DOA-ESTIMATION BASED ON A COMPLEX WATSON KERNEL METHOD

Lukas Drude, Florian Jacob, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany

ABSTRACT

This contribution presents a Direction of Arrival (DoA) estimation algorithm based on the complex Watson distribution to incorporate both phase and level differences of captured microphone array signals. The derived algorithm is reviewed in the context of the Generalized State Coherence Transform (GSCT) on the one hand and a kernel density estimation method on the other hand. A thorough simulative evaluation yields insight into parameter selection and provides details on the performance for both directional and omni-directional microphones. A comparison to the well known Steered Response Power with Phase Transform (SRP-PHAT) algorithm and a state of the art DoA estimator which explicitly accounts for aliasing, shows in particular the advantages of presented algorithm if inter-sensor level differences are indicative of the DoA, as with directional microphones.

Index Terms— Direction of Arrival, sensor array, directional statistics, complex Watson distribution, directional sensors

1. INTRODUCTION

Acoustic sensor networks are of great importance for many signal processing applications, such as advanced teleconferencing systems, distributed hearing-aids and monitoring and surveillance systems. An important signal processing task to be carried out on such networks is acoustic speaker localization and tracking. Most speaker localization algorithms employ Time difference of Arrival (TDoA) or DoA estimates, and the localization performance in reverberant and noisy environments critically depends on the quality of these estimates. Popular DoA estimation algorithms are the Multiple Signal Classification (MUSIC) algorithm [1], which relies on an eigendecomposition of the power spectral density matrix of the microphone signals, and the SRP-PHAT algorithm [2], which is the generalization of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) algorithm [3] for more than two microphones. GCC-PHAT allows a direct computation of the DoA using the delay corresponding to the maximum of cross correlation function of the sensor signals, while MUSIC and SRP-PHAT require an evaluation of a model for each possible candidate direction.

This work was in part supported by Deutsche Forschungsgemeinschaft under contract no. Ha3455/7-2.

While these algorithms do not explicitly treat reverberation, a possibility to improve performance in reverberant environments is to extract the signal components related to the direct sound propagation path followed by an ordinary TDoA estimation based on the GCC-PHAT algorithm [4].

A fairly recent approach, which proved effective in reverberant environments, is proposed in [5]. It is a special case of the GSCT [6], and it employs a cosine distance measure between the observations and anechoic candidate models, i.e., the expected observations in an anechoic sound field originating from a candidate position. With the cosine distance it accounts for the spatial aliasing problem, which arises for frequencies, where the sensor distance is higher than half a wavelength.

The algorithm proposed here is in some sense similar, as it also accounts for spatial aliasing and uses a nonlinearity. However, it is derived from a statistical model. It makes use of the complex Watson distribution to describe the probabilities of each observation to be generated by one of the anechoic candidate models. The Watson distribution is a distribution of complex vectors which are normalized to unit length. It found early applications in the field of image recognition [7], and it has appeared in the area of beamforming and Blind Source Separation (BSS) as an alternative to the complex Gaussian distribution to model microphone array signals in the Short Time Fourier Transform (STFT) domain [8–11].

The motivation to model the observations as unit-length vectors instead of a vector of complex STFT coefficients is the fact that the length of each STFT coefficient vector is mainly caused by the source. On the contrary, the differences between the individual coefficients are related to the level and phase differences, which are mainly caused by the transmission path. Further, the distribution naturally accounts for aliasing due to its invariance with respect to multiplication of vector components with $\mathrm{e}^{\mathrm{j}2\pi}$.

In this paper we formulate DoA estimation as Maximum Likelihood (ML) estimation involving the Watson distribution. This results in a statistically motivated distance measure, which resembles the beamforming concept. A main improvement is the fact, that the proposed algorithm explicitly exploits level differences between microphones, since it does not solely rely on phase differences. This leads to an improved DoA estimation, if the level differences are indicative of the DoA, as is the case with directional microphones. Additionally,

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

section 4-5 explains how the proposed method can be viewed either from a kernel density estimation or GSCT point of view and therefore unifies the different views.

2. OBSERVATION MODEL

Let us consider a reverberated speech signal $s(t)$ captured by an array of D microphones:

$$x_d(t) = h_{d,\mathrm{direct}}(t) * s(t) + h_{d,\mathrm{rev}}(t) * s(t) + n_d(t), \quad (1)$$

where $h_{d,\mathrm{direct}}(t)$ covers the impulse response due to the time delay and microphone directivity of the direct path, $h_{d,\mathrm{rev}}(t)$ covers the reverberation and microphone directivity for the reflections and $n_d(t)$ is a noise term for each sensor $d = 1, \ldots, D$.

This time domain model can be approximated by a multiplicative transfer function model in STFT domain:

$$\mathbf{X}(m,k) = \mathbf{H}_{\mathrm{direct}}(k) S(m,k) + \tilde{\mathbf{N}}(m,k), \quad (2)$$

where $\mathbf{X}(m,k)$ is the vector of observations in STFT domain: $\mathbf{X}(m,k) = [X_d(m,k)]_{1\leq d\leq D}$, and where $X_d(m,k) = \mathrm{STFT}\{x_d(t)\}$. $\mathbf{H}_{\mathrm{direct}}(k)$ is the frequency bin dependent transfer function corresponding to the line-of-sight propagation path, and $S(m,k)$ is the source signal. The vector $\tilde{\mathbf{N}}(m,k)$ comprises both the noise term and the reverberation, which, in a first approximation, can be assumed to be uncorrelated to the direct signal component [12].

The varying power due to the speaker is removed by a unit-norm normalization, which maintains level and phase differences:

$$\mathbf{Y}(m,k) = \mathbf{X}(m,k) / \|\mathbf{X}(m,k)\|, \quad (3)$$

where $\|\cdot\| = \sqrt{\mathbf{X}^{\mathrm{H}}\mathbf{X}}$. The observations $\mathbf{Y}(m,k)$ are now vectors on the complex unit hyper-sphere, i.e., $\mathbf{Y}^{\mathrm{H}}\mathbf{Y} = 1$. Please note, that this normalization does not effect the relative level differences between the channels. The relative level differences can therefore still be used for the DoA estimation, especially when the sensors are directive.

In a source localization or DoA estimation, one is interested in the direct, line-of-sight signal component, impinging on the microphone from the source location. Observations $\mathbf{Y}$ originating from the direct path have similar level and phase differences as in the anechoic scenario. Thus, we derive the anechoic model vector $\mathbf{W}(k,\mathbf{p})$ which aims to model $\mathbf{H}_{\mathrm{direct}}(k)$ for all possible source locations $\mathbf{p}$ relative to the sensor array center.

The TDoA relative to the sensor array center is given by

$$\tau_d(\mathbf{p}) = (\|\mathbf{p} - \mathbf{m}_d\| - \|\mathbf{p}\|) \cdot c^{-1}, \quad (4)$$

where c is the speed of sound and $\mathbf{m}_d$ is the location of sensor d relative to the sensor array center. The anechoic model attenuation is given by

$$A_d(\mathbf{p}) = \alpha + (1-\alpha)\frac{\mathbf{n}_d^{\mathrm{T}}(\mathbf{p} - \mathbf{m}_d)}{\|\mathbf{n}_d\| \, \|\mathbf{p} - \mathbf{m}_d\|}, \quad (5)$$

where α governs the microphone directivity (i.e. $\alpha = 1$ for an omni-directional and $\alpha = 0.75$ for a sub-cardioid microphone), and $\mathbf{n}_d$ is the look-direction of each sensor. This equation can be replaced by, possibly frequency dependent, measured sensor directivity patterns. For simplicity reasons, the inverse square law for sound radiation has not been applied. Nevertheless, it can easily be incorporated as an additional factor in equation 5.

Finally, the anechoic unit-norm normalized model is composed:

$$\widetilde{\mathbf{W}}(k,\mathbf{p}) = \left[A_d(\mathbf{p}) \mathrm{e}^{-2\pi \mathrm{j} f(k) \tau_d(\mathbf{p})}\right]_{1\leq d\leq D}, \quad (6)$$

$$\mathbf{W}(k,\mathbf{p}) = \widetilde{\mathbf{W}}(k,\mathbf{p}) / \left\|\widetilde{\mathbf{W}}(k,\mathbf{p})\right\|, \quad (7)$$

where $f(k)$ is the center frequency of the k-th frequency bin. In the later, the anechoic model is used to identify observations which are most likely caused by the line-of-sight propagation.

3. STATISTICAL MODEL

Since both the anechoic model vectors and the observations are mapped to the complex unit hyper-sphere, a probability distribution accounting for this fact has to be used.

The complex Watson Probability Density Function (PDF) is defined on the complex unit hyper-sphere:

$$p(\mathbf{Y}; \kappa, \mathbf{W}) = \frac{1}{c_{\mathrm{W}}(\kappa)} \mathrm{e}^{\kappa |\mathbf{W}^{\mathrm{H}}\mathbf{Y}|^2}, \quad (8)$$

where $\mathbf{W}$ is the mode vector, κ is a real-valued concentration parameter and $c_{\mathrm{W}}(\kappa)$ is a normalization constant [7]. Note that the value of $\kappa = 0$ corresponds to a uniform distribution on the hyper-sphere.

It becomes apparent, that the distribution is invariant to multiplication of $\mathbf{Y}$ or $\mathbf{W}$ with a complex scalar of unit norm and therefore is not influenced by phase terms introduced by the source. Furthermore, the probability of $\mathbf{Y}$ does not change, if individual vector components are multiplied with $\mathrm{e}^{\mathrm{j}2\pi}$ and, thus, the distribution naturally accounts for the aliasing problem. The distance measure $|\mathbf{W}^{\mathrm{H}}\mathbf{Y}|^2$ within the distribution has a plausible form, since it is equivalent to the normalized response power of a beamformer. It is therefore easier to motivate than the cosine-distance measure based solely on TDoAs which is used in [5].

Now, the PDF values $p(\mathbf{Y}(m,k); \kappa, \mathbf{W}(k,\mathbf{p}))$ can be averaged over all observations and calculated for each possible source position:

$$J(\mathbf{p}) = \frac{1}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} p(\mathbf{Y}(m,k); \kappa, \mathbf{W}(k,\mathbf{p})), \quad (9)$$

where M and K are the number of time frames and frequency bins, respectively. Finally, the estimated source position is chosen by a maximum selection, yielding the ML estimate.

4. RELATION TO KERNEL DENSITY ESTIMATION

A kernel density estimator yields an estimate of a probability density function $\hat{p}_k(\widetilde{\mathbf{Y}})$ by summing over M kernel functions, each centered at an observation $\mathbf{Y}(m,k)$ [13]:

$$\hat{p}_k(\widetilde{\mathbf{Y}}) = \sum_{m=1}^{M} \frac{1}{c_{\mathrm{W}}(\kappa)} \mathrm{e}^{\kappa \left|\mathbf{Y}^{\mathrm{H}}(m,k)\widetilde{\mathbf{Y}}\right|^2}, \quad \widetilde{\mathbf{Y}}^{\mathrm{H}}\widetilde{\mathbf{Y}} = 1. \quad (10)$$

If we now evaluate $\hat{p}_k(\widetilde{\mathbf{Y}})$ with $\widetilde{\mathbf{Y}} = \mathbf{W}(k,\mathbf{p})$ at all possible anechoic candidate positions $\mathbf{p}$ and average over all frequency bins k, we achieve an estimate of the scaled probabilities of each candidate and result in the score function $J(\mathbf{p})$ of (9).

5. RELATION TO STATE COHERENCE TRANSFORM

Nesta and Omologo presented the GSCT as a ML estimator of the data given all possible state models [6]. They defined a state to be the parameters encoding the properties of the acoustic signal propagation. Furthermore, they proposed to calculate state models depending on time differences only, whereas we generalized this to include arbitrary attenuations (level differences).

They then chose the Euclidean norm of the difference between the observation as predicted by the model and the true observations as the distance measure. In our case, the complex Watson distribution incorporates the squared absolute value of the scalar product of the normalized observation and the anechoic model. This appears to be a more natural choice, due to the fact that the response power of the beamformer is exactly that.

Finally, they proposed to apply the non-linearity $g(x) = 1 - \tanh(\alpha x/\sqrt{4D})$ to the distance arguing that this non-linearity increases the resolution of the GSCT likelihood. Here, this heuristically chosen non-linearity is replaced by the exponential function due to the complex Watson distribution. Only the scalar κ remains heuristic instead of an entire function.

6. EXPERIMENTAL EVALUATION

The proposed algorithm is evaluated by simulating a reverberant enclosure using the image method [14] with speech samples taken from the TIMIT database [15]. An isotropic noise field is generated using an isotropic noise field generator [16, 17]. Room sizes are randomly sampled between 4 and 5 m edge length. The sensor array is randomly placed approximately in the center of the room. The true source position is randomly chosen on a circle with radius 1.5 m centered around the randomly rotated sensor array. The DoA estimation statistics given in the following are the average of 1000 simulated configurations.

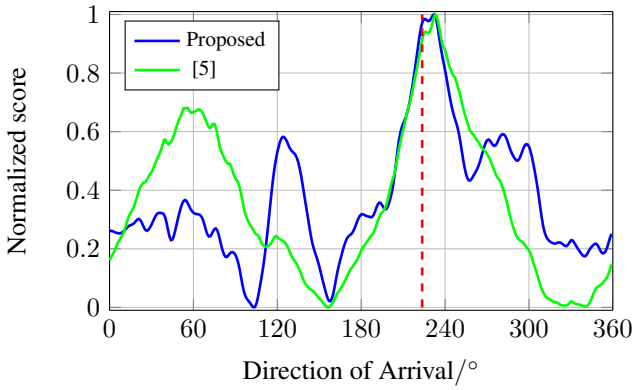

Fig. 1: Normalized score functions for [5] and the proposed algorithm with the true DoA in red. Both algorithms suffer due to the spherical noise and high reverberation and do not estimate the true source position exactly.

In the simulations, instead of calculating models for arbitrary source positions $\mathbf{p}$, we evaluate models for all DoAs, with a resolution of 1° in the farfield. We deliberately decided to compare the algorithms in terms of a one-dimensional search to allow a simplified comparison, although all three algorithms are appropriate for three-dimensional search. We therefore calculated the farfield models, since the distance of the source to the sensors is assumed to be unknown.

Signal duration	1 s
Sampling rate	16 kHz
Reverberation time T_{60}	400 ms
Number of sensors D	6
Radius of circular array	10 cm
Sensor directivity	Omni-directional
SNR	10 dB
STFT size and shift	1024 and 256 samples
STFT window function	Blackman

Table 1: Standard simulation parameters if not explicitly altered in given plot.

Algorithm [5] and the proposed algorithm allow to easily incorporate observation selection schemes by only considering observations for which a given criterion is met. An a posteriori SNR is calculated by dividing the energy in a given bin by a noise energy estimate from a noise tracker. Subsequently, local and global smoothing windows are applied to the a posteriori SNR values. Finally, voice-activity-detection estimates are ob-

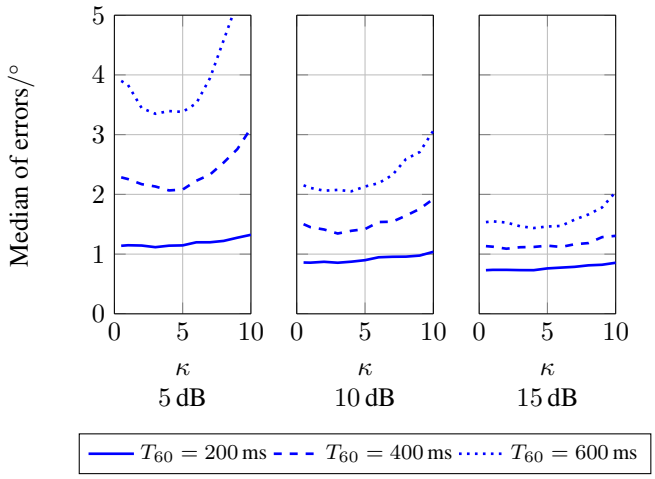

Fig. 2: Evaluation of optimal design parameter κ with respect to different SNR and reverberation conditions for the proposed algorithm.

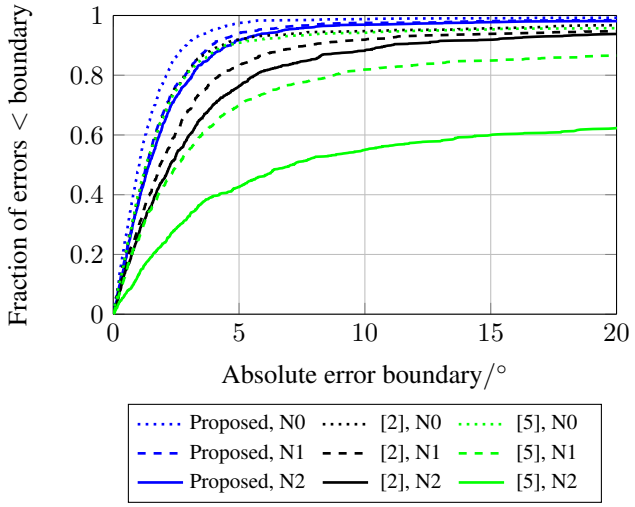

Fig. 3: DoA estimation performance in spectrally and spatially white noise (N0), in spectrally pink and spatially white noise (N1), and in a spectrally white and spatially isotropic noise field (N2) for the different estimators.

tained by thresholding the local and global smoothing results. This method is a stripped down version of the speech-presence-probability estimator given [18, Section 44.5]. Note that the SRP-PHAT algorithm does not easily allow an observation selection.

Figure 1 shows example score functions for [5] and the proposed score function, where both are normalized to fit into a common plot.

The concentration parameter of the complex Watson distribution governs the smoothness of the score function. Higher values result in a peaky kernel, meaning that more focus is put on observations which are close to the candidate steering vectors, while lower values of κ lead to smoother score functions, however sacrificing resolution.

We first evaluated whether the parameter can be chosen independently of the SNR and reverberation conditions. We therefore keep the noise type fixed and vary κ for different conditions. The results are presented in Figure 2, where we chose the median instead of the mean square error as a scalar DoA estimation error measure motivated by the fact that drastic outliers govern mean square error. It can be observed that $\kappa = 5$ is a good choice fairly independent of the noise and reverberation conditions. Thus, we chose $\kappa = 5$ for all other simulations. This is also the value that is underlying the score function depicted in Figure 1. There is a similar parameter governing the smoothness of the score function in the algorithm by [5]. It is set to $\alpha = 1$ in all simulations. All further parameters are listed in Table 1.

We now analyze the impact of different stationary noise types on the proposed estimator and compare it to the results obtained by the SRP-PHAT algorithm [2] and the GSCT variant [5]. Figure 3 shows the cumulative histogram of errors. As an example of how to read this figure, one can see that 80 % of the errors of the proposed algorithm are below 3° in an isotropic noise field.

Using the cumulative histogram is motivated by the fact that drastic outliers (i.e. 180° error) greatly influence the mean squared error whereas the cumulative histogram allows to visualize the distribution of errors.

All algorithms achieve the highest performance in spectrally pink and spatially white noise conditions (N1). Spatially white noise affects the algorithms more because higher frequencies carry less speech power and are therefore less reliable in white noise conditions (N0). A spectrally white and spatially isotropic noise field is clearly the harshest condition (N2).

We finally evaluate the performance of all three algorithms for directional microphones. We compare frequency independent omni-directional, sub-cardioid and cardioid directivity patterns.

The blue curves in Figure 4 show the performance of the proposed algorithm. The different directivities do not influence the performance much, since the model mode vectors incorporate the attenuations as indicated in equation (5).

The black and green curves show the performance of [2] and [5], respectively. Both do not incorporate level differences as features, leading to degraded performance, in particular for cardioid microphones.

7. CONCLUSION

In this paper we have proposed a source position or DoA estimator based on a complex Watson kernel. The estimator is statistically motivated and, unlike the estimator proposed in [5], does not rely on heuristically defined non-linearities and distance measures. This contribution can be seen as a motivation to use the complex Watson distribution not only in

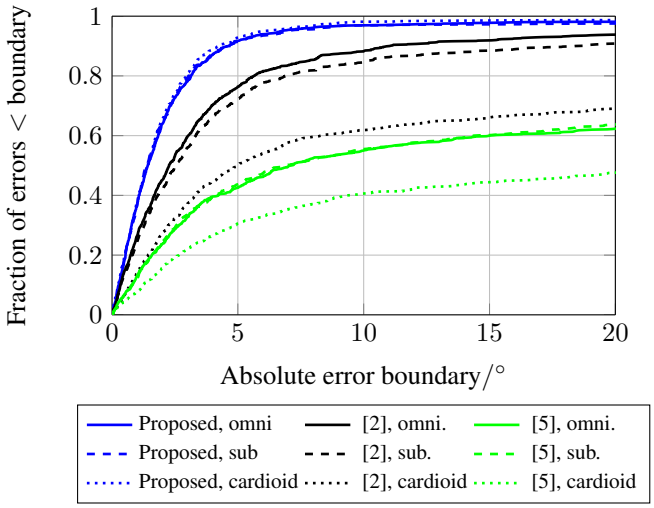

Fig. 4: Performance of all algorithms with given sensor directivity.

the context of BSS but also in DoA estimation. The proposed estimator allows to incorporate microphone level differences in addition to time differences. It is capable of dealing with time frequency selection schemes, which is not directly possible in well known algorithms such as MUSIC and SRP-PHAT. When the level differences contain information about the source direction and the sensor directivities are known, the proposed algorithm clearly outperforms the alternatives.

REFERENCES

[1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.

[3] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[4] S. Mosayyebpour, H. Lohrasbipeydeh, M. Esmaeili, and T. A. Gulliver, "Time delay estimation via minimum-phase and all-pass component processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, 2013, pp. 4285–4289.

[5] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Latent Variable Analysis and Signal Separation*, pp. 1–8. Springer, 2010.

[6] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional TDoA estimation of multiple sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 246–260, Jan 2012.

[7] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[8] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 241–244.

[9] L. Drude, A. Chinaev, D. H. Tran Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6834–6838.

[10] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3238–3242.

[11] I. Jafari, R. Togneri, and S. Nordholm, "On the use of the Watson mixture model for clustering-based underdetermined blind source separation," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 988–992.

[12] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[13] C. M. Bishop et al., *Pattern recognition and machine learning*, vol. 1, Springer New York, 2006.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CDROM*, U.S. Department of Commerce, 1993.

[16] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[17] E. A. P. Habets and S. Gannot, "Comments on Generating sensor signals in isotropic noise fields," 2010.

[18] J. Benesty, M. Mohan. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, 2008.