A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling

Maximo Cobos, *Member, IEEE*, Amparo Marti, *Student Member, IEEE*, and Jose J. Lopez, *Member, IEEE*

***Abstract*—The Steered Response Power – Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. However, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this letter, we introduce an effective strategy that extends the conventional SRP-PHAT functional with the aim of considering the volume surrounding the discrete locations of the spatial grid. As a result, the modified functional performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid. To this end, the Generalized Cross-Correlation (GCC) function corresponding to each microphone pair must be properly accumulated according to the defined microphone setup. Experiments carried out under different acoustic conditions confirm the validity of the proposed approach.**

***Index Terms*—Microphone array, sound source localization, SRP-PHAT.**

I. Introduction

SOUND source localization under high noise and reverberation still remains a very challenging task. To this end, microphone arrays are commonly employed in many sound processing applications such as videoconferencing, hands-free speech acquisition, digital hearing aids, video-gaming, autonomous robots and remote surveillance. Algorithms for sound source localization can be broadly divided into indirect and direct approaches [1]. Indirect approaches usually follow a two-step procedure: they first estimate the *Time Difference Of Arrival* (TDOA) [2] between microphone pairs and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches perform TDOA estimation and source localization in one single step by scanning a set of candidate source locations and selecting the most likely position as an estimate of the source location. In addition, information theoretic approaches have also shown to be significantly powerful in source localization tasks [3].

The *Steered Response Power – Phase Transform* (SRP-PHAT) algorithm is a direct approach that has been shown to be very robust under difficult acoustic conditions [4]–[6]. The algorithm is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [7]. Very interesting modifications and optimizations have already been proposed to deal with this problem, such as those based on Stochastic Region Contraction (SRC) [8] and coarse-to-fine region contraction [9], achieving a reduction in computational cost of more than three orders of magnitude.

In this letter, we propose a different strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the *Generalized Cross Correlation* (GCC) lag space corresponding to the volume surrounding each point of the grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

II. The SRP-PHAT Algorithm

Consider the output from microphone l, $m_l(t)$, in an M microphone system. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt \qquad (1)$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [7] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega \qquad (2)$$

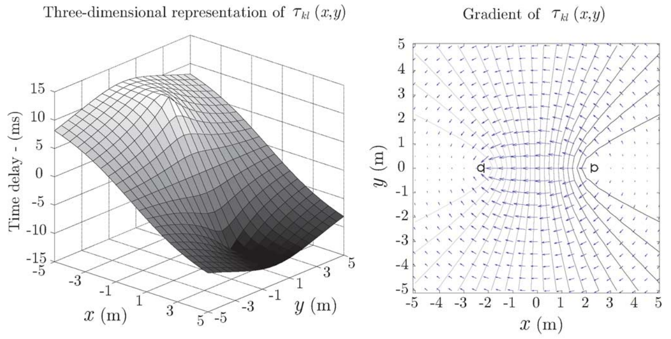
Manuscript received September 06, 2010; revised October 22, 2010; accepted October 27, 2010. Date of publication November 11, 2010; date of current version December 16, 2010. This work was suported by the The Spanish Ministry of Science and Innovation supported this work under the project TEC2009-14414-C03-01. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Constantine L. Kotropoulos.

The authors are with the Institute of Telecommunications and Multimedia Applications, Universidad Politécnica de Valencia, 46022 Valencia, Spain (e-mail: mcobos@iteam.upv.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2010.2091502

1070-9908/$26.00 © 2010 IEEE



(a) (b)

Fig. 1. Example of IMTDF. (a) Representation for the plane $z = 0$ with microphones located at $[-2, 0, 0]$ and $[2, 0, 0]$. (b) Gradient.

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega)$ is a combined weighting function in the frequency domain. The phase transform (PHAT) [10] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega)M_l^*(\omega)|}. \tag{3}$$

Taking into account the symmetries involved in the computation of (1) and removing some fixed energy terms [7], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right) \tag{4}$$

where $\tau_{kl}(\mathbf{x})$ is the *inter-microphone time-delay function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c} \tag{5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\mathbf{x}_s = \arg\max_{\mathbf{x}\in G} P'_n(\mathbf{x}). \tag{6}$$

III. The Inter-Microphone Time Delay Function

As commented in the previous section, the IMTDF plays a very important role in the source localization task. This function can be interpreted as the spatial distribution of possible TDOAs resulting from a given microphone pair geometry.

The function $\tau_{kl}(\mathbf{x})$ is continuous in $\mathbf{x}$ and changes rapidly at points close to the line connecting both microphone locations. Therefore, a pair of microphones used as a time-delay sensor is maximally sensible to changes produced over this line [11]. An example function is depicted in Fig. 1(a) for the plane $z = 0$, with $\mathbf{x}_k = [-2, 0, 0]$ and $\mathbf{x}_l = [2, 0, 0]$. The gradient of the function is shown in Fig. 1(b).

It is useful here to remark that the equation $|\tau_{kl}(\mathbf{x})| = C$, with C being a positive real constant, defines a hyperboloid in space with foci on the microphone locations $\mathbf{x}_k$ and $\mathbf{x}_l$. Moreover, the set of continuous confocal half-hyperboloids $\tau_{kl}(\mathbf{x}) = C$ with $C \in [-C_{\max}, C_{\max}]$, being $C_{\max} = (1/c)\|\mathbf{x}_k - \mathbf{x}_l\|$, spans the whole 3-D space.

Theorem: Given a volume V in space, the IMTDF for points inside V, $\tau_{kl}(\mathbf{x} \in V)$, takes only values in the continuous range $[\min\left(\tau_{kl}(\mathbf{x} \in \partial V)\right), \max\left(\tau_{kl}(\mathbf{x} \in \partial V)\right)]$, where ∂V is the boundary surface that encloses V.

Proof: Let us assume that a point inside V, $\mathbf{x}_0 \in V$, takes the maximum value in the volume, i.e., $\tau_{kl}(\mathbf{x}_0) = \max\left(\tau_{kl}(\mathbf{x} \in V)\right) = C_{\max_V}$. Since there is a half-hyperboloid that goes through each point of the space, all the points besides $\mathbf{x}_0$ satisfying $\tau_{kl}(\mathbf{x}) = C_{\max_V}$ will also take the maximum value. Therefore, all the points on the surface resulting from the intersection of the volume and the half-hyperboloid will take this maximum value, including those pertaining to the boundary surface ∂V. The existence of the minimum in ∂V is similarly deduced.

The above property is very useful to understand the advantages of the approach presented in this letter. Note that the SRP-PHAT algorithm is based on accumulating the values of the different GCCs at those time lags coinciding with the theoretical inter-microphone time delays, which are only computed at discrete points of a spatial grid. However, as described before, it is possible to analyze a complete spatial volume by scanning the time-delays contained in a range defined by the maximum and minimum values on its boundary surface. In the next section, we describe how this knowledge can be included in the localization algorithm to increase its robustness.

IV. Proposed Approach

Let us begin the description of the proposed approach by analyzing a simple case where we want to estimate the location $\mathbf{x}_s$ of a sound source inside an anechoic space. In this simple case, the GCCs corresponding to each microphone pair are delta functions centered at the corresponding inter-microphone time-delays: $R_{m_k m_l}(\tau) = \delta(\tau - \tau_{kl}(\mathbf{x}_s))$. For example and without loss of generality, let us assume a setup with $M = 3$ microphones, as depicted in Fig. 2(a). Then, the source position would be that of the intersection of the three half-hyperboloids $\tau_{kl}(\mathbf{x}) = \tau_{kl}(\mathbf{x}_s)$, with $(k, l) \in \{(1, 2), (1, 3), (2, 3)\}$. Consider now that, to localize the source, a spatial grid with resolution $r = 1$ m is used as shown in Fig. 2(a). Unfortunately, the intersection does not match any of the sampled positions, leading to an error in the localization task. Obviously, this problem would have been easier to solve with a two step localization approach, but the above example shows the limitations imposed by the selected spatial sampling in SRP-PHAT, even in optimal acoustic conditions. This is not the case of the approach followed to localize the source in Fig. 2(b) where, using the same spatial grid, the GCCs have been integrated for each sampled position in a range that covers their volume of influence. A darker gray color indicates a greater accumulated value and, therefore, the darkest area is being correctly identified as the one containing the true sound source location. This new modified functional is expressed as follows

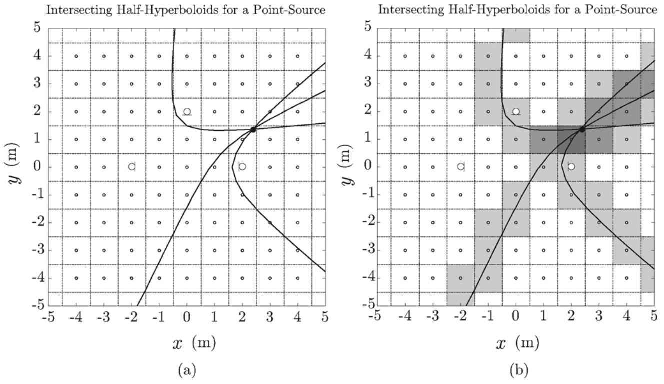


Fig. 2. Intersecting half-hyperboloids and localization approaches. (a) Conventional SRP-PHAT. (b) Proposed.

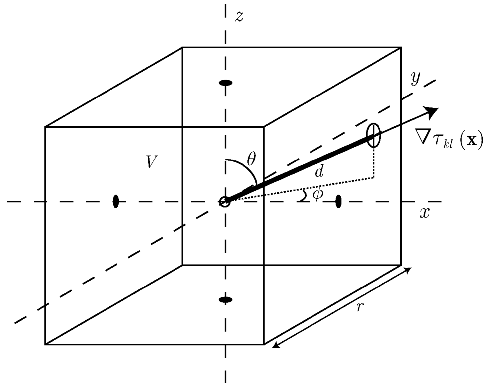


Fig. 3. Volume of influence of a point in a rectangular grid.

$$P_n''(\mathbf{x}) = \sum_{k=1}^{M}\sum_{l=k+1}^{M}\sum_{\tau=L_{kl1}(\mathbf{x})}^{L_{kl2}(\mathbf{x})} R_{m_k m_l}(\tau). \tag{7}$$

The problem is to determine correctly the limits $L_{kl1}(\mathbf{x})$ and $L_{kl2}(\mathbf{x})$, which depend on the specific IMTDF resulting from each microphone pair. The computation of these limits is explained in the next subsection.

A. Computation of Accumulation Limits

As explained in Section III, the IMTDF inside a volume can only take values in the range defined by its boundary surface. Therefore, for each point of the grid, the problem of finding the GCC accumulation limits of its volume of influence can be simplified to finding the maximum and minimum values on the boundary. To this end, it becomes useful to study the direction of the greatest rate of increase at each grid point, which is given by the gradient

$$\nabla\tau_{kl}(\mathbf{x}) = [\nabla_x\tau_{kl}(\mathbf{x}), \nabla_y\tau_{kl}(\mathbf{x}), \nabla_z\tau_{kl}(\mathbf{x})] \tag{8}$$

where each component of the gradient vector can be calculated with

$$\nabla_\gamma\tau_{kl}(\mathbf{x}) = \frac{\partial\tau_{kl}(\mathbf{x})}{\partial\gamma} = \frac{1}{c}\left(\frac{\gamma-\gamma_k}{\|\mathbf{x}-\mathbf{x}_k\|} - \frac{\gamma-\gamma_l}{\|\mathbf{x}-\mathbf{x}_l\|}\right) \tag{9}$$

where γ denotes either x, y or z. The accumulation limits for a symmetric volume surrounding a point of the grid can be calculated by taking the product of the magnitude of the gradient and the distance d that exists from the point to the boundary following the gradient's direction:

$$L_{kl1}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) - \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d, \tag{10}$$

$$L_{kl2}(\mathbf{x}) = \tau_{kl}(\mathbf{x}) + \|\nabla\tau_{kl}(\mathbf{x})\| \cdot d \tag{11}$$

Fig. 3 depicts the geometry for a rectangular grid with spatial resolution r. For this cubic geometry, the distance d can be expressed as

$$d = \frac{r}{2}\min\left(\frac{1}{|\sin\theta\cos\phi|}, \frac{1}{|\sin\theta\sin\phi|}, \frac{1}{|\cos\theta|}\right) \tag{12}$$

where

$$\theta = \cos^{-1}\left(\frac{\nabla_z\tau_{kl}(\mathbf{x})}{\|\nabla\tau_{kl}(\mathbf{x})\|}\right), \tag{13}$$

$$\phi = \mathrm{atan}_2(\nabla_y\tau_{kl}(\mathbf{x}), \nabla_x\tau_{kl}(\mathbf{x})) \tag{14}$$

being $\mathrm{atan}_2(y,x)$ the quadrant-resolving arctangent function.

B. Computational Cost

Let L be the DFT length of a frame and $Q = M(M-1)/2$ the number of microphone pairs. The computational cost of SRP-PHAT is given by [5]:

$$\begin{aligned}\mathrm{SRP-PHAT}_{cost} \approx\ & [6.125Q^2 + 3.75Q]L\log_2 L \\ & + 15LQ(1.5Q-1) + (45Q^2 - 30Q)\nu' \end{aligned} \tag{15}$$

where ν' is the average number of functional evaluations required to find the maximum of the SRP space. Since the cost added by the modified functional is negligible and the frequency-domain processing of our approach remains the same as the conventional SRP-PHAT algorithm, the above formula is valid for both approaches. Moreover, since the accumulation limits can be precomputed before running the localization algorithm, the associated processing does not involve additional computation effort. However, as it will be shown in the next subsection, the advantage of the proposed method relies on the reduced number of required functional evaluations ν' for detecting the true source location, which results in an improved computational efficiency.

V. Experiments

Different experiments with real and synthetic recordings were conducted to compare the performances of the conventional SRP-PHAT algorithm, the SRC algorithm and our proposed method. First, the *Roomsim* Matlab package [12] was used to simulate an array of six microphones placed on the walls of a shoe-box-shaped room with dimensions 4 m × 6 m × 2 m (Fig. 4(a)). The simulations were repeated with two different reverberation times ($\mathrm{T}_{60} = 0.2$ s and $\mathrm{T}_{60} = 0.7$ s), considering 30 random source locations and different signal-to-noise ratio (SNR) conditions. The resultant recordings were processed with three different spatial grid resolutions in the case of SRP-PHAT and the proposed method ($r_1 = 0.01$ m, $r_2 = 0.1$ m and $r_3 = 0.5$ m). Note that the number of functional evaluations

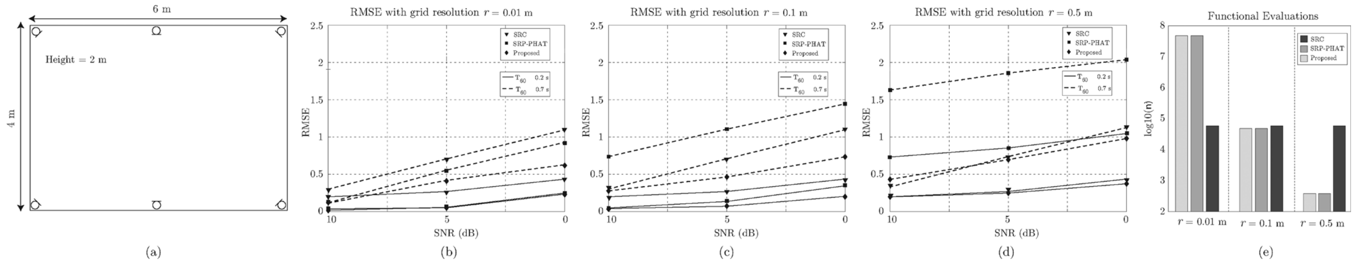


Fig. 4. Results with simulations. (a) Setup. (b) $r = 0.01\ \mathrm{m}$. (c) $r = 0.1\ \mathrm{m}$. (d) $r = 0.5\ \mathrm{m}$. (e) Functional evaluations.

TABLE I
RMSE FOR THE REAL-DATA EXPERIMENT

r ν'	0.01 $802 \cdot 10^5$	0.1 $802 \cdot 10^2$	0.5 641
SRP-PHAT	RMSE = 0.29	RMSE = 0.74	RMSE = 1.82
Proposed	RMSE = 0.21	RMSE = 0.29	RMSE = 0.31
SRC	RMSE = 0.34 ($\nu' = 58307$)		

ν' depends on the selected value of r, having $\nu'_1 = 480 \times 10^5$, $\nu'_2 = 480 \times 10^2$ and $\nu'_3 = 384$. The implementation of SRC was the one made available by Brown University's LEMS at http://www.lems.brown.edu/array/download.html, using 3000 initial random points. The processing was carried out using a sampling rate of 44.1 kHz, with time windows of 4096 samples of length and 50% overlap. The simulated sources were male and female speech signals of length 5 s with no pauses. The averaged results in terms of *Root Mean Squared Error* (RMSE) are shown in Fig. 4(b)–(d). Since SRC does not depend on the grid size, the SRC curves are the same in all these graphs. As expected, all the tested systems perform considerably better in the case of low reverberation and high SNR. For the finest grid, it can be clearly observed that the performance of SRP-PHAT and the proposed method is almost the same. However, for coarser grids, our proposed method is only slightly degraded, while the performance of SRP-PHAT becomes substantially worse, specially for low SNRs and high reverberation. SRC has similar performance to SRP-PHAT with $r = 0.01\ \mathrm{m}$. Therefore, our proposed approach performs robustly with higher grid sizes, which results in a great computational saving in terms of functional evaluations, as depicted in Fig. 4(e).

On the other hand, a real setup quite similar to the simulated one was considered to study the performance of the method in a real scenario. Six omnidirectional microphones were placed at the four corners and at the middle of the longest walls of a videoconferencing room with dimensions 5.7 m × 6.7 m × 2.1 m and 12 seats. The measured reverberation time was $T_{60} = 0.28$ s. The processing was the same as with the synthetic recordings, using continuous speech fragments obtained from the 12 seat locations. The results are shown in Table I and confirm that our proposed method performs robustly using a very coarse grid. Although similar accuracy to SRC is obtained, the number of functional evaluations is significantly reduced.

VI. CONCLUSION

This letter presented a robust approach to sound source localization based on a modified version of the well-known SRP-PHAT algorithm. The proposed functional is based on the accumulation of GCC values in a range that covers the volume surrounding each point of the defined spatial grid. The GCC accumulation limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. Our results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magnitude in the required number of functional evaluations, with further computational saving than SRC. This reduction has been shown to be sufficient for the development of real-time source localization applications.

REFERENCES

[1] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*. Hoboken, NJ: Wiley, 2008, pp. 135–166.

[2] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.

[3] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Process.*, vol. 12, no. 8, pp. 561–564, 2005.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001, pp. 157–180.

[5] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, III, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 593–606, 2005.

[6] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 4, pp. 338–347, 2003.

[7] J. H. DiBiase, "A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments using Microphone Arrays," Ph.D. dissertation, Brown Univ., Providence, RI, 2000.

[8] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, HI, Apr. 2007.

[9] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, New Paltz, NY, Oct. 2007.

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

[11] E. A. P. Habets and P. C. W. Sommen, "Optimal microphone placement for source localization using time delay estimation," in *Proc. 13th Annu. Workshop on Circuits, Systems and Signal Processing (ProRISC 2002)*, Veldhoven, The Netherlands, 2002.

[12] D. R. Campbell, Roomsim: A MATLAB Simulation Shoebox Room Acoustics 2007 [Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim