TWO-MICROPHONE SOURCE SEPARATION ALGORITHM BASED ON STATISTICAL MODELING OF ANGLE DISTRIBUTIONS

Chanwoo Kim [1], *Charbel Khawand* [2] *and Richard M. Stern* [3]

Windows Phone Division
Microsoft Corporation[1,2], Redmond WA 98052 USA
Department of Electrical and Computer Engineering
Carnegie Mellon University[3], Pittsburgh PA 15213 USA

ABSTRACT

In this paper we present a novel two-microphone sound source separation algorithm, which selects speech from the target speaker while suppressing signals from interfering sources. In this algorithm, which is refered to as SMAD-CW, we first estimate the direction of sound sources for each time-frequency bin using phase differences in the spectral domain. For each frame we assume that the angle distribution is a mixture of two distributions, one from the target and the other from the dominant noise source. For each mixture component we use the von Mises distribution, which is a close approximation to the wrapped normal distribution. The expectation-maximization (EM) algorithm is employed to obtain parameters of this mixture distribution. Using this statistical model, we perform maximum *a posteriori* (MAP) hypothesis testing in order to obtain appropriate binary masks. We demonstrate that the algorithm described in this paper provides speech recognition accuracy that is significantly better than that obtained using conventional approaches.

Index Terms— Robust speech recognition, signal separation, interaural time difference, statistical modeling, binaural hearing, von Mises distribution

1. INTRODUCTION

Speech recognition systems have significantly improved in recent years, and they have been used in many applications. Even though we can obtain high speech recognition accuracy in clean environments using state-of-the-art speech recognition systems, performance seriously degrades in noisy environments. Thus, noise robustness remains a critical issue for speech recognition systems that are used for real consumer products in difficult acoustical environments.

Many algorithms have been developed to address these problems, and a number of them have proved to be of significant value in reducing the impact stationary noise. Nevertheless, improvement in non-stationary noise remains elusive. An alternative approach is signal separation based on analysis of differences in arrival time (*e.g.* [1, 2]). It is well known that the human binaural system is remarkable in its ability to separate speech from interfering sources (*e.g.* [3]). Motivated by these observations, many models and algorithms have been developed using interaural time differences (ITDs), interaural intensity difference (IIDs), interaural phase differences (IPDs), and other cues (*e.g.* [1, 2, 4]). IPD and ITD have been extensively used in binaural processing because this information can be easily obtained by spectral analysis (*e.g.* [5]). In the present approach, we use *statistical modeling of angle distributions with channel weighting* (SMAD-CW) instead of a fixed threshold to determine which signal components belong to the target signal and which components are part of the background noise.

This research was supported by the Microsoft Corporation and by the NSF (Grant IIS-I0916918). The authors are grateful to Prof. Bhiksha Raj and Kshitiz Kumar for many useful discussions.

2. STRUCTURE OF THE SMAD-CW ALGORITHM

The SMAD-CW algorithm crudely emulates selected aspects of human binaural processing and is summarized by the block diagram of Fig. 1. While the description below assumes a sampling rate of 16 kHz and 4 cm between the two microphones, the algorithm is easily modified to accommodate other sampling frequencies and microphone separations. In our discussion we assume that the location of the target source is known *a priori*, and lies along the perpendicular bisector of the line between the two microphones.

Short-time Fourier transforms (STFTs) are performed on the signals from the left and right microphonesusing Hamming windows of duration 75 ms, 37.5 ms between successive frames, and a DFT size of 2048. The choice of a rather long window has been discussed previously (*e.g.* [5]). For each time-frequency bin, the direction of the sound source is estimated indirectly by comparing the phase information from the two microphones. Either the angle or ITD information is used as a statistic to represent the direction of the sound source, as described in Sec. 3.1.

Most conventional algorithms using a pair of microphones compare the signal components in each time-frequency bin to a threshold angle or ITD to determine whether the signal component in each time-frequency bin is likely to originate from the target or a noise source (*e.g.* [1, 5]). The SMAD-CW algorithm, in contrast, models the angle distribution for each frame as a mixture of two Von Mises distributions; one from the target and the other from the noise source. The von Mises distribution, which is a close approximation to the wrapped normal distribution, is used rather than the well-known Gaussian distribution because the angle is limited between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Parameters of the distribution are estimated using the expectation-maximization (EM) algorithm, as described in Sec. 3.2.

After obtaining parameters of the angle distribution, we perform maximum *a posteriori* (MAP) testing on each time-frequency bin. From these results binary masks are constructed based on whether a specific time-frequency bin is likely to be occupied by the target distribution or the noise distribution. Hence, SMAD-CW employs a soft decision approach based on statistical hypothesis testing.

To obtain better speech recognition accuracy in noisy environments, we apply the gammatone channel weighting approach intro-

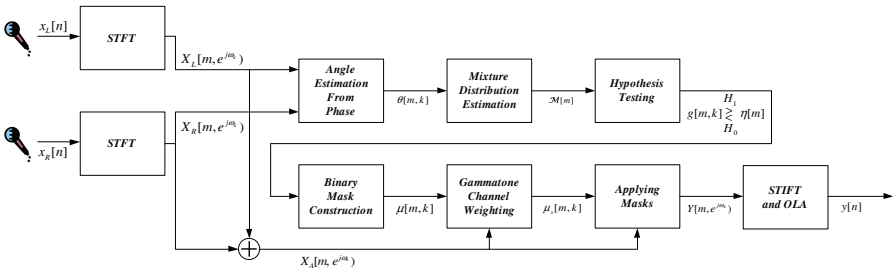

Fig. 1. *The block diagram of a sound source separation system based on the Statistical Modeling of Angles and likelihood Ratio Testing (SMAD) algorithm.*

duced in [5] rather than directly applying the binary mask. In gammatone channel weighting, the ratio of power after applying the binary mask to the original power is obtained for each channel, which is subsequently used to modify the original input spectrum, as described in Sec. 3.4. Finally, the time domain signal is obtained by the overlap-add (OLA) method. The SWAD algorithm with channel weighting is referred to as SMAD-CW. Each component of the SWAD-CW algorithm is described in further detail in Sec. 3.

3. COMPONENTS OF SMAD-CW PROCESSING

3.1. Estimation of the angle for each frequency index

In each frame, the phase differences between the left and right spectra are used to estimate the intermicrophone time difference (ITD), and subsequently the angle of the sound source, as described previously in [5] and elsewhere. Let $X_L[m, e^{j\omega_k})$ and $X_R[m, e^{j\omega_k})$ represent the STFT of the signals from the left and right microphones, respectively, where $w_k = 2\pi k/N$ and N is the FFT size. We refer to $\tau[m, k]$ as the ITD at frame index m and frequency index k. We obtain the relationship:

$$\phi[m,k] \triangleq \angle X_R[m, e^{j\omega_k}) - \angle X_L[m, e^{j\omega_k}) = w_k \tau[m,k] + 2\pi l \tag{1}$$

where l is an integer chosen such that

$$w_k \tau[m,k] = \begin{cases} \phi[m,k], & \text{if } |\phi[m,k]| \le \pi \\ \phi[m,k] - 2\pi, & \text{if } \phi[m,k] \ge \pi \\ \phi[m,k] + 2\pi, & \text{if } \phi[m,k] < -\pi \end{cases} \tag{2}$$

(In these discussions we consider only values of the frequency index k that correspond to positive frequency components, $0 \le k \le \pi/2$.) We use Eq. (1) and (2) to obtain $\tau[m, k]$ from the measured values $\angle X_R[m, e^{j\omega_k})$ and $\angle X_L[m, e^{j\omega_k})$.

If a sound source is located along a line of angle $\theta[m, k]$ with respect to the perpendicular bisector to the line between the microphones, geometric considerations determine the ITD $\tau[m, k]$ to be

$$\tau[m,k] = d \sin(\theta[m,k]) f_s / c_{air} \tag{3}$$

where c_{air} is the speed of sound in air (assumed to be 340 m/s in our work) and f_s is the sampling rate.

While in principle $|\tau[m, k]|$ cannot be larger than $\tau_{max} = f_s d/c_{air}$ from Eq. (3), in real environments $|\tau[m, k]|$ may be larger than τ_{max} because of approximations in the assumptions that were made if the ITD is estimated directly from Eq. (1) and (2). For this reason we limit $\tau[m, k]$ to lie between between $-\tau_{max}$ and τ_{max} and we refer to this limited ITD estimate as $\tilde{\tau}[m, k]$. The estimated angle $\theta[m, k]$ is obtained from $\tilde{\tau}[m, k]$ using

$$\theta[m,k] = \text{asin}\left(\frac{c_{air}\tilde{\tau}[m,k]}{f_s d}\right) \tag{4}$$

3.2. Statistical modeling of the angle distribution

For each frame, the distribution of estimated angles $\theta[m, k]$ is modeled as a mixture of the target and the noise distributions:

$$\begin{aligned} f_T(\theta|\mathcal{M}[m]) \quad = \quad & c_0[m] f_0(\theta|\mu_0[m], \kappa_0[m]) \\ & + c_1[m] f_1(\theta|\mu_1[m], \kappa_1[m]) \end{aligned} \tag{5}$$

where $c_1[m]$ and $c_0[m]$ are the mixture coefficients, and $\mathcal{M}[m]$ is the set of parameters of the mixture distribution. In this section we use the subscript 0 to represent the noise and the subscript 1 to represent the target. Specifically,

$$\mathcal{M}[m] = \left\{c_1[m], \mu_0[m], \mu_1[m], \kappa_0[m], \kappa_1[m]\right\} \tag{6}$$

$f_1(\theta|\mu_1[m], \kappa_1[m])$ and $f_0(\theta|\mu_0[m], \kappa_0[m])$ are given as follows:

$$f_0(\theta|\mu_0[m], \kappa_0[m]) = \frac{\exp(\kappa_0[m]\cos(2\theta - \mu_0[m]))}{\pi I_0(\kappa_0[m])} \tag{7a}$$

$$f_1(\theta|\mu_1[m], \kappa_1[m]) = \frac{\exp(\kappa_1[m]\cos(2\theta - \mu_1[m]))}{\pi I_0(\kappa_1[m])} \tag{7b}$$

The coefficient $c_0[m]$ follows directly from the constraint $c_0[m] + c_1[m] = 1$. Since the parameters $\mathcal{M}[m]$ cannot be directly estimated in closed form, we obtain them using the EM algorithm. We impose the following constraints in parameter estimation:

$$0 \le |\mu_1[m]| \le \theta_0 \tag{8a}$$

$$\theta_0 \le |\mu_0[m]| \le \frac{\pi}{2} \tag{8b}$$

$$\theta_0 \le |\mu_1[m] - \mu_0[m]| \tag{8c}$$

where θ_0 is a fixed angle that equals $15\pi/180$ in the present work. This constraint is applied both in the initial stage and the update stage explained below. Without this constraint $\mu_0[m]$ and $\kappa_0[m]$ may converge to the target mixture or $\mu_1[m]$ and $\kappa_1[m]$ may converge to the interference mixture, which would be problematical.

Initial parameter estimation: To obtain the initial parameters of $\mathcal{M}[m]$, we consider the following two partitions of the frequency index k

$$\mathcal{K}_0[m] = \left\{k \middle| \left|\theta[m,k]\right| \ge \theta_0, 0 \le k \le N/2\right\} \tag{9a}$$

$$\mathcal{K}_1[m] = \left\{k \middle| \left|\theta[m,k]\right| < \theta_0, 0 \le k \le N/2\right\} \tag{9b}$$

In this initial step, we assume that if the frequency index k belongs to $\mathcal{K}_1[m]$, then this time-frequency bin is dominated by the target; therwise, we assume that it is dominated by the noise. This initial step is similar to approaches using a fixed threshold. Consider a variable $z[m, k]$ which is defined as follows:

$$z[m,k] = e^{j2\theta[m,k]} \tag{10}$$

Let us define the weighted average $\bar{z}_j^{(0)}[m], j = 0, 1$:

$$\bar{z}_j^{(0)}[m] = \frac{\sum_{k=0}^{N/2} \rho[m,k]\mathbb{I}\left(\theta[m,k] \in \mathcal{K}_j\right) z[m,k]}{\sum_{k=0}^{N/2} \rho[m,k]\mathbb{I}\left(\theta[m,k] \in \mathcal{K}_j\right)}, \tag{11}$$

where $\mathbb{I}$ is the indicator function. The following equations ($j = 0, 1$) are used in analogy to Eq. (17).

$$c_j^{(0)}[m] = \frac{\sum_{k \in \mathcal{K}_j} \rho[m,k]}{\sum_{k=0}^{N/2} \rho[m,k]} \tag{12a}$$

$$\mu_j^{(0)}[m] = \text{Arg}\left(\bar{z}_j^{(0)}[m]\right) \tag{12b}$$

$$\frac{I_1(\kappa_j^{(0)}[m])}{I_0(\kappa_j^{(0)}[m])} = |\bar{z}_j^{(0)}[m]| \tag{12c}$$

where $I_0(\kappa_j)$ and $I_1(\kappa_j)$ are modified Bessel functions of the zeroth and first order. For the first frame ($m = 0$), we initialize the variables for the target by $\mu_1^{(0)}[0] = 0$ and $\kappa_1^{(0)}[0] = 200$, which are typical values from the actual target utterances. This is done because, in the first several frames, there might not be any target speech at all.

Parameter update: The E-step is given as follows:

$$\tilde{Q}(\mathcal{M}[m], \mathcal{M}^{(t)}[m]) = \sum_{k=0}^{N/2} \rho[m,k] E\Big[\log f_T\left(\theta[m,k], s[m,k] \Big| \theta[m,k], \mathcal{M}^{(t)}[m]\right) \Big] \tag{13}$$

where $\rho[m,k]$ is a weighting coefficient defined by $\rho[m,k] = |X_A[m, e^{j\omega_k}]|^2$, and $s[m,k]$ is the latent variable denoting whether the k^{th} frequency element originates from the target source or the noise source. $X_A[m, e^{j\omega_k}]$ is defined by:

$$X_A[m, e^{j\omega_k}] = \Big[X_L[m, e^{j\omega_k}) + X_R[m, e^{j\omega_k})\Big]/2 \tag{14}$$

Given the current estimated model $\mathcal{M}^{(t)}[m]$, we define the conditional probability $T_j^{(t)}[m,k], j = 0, 1$ as follows:

$$\begin{aligned} T_j^{(t)}[m,k] &= P(s[m,k] = j | \theta[m,k], \mathcal{M}^{(t)}[m]), \\ &= \frac{c_j^{(t)} f_j(\theta[m,k]|\mu_j, \kappa_j)}{\sum_{j=0}^{1} c_j^{(t)} f_j(\theta[m,k]|\mu_j, \kappa_j)} \end{aligned} \tag{15}$$

Let us define the weighted mean of $\bar{z}_j^{(t)}[m], j = 0, 1$ as follows:

$$\bar{z}_j^{(t)}[m] = \frac{\sum_{k=0}^{N/2} \rho[m,k] T_j^{(t)}[m,k] z[m,k]}{\sum_{k=0}^{N/2} \rho[m,k] T_j^{(t)}[m,k]} \tag{16}$$

Using Eqs. (15) and (16), it can be shown that the following update equations ($j = 0, 1$) maximize Eq. (13):

$$c_j^{(t+1)}[m] = \frac{\sum_{k=0}^{\frac{N}{2}} \rho[m,k] T_j^{(t)}[m,k]}{\sum_{k=0}^{\frac{N}{2}} \rho[m,k]} \tag{17a}$$

$$\mu_j^{(t+1)}[m] = \text{Arg}\left(\bar{z}_j^{(t)}[m]\right) \tag{17b}$$

$$\frac{I_1(\kappa_j^{(t+1)}[m])}{I_0(\kappa_j^{(t+1)}[m])} = |\bar{z}_j^{(t)}[m]| \tag{17c}$$

Assuming that the target speaker does not move rapidly with respect to the microphone, we apply the following smoothing to improve performance:

$$\tilde{\mu}_1[m] = \lambda \mu_1[m-1] + (1-\lambda)\mu_1[m] \tag{18}$$

$$\tilde{\kappa}_1[m] = \lambda \kappa_1[m-1] + (1-\lambda)\kappa_1[m] \tag{19}$$

with the forgetting factor λ equal to 0.95. The parameters $\tilde{\mu}_1[m]$ and $\tilde{\kappa}_1[m]$ are used instead of $\mu_1[m]$ and $\kappa_1[m]$ in subsequent iterations. This smoothing is not applied to the representation of the noise source.

3.3. Hypothesis Testing

Using the obtained model $\mathcal{M}[m]$ and Eq. (7), we obtain the following MAP decision criterion:

$$g[m,k] \underset{H_0}{\overset{H_1}{\gtrless}} \eta[m] \tag{20}$$

where $g[m,k]$ and $\eta[m]$ are defined as follows:

$$\begin{aligned} g[m,k] = \kappa_1[m]\cos\left(2\theta[m,k] - \mu_1[m]\right) \\ -\kappa_0[m]\cos\left(2\theta[m,k] - \mu_0[m]\right) \end{aligned} \tag{21}$$

$$\eta[m] = \ln\left(\frac{I_0(\kappa_1[m])c_0[m]}{I_0(\kappa_0[m])c_1[m]}\right) \tag{22}$$

Using Eq. (20) we construct a binary mask $w_b[m,k]$ for each frequency index k as follows:

$$w_b[m,k] = \begin{cases} 1 & \text{if } g[m,k] \geq \eta[m] \\ 0 & \text{if } g[m,k] < \eta[m] \end{cases} \tag{23}$$

Processed spectra are obtained by applying the mask $w_b[m,k]$, and speech is resynthesized using the IFFT and OLA. This approach (without the channel weighting described in Sec. 3.4) is referred to as SMAD reconstruction.

3.4. Applying channel weighting

To reduce the impact of discontinuities associated with binary masks, we obtain a weighting coefficient for each channel. Each of these channels is associated with $H_l\left(e^{j\omega_k}\right)$, the frequency response of one of a set of gammatone filters, as specified in [6]. Let $w[m,l]$ be the square root of the ratio of the output power to the input power for frame index m and channel index l:

$$w[m,l] = \max\left(\sqrt{\frac{\sum_{k=0}^{\frac{N}{2}-1}\left|X_A[m, e^{j\omega_k}) w_b[m,k] H_l\left(e^{j\omega_k}\right)\right|^2}{\sum_{k=0}^{\frac{N}{2}-1}\left|X_A[m, e^{j\omega_k}) H_l\left(e^{j\omega_k}\right)\right|^2}}, \delta\right) \tag{24}$$

where δ is a flooring coefficient that is set to 0.01 in the present implementation. Note that the channel weighting coefficient $w[m,l]$ is somewhat different from the coefficient in our previous paper [5]. Using $w[m,l]$, speech is resynthesized in the same fashion as in [5].

4. EXPERIMENTAL RESULTS

In this section we present experimental results using the SMAD-CW algorithm described in this paper. To evaluate the effectiveness of

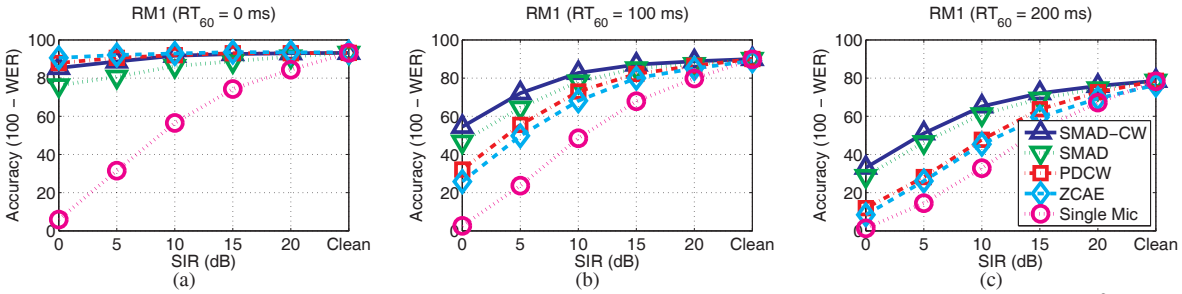

Fig. 2. Comparison of recognition accuracy for the DARPA RM database corrupted by an interfering speaker placed at 30° with respect to the perpendicular bisector to the line connecting the two microphones with three reverberation times: (a) 0 ms, (b) 100 ms, and (c) 200 ms.

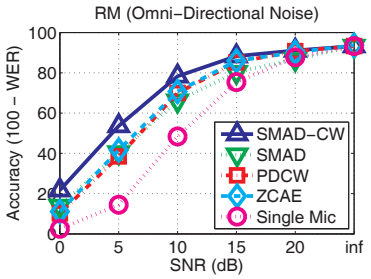

Fig. 3. Speech recognition accuracy using different algorithms in the presence of natural real-world noise.

the statistical modeling of angle distributions and channel weighting, we compare performance of the SMAD-CW, SMAD, with the state-of-the art PDCW algorithm, as well as the baseline processing provided by the ZCAE algorithm [2] using binary masking. For ZCAE processing, we use zero-phase gammatone filter coefficients as described in [7].

Speech recognition experiments were performed using the reconstructed signals obtained as in Sec. 3 in conjunction with conventional MFCC features implemented as in `sphinx_fe` in `sphinxbase 0.4.1`. For acoustic model training we used `SphinxTrain 1.0`, and decoding was performed suing the `CMU Sphinx 3.8`.We used subsets of 1600 utterances and 600 utterances, respectively, from the DARPA Resource Management (RM1) database for training and testing. A bigram language model was used in all experiments. In all experiments, we used feature vectors of length of 39 including delta and delta-delta features. We assumed that the distance between two microphones is 4 cm.

The first set of experiments was conducted using simulated reverberant environments in which the target speaker is masked by a single interfering speaker. We assumed that the target is located along the perpendicular bisector of the line between two microphones, so $\theta_T = 0^\circ$. We assume that the interfering source is located at $\theta_I = 30^\circ$. Reverberation simulations were accomplished using the *Room Impulse Response* open source software package [8] based on the image method [9]. In the experiments in this section, we assumed room dimensions of $5 \times 4 \times 3$ m, with microphones that are located at the center of the room. Both the target and interfering sources are 1.5 m away from the microphone. For the fixed-ITD-threshold systems PDCW and ZCAE, we used the threshold angle $\theta_{TH} = 15^\circ$. As shown in Fig. 2(a), in the absence of reverberation at 0-dB signal-to-interference ratio (SIR), the fixed-ITD-threshold systems PDCW and ZCAE and the SMAD-CW system provide comparable performance. In contrast, the SMAD-CW system provides substantially better performance than the PDCW signal separation system in the presence of reverberation.

In the second set of experiments, we added noise recorded in real environments with real two-microphone hardware in locations such as a public market, a food court, a city street and a bus stop. These real noises were digitally added to the clean test set of the DARPA RM database. Fig. 3 shows the speech recognition accuracy obtained for these data. Again we observe that SMAD-CW shows the best performance by a significant margin, and the SMAD, PDCW and ZCAE provide similar but worse) performance.

The MATLAB code for the SMAD-CW algorithm can be found at `http://www.cs.cmu.edu/~robust/archive/algorithms/SMAD_ICASSP2012/`. We note that a US patent application has been applied for part of this work by the Microsoft Corporation.

5. REFERENCES

[1] S. Srinivasan, M. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Comm.*, vol. 48, pp. 1486–1501, 2006.

[2] H. Park, and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings," *Speech Communication*, vol. 51, no. 1, pp. 15–25, Jan. 2009.

[3] W. Grantham, "Spatial hearing and related phenomena," in *Hearing*, B. C. J. Moore, Ed. Academic, 1995, pp. 297–345.

[4] P. Arabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Tran. Systems, Man, and Cybernetics-Part B:*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.

[5] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *INTERSPEECH-2009*, Sept. 2009, pp. 2495–2498.

[6] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, (in submission).

[7] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May. 2011, pp. 5072–5075.

[8] S. G. McGovern, "A model for room acoustics," http://2pi.us/rir.html.

[9] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.