-field noise condition is considered in [8]. Figure 2 shows the acoustic maps at a specific analysis block for the female speech in position s_1 (Test 1). Table 2 shows some statistics of the RBFN considering three frequency ranges. The percentage of rejection was evaluated using equation (17), and the percentage of RBFN error was evaluated testing the training set. The error was defined by an absolute different value of the weighting factor γ_f greater than 0.3. We can note that the number of incorrect maps increases at low and high frequencies, and that the RBFN error is grater in the range 80-2000 Hz.

5. Conclusions

An incoherent combination of normalized narrowband MVDR map based on a machine learning approach is proposed to mitigate the effect of incorrect narrowband power spectrum due to SNR variability at each frequency. The WMVDR consists on applying a selection of narrowband map using a RBFN classifier, which is trained on marginal distributions of response power. Preliminary experiments show that a supervised learning component trained to select the useful frequency maps to use in the acoustic information fusion can improve the performance of spatial speaker localization using microphone arrays. Future work includes the use of different learning algorithms and a validation of a simulated training system in a real-world scenario.

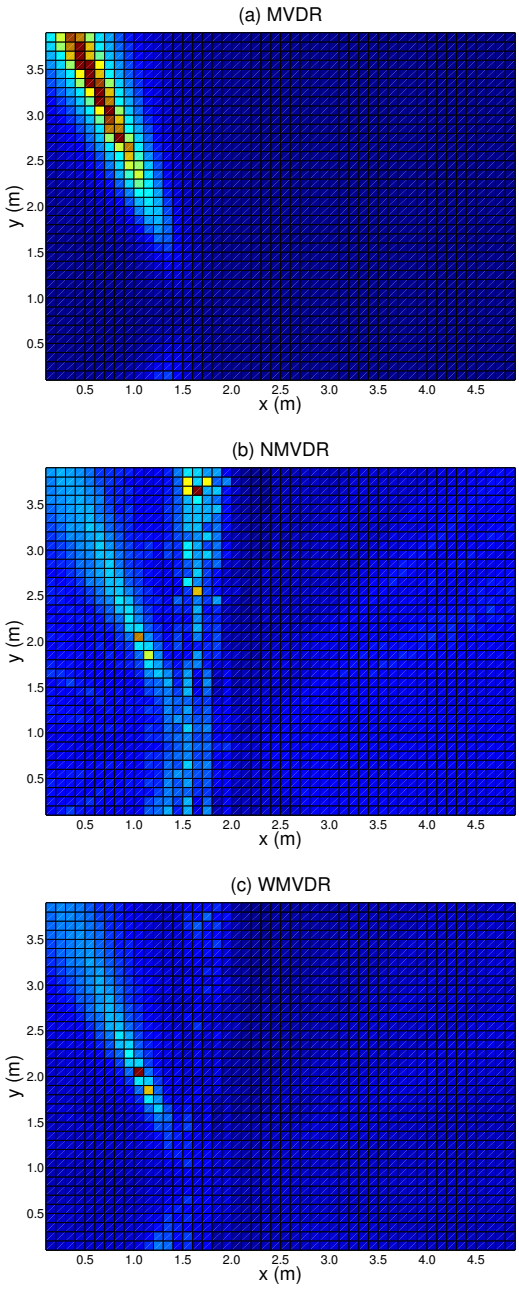

Figure 2: *Comparison of acoustic map estimates at a specific analysis block for the female speech in position s_1 (Test 1). The proposed WMVDR (c) algorithm localizes the source at the correct position, while the MVDR (a) and the NMVDR (b) algorithms provide an acoustic map with maximum value at incorrect position.*

6. References

[1] C. Seguraa, A. Abad, J. Hernando, and C. Nadeu, "Multispeaker localization and tracking in intelligent environments," *Lecture Notes in Computer Science*, vol. 4625, pp. 82–90, 2008.

[2] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 3, pp. 799 –807, 2008.

[3] D. Salvati and S. Canazza, "Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 507–510, 2013.

[4] O. Thiergart, G. D. Galdo, M. Taseska, and E. A. P. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

[5] D. Salvati and S. Canazza, "Incident signal power comparison for localization of concurrent multiple acoustic sources," *The Scientific World Journal*, vol. 2014, pp. 1–13, 2014.

[6] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, ch. Robust localization in reverberant rooms.

[7] M. R. Azimi-Sadjadi, A. Pezeshki, and N. Roseveare, "Wideband DOA estimation algorithms for multiple moving sources using unattended acoustic sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 4, pp. 1585–1599, 2008.

[8] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, 2014.

[9] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[10] E. Di Claudio and R. Parisi, "WAVES: Weighted average of signal subspaces for robust wideband direction finding," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2179–2190, 2001.

[11] Y. Yoon, L. M. Kaplan, and J. H. McClellan, "TOPS: New DOA estimator for wideband signals," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 791–802, 2006.

[12] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[13] D. S. Broomhead and D. Lowe, "Radial basis functions, multivariable functional interpolation and adaptive networks," DTIC Document, Tech. Rep., 1998.

[14] T. Nishino and K. Takeda, "Binaural sound localization for untrained directions based on a gaussian mixture model," in *Proceedings of the European Signal Processing Conference*, no. 1–5, 2008.

[15] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, no. 1–4, 2013.

[16] H. Kayser and J. Anemuller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2014, pp. 99–103.

[17] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[18] B. S. Morse, S. T. Yoo, P. Rheingans, D. T. Chen, and K. R. Subramanian, "Interpolating implicit surfaces from scattered surface data using compactly supported radial basis functions," in *Proceedings of the ACM SIGGRAPH Courses*, 2005.

[19] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Transactions on Neural Networks*, vol. 2, no. 2, pp. 302–309, 1991.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.