Linear and Parametric Microphone Array Processing

Part 1 - Introduction

Emanuël A. P. Habets[1] and Sharon Gannot[2]

[1] International Audio Laboratories Erlangen, Germany
A joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS

[2] Faculty of Engineering, Bar-Ilan University, Israel

ICASSP 2013, Vancouver, Canada

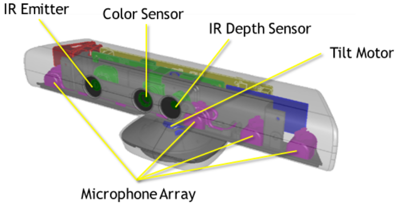
Presenters

Prof. Dr. Emanuël Habets (Emanuel.Habets@audiolabs-erlangen.de)

- **2007** PhD degree from the Technische Universiteit Eindhoven, The Netherlands
- **2007-2008** Postdoctoral Fellow at the Technion and Bar-Ilan University, Israel
- **2009-2010** Research Fellow at Imperial College London, UK
- **2010 - present** Professor at the University of Erlangen-Nuremberg (FAU), Germany
- **2010 - present** Group Manager and Chief Scientist for Spatial Audio Processing at Fraunhofer IIS (Home of mp3), Germany

Prof. Dr. Sharon Gannot (Sharon.Gannot@biu.ac.il)

- **2000** PhD degree from Tel-Aviv University, Israel
- **2001** Postdoctoral Fellow at K.U. Leuven, Belgium
- **2002-2003** Researcher / Lecturer at the Technion - Israel Institute of Technology, Israel
- **2004-present** Professor at Bar-Ilan University (BIU), Israel


IR Emitter
Color Sensor
IR Depth Sensor
Tilt Motor
Microphone Array

Overview

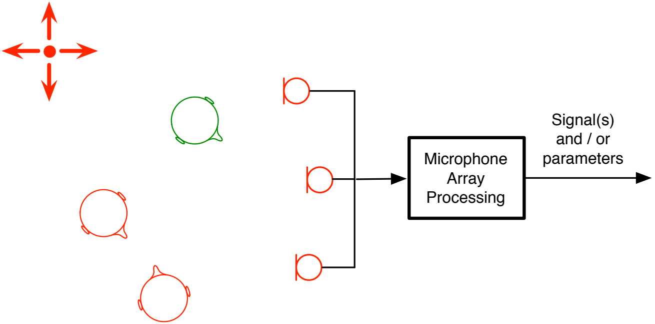
1. Introduction and Motivation

Problems we can (potentially) solve using microphone array processing:

- Extract desired sounds that are corrupted by interfering sounds
  - **Noise reduction**
  - **Reverberation reduction**
  - Echo reduction
- Localize sound sources
- Determine the number of (active) sound sources

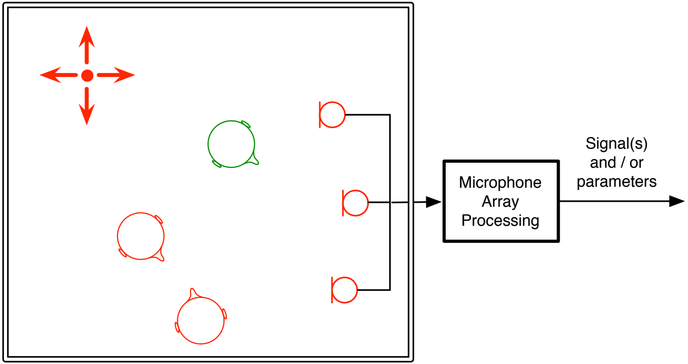
Applications:

- Hands-free communication systems
- Hands-free human-machine interfaces (e.g., TVs, smartphones)
- Teleconferencing systems
- Hearing-aid devices
- Assistive listening devices
- ...



Interferers:

- Spatially coherent noise (e.g., used to model sound sources)
- Spatially incoherent noise (e.g., used to model sensor noise)
- Diffuse noise (e.g., used to model reverberation, car noise, cocktail-party noise and babble noise)



Interferers:

- Spatially coherent noise (e.g., used to model sound sources)
- Spatially incoherent noise (e.g., used to model sensor noise)
- Diffuse noise (e.g., used to model reverberation, car noise, cocktail-party noise and babble noise)

Why is microphone array processing so different from antenna array processing?

Challenges (to name a few):

- Speech signals are wideband and highly non-stationary
- Noise often has the same spectral characteristics as the desired sounds
- Room reverberation / diffuse sound field
- Time-varying spatial characteristics
- Number of sensors and placement is usually restricted
- The human ear has a very high dynamic range
- Knowing what is desired and what is undesired

AUDIO LABS

Why is microphone array processing so different from antenna array processing?

Challenges (to name a few):

- Speech signals are wideband and highly non-stationary
- Noise often has the same spectral characteristics as the desired sounds
- Room reverberation / diffuse sound field
- Time-varying spatial characteristics
- Number of sensors and placement is usually restricted
- The human ear has a very high dynamic range
- Knowing what is desired and what is undesired

Overview

2. Spatial Processing Approaches

We can divide existing approaches into three categories:

- **Linear Spatial Processing** A linear filter is applied to the observed microphone signals. The filter is based on, for example, the estimated second-order statistics of the observed and noisy signals. In many cases, estimates of the (relative) acoustic transfer functions are employed. Can be applied in both centralized and distributed manner.
- **Parametric Spatial Processing** A perceptually or physically motivated parametric sound field model is assumed. The model parameters such as the direction-of-arrival, position and signal-to-diffuse ratio are estimated using multiple microphones. Based on these parameters, a time and frequency dependent gain is computed and applied to a reference signal (e.g., one of the microphones or fixed beamformer).
- **Informed Spatial Processing** The main idea behind informed spatial filtering is to incorporate relevant information about the specific problem into the design of spatial filters and the estimation of the required statistics and/or propagation vector(s).

2. Spatial Processing Approaches

We can divide existing approaches into three categories:

- **Linear Spatial Processing** A linear filter is applied to the observed microphone signals. The filter is based on, for example, the estimated second-order statistics of the observed and noisy signals. In many cases, estimates of the (relative) acoustic transfer functions are employed. Can be applied in both centralized and distributed manner.

- **Parametric Spatial Processing** A perceptually or physically motivated parametric sound field model is assumed. The model parameters such as the direction-of-arrival, position and signal-to-diffuse ratio are estimated using multiple microphones. Based on these parameters, a time and frequency dependent gain is computed and applied to a reference signal (e.g., one of the microphones or fixed beamformer).

- **Informed Spatial Processing** The main idea behind informed spatial filtering is to incorporate relevant information about the specific problem into the design of spatial filters and the estimation of the required statistics and/or propagation vector(s).

2. Spatial Processing Approaches

We can divide existing approaches into three categories:

- **Linear Spatial Processing** A linear filter is applied to the observed microphone signals. The filter is based on, for example, the estimated second-order statistics of the observed and noisy signals. In many cases, estimates of the (relative) acoustic transfer functions are employed. Can be applied in both centralized and distributed manner.

- **Parametric Spatial Processing** A perceptually or physically motivated parametric sound field model is assumed. The model parameters such as the direction-of-arrival, position and signal-to-diffuse ratio are estimated using multiple microphones. Based on these parameters, a time and frequency dependent gain is computed and applied to a reference signal (e.g., one of the microphones or fixed beamformer).

- **Informed Spatial Processing** The main idea behind informed spatial filtering is to incorporate relevant information about the specific problem into the design of spatial filters and the estimation of the required statistics and/or propagation vector(s).

2. Spatial Processing Approaches

Outline of today's tutorial:

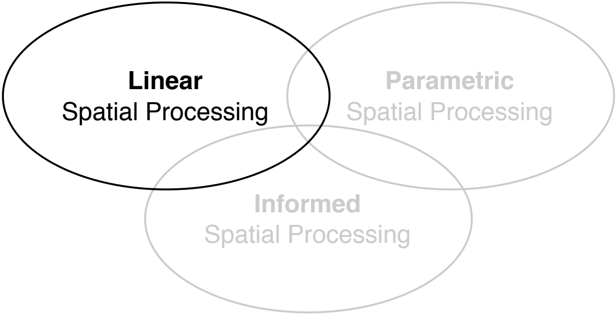


Figure: Different spatial processing approaches.

2. Spatial Processing Approaches

Outline of today's tutorial:

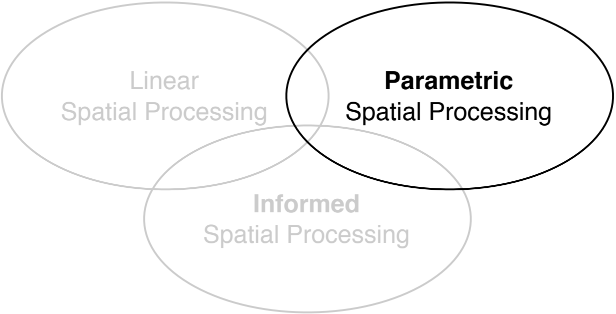


Figure: Different spatial processing approaches.

2. Spatial Processing Approaches

Outline of today's tutorial:

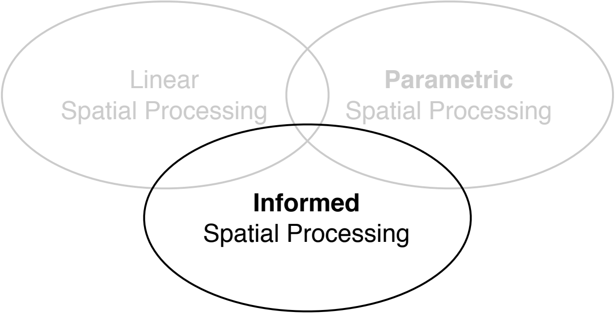


Figure: Different spatial processing approaches.

2. Spatial Processing Approaches

Different **objectives**:

- Estimate the anechoic signal as received by one of the microphones.
- Estimate the reverberant signal as received by one of the microphones (See, for example, [Gannot et al., 2001, Benesty et al., 2008, Benesty et al., 2011]).
- Estimate the signal provided by a *signal-independent beamformer* or *single-channel/multichannel equalizer* (See, for example, [Habets et al., 2010, Habets and Benesty, 2013]).
- ...

Different **optimization criteria**:

- Minimum mean squared error → Multichannel Wiener filter (MWF).
- Constrained minimization → Parametric MWF (a.k.a. speech-distortion weighted MWF).
- Constrained minimization → Minimum variance distortionless response (MVDR) beamformer.
- Constrained minimization → Linearly constrained minimum variance (LCMV) beamformer.

2. Spatial Processing Approaches

Different **objectives**:

- Estimate the anechoic signal as received by one of the microphones.
- Estimate the reverberant signal as received by one of the microphones (See, for example, [Gannot et al., 2001, Benesty et al., 2008, Benesty et al., 2011]).
- Estimate the signal provided by a *signal-independent beamformer* or *single-channel/multichannel equalizer* (See, for example, [Habets et al., 2010, Habets and Benesty, 2013]).
- ...

Different **optimization criteria**:

- Minimum mean squared error $\rightarrow$ Multichannel Wiener filter (MWF).
- Constrained minimization $\rightarrow$ Parametric MWF (a.k.a. speech-distortion weighted MWF).
- Constrained minimization $\rightarrow$ Minimum variance distortionless response (MVDR) beamformer.
- Constrained minimization $\rightarrow$ Linearly constrained minimum variance (LCMV) beamformer.

Some important facts:

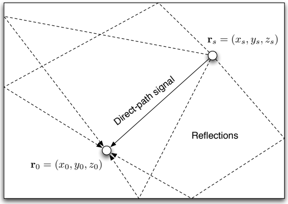
- All filters (expect for the LCMV) maximize the subband output signal-to-noise ratio.
- All filters (expect for the LCMV) are equal up to a frequency-dependent scaling factor.
- All filters are different in terms of the amount of noise reduction and speech distortion.
- Depending on the assumed propagation model, we can find different filter expressions such as, for example, the rank-1 multichannel Wiener filter.

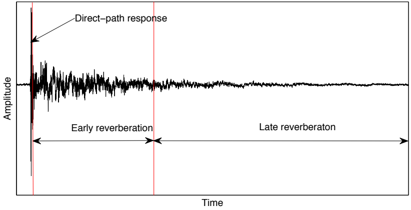
Overview

3.1 Room Impulse Response

Room impulse responses consist of:

- Direct path
- Early reflections → Haas effect, precedence effect and coloration
- Late reflections → can reduce speech intelligibility





Further reading: [Kuttruff, 2000].

3.1 Room Impulse Response

- Room impulse response:

$$h_{(\mathbf{r}_\mathrm{o},\mathbf{r}_\mathrm{s})}(t) = \sum_{i=1}^{\infty} g_i(t) * \delta(t - \tau_i),$$

  where τ_i denotes the time-of-arrival of the i-th reflection and $g_i(t)$ denotes the impulse response of the i-th reflection.

- Room transfer function:

$$H_{(\mathbf{r}_\mathrm{o},\mathbf{r}_\mathrm{s})}(\omega) = \sum_{i=1}^{\infty} G_i(\omega) \exp(-j\omega\tau_i).$$

  where $G(\omega)$ denotes the Fourier transform of $g_i(t)$.

- Statistical models have been proposed to model the RIR in [Polack, 1988, Jot et al., 1997] and the RTF in [Schroeder, 1962, Schroeder, 1987].

3.1 Room Impulse Response

One popular way to simulate RIRs is to use image sources [Allen and Berkley, 1979]:

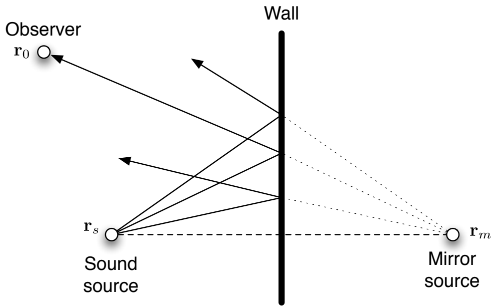


Figure: Source image method.

An implementation is available at:
`http://home.tiscali.nl/ehabets/rir_generator.html`

The energy decay curve (EDC) is defined as:

$$\mathrm{EDC}(t) = \int_{t}^{\infty} h^2_{(\mathbf{r}_\mathrm{O},\mathbf{r}_\mathrm{S})}(t')\,\mathrm{d}t'.$$

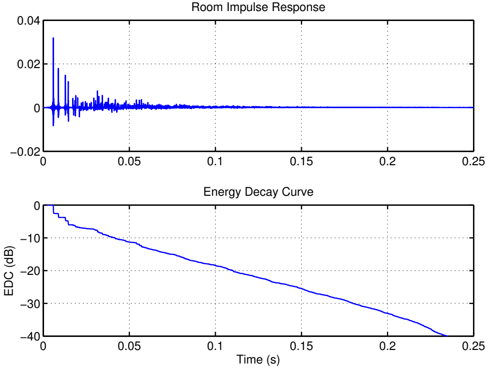


Figure: Example of an energy decay curve.

3.2 Reverberation Time

- The reverberation time quantifies the severity of reverberation within a room, and is often denoted by RT_{60}.
- It is defined as the time that is necessary for a 60 dB decay of the sound energy after switching off a sound source.

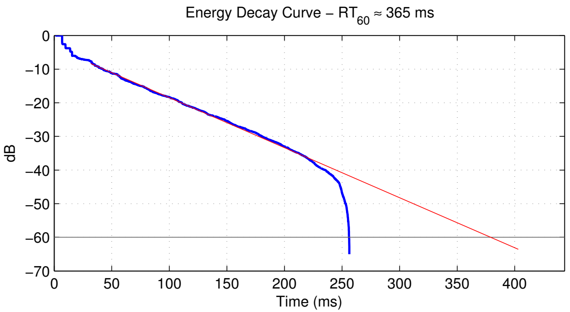


Figure: Determining the reverberation time.

3.3 Spatial Coherence

Definition of the (complex) spatial coherence:

$$\Gamma_{X_1X_2}(\omega) = \frac{\int_{\mathbb{A}} P_{X_1X_2}(\omega)\ d\mathbb{A}}{\int_{\mathbb{A}} \sqrt{P_{X_1}(\omega)P_{X_2}(\omega)}\ \mathrm{d}\mathbb{A}}.$$

The mean-squared coherence is given by $|\Gamma_{X_1X_2}(\omega)|^2$.

Coherent sound field:

$$\Gamma_{X_1X_2}(\omega) = \frac{P_{X_1X_2}(\omega)}{\sqrt{P_{X_1}(\omega)P_{X_2}(\omega)}} = e^{-j\,\frac{\omega}{c}\,d\cos(\phi)},$$

where $P_{X_1} = P_{X_2}$ and $P_{X_1X_2} = P_{X_1}e^{-j\,\frac{\omega}{c}\,d\cos(\phi)}$.

Incoherent sound fields:

$$\Gamma_{X_1X_2}(\omega) = 0,$$

because $P_{X_1X_2} = 0$.

Cylindrically isotropic sound field (2D diffuse) with $\mathrm{d}\mathbb{A} = r\ \mathrm{d}\phi$ and $A = 2\pi r$:

$$\Gamma_{X_1 X_2}(\omega) = \frac{1}{2\pi}\int_0^{2\pi} e^{-j\frac{\omega}{c}d\cos\phi}\ \mathrm{d}\phi$$
$$= J_0(\omega d/c),$$

where $J_0(\,\cdot\,)$ is the zero-order Bessel function of the first kind.

Spherically Isotropic sound field (3D diffuse) with $\mathrm{d}\mathbb{A} = r^2 sin(\phi)\ \mathrm{d}\phi\ \mathrm{d}\theta$ and $A = 4\pi r^2$:

$$\Gamma_{X_1 X_2}(\omega) = \frac{1}{4\pi r^2}\int_0^{2\pi}\int_0^{\pi} e^{-j\frac{\omega}{c}d\cos\phi} r^2 \sin(\phi)\ \mathrm{d}\phi\ \mathrm{d}\theta$$
$$= \frac{\sin(\omega d/c)}{\omega d/c}.$$

3.4 Simulators

- RIR generator:
  http://home.tiscali.nl/ehabets/rir_generator.html

- Signal generator (time-varying RIRs):
  http://home.tiscali.nl/ehabets/signal_generator.html

- Spherical microphone array RIR generator: [Jarrett et al., 2012]:
  http://home.tiscali.nl/ehabets/smirgen.html

  Note: This simulator can also be used as a mouth simulator!

- Spherical and cylindrical isotropic noise generator [Habets and Gannot, 2007]:
  http://home.tiscali.nl/ehabets/publications/Habets2007b.html

- Generating nonstationary multisensor signals (such as babble speech) that exhibit a pre-defined spatial coherence [Habets et al., 2008]:
  http://home.tiscali.nl/ehabets/publications/Habets2008.html

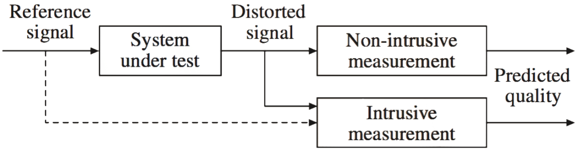
Overview

4. Evaluation

Quality assessment by:

1. **Subjective listening test**
   - Extremely valuable but time consuming and expensive.
   - The test needs to be carefully designed.

2. **Objective quality measures**
   - Quantify the quality by measuring a “distance” between the original and processed signals.
   - Objective measure are most useful when there is a high correlation with subjective listening test results.
   - For that reason, many objective measures exploit aspects of the auditory system.



4.1 Subjective Listening Test

The mean opinion score (MOS) is a widely used and accepted criterion for speech coder assessment. Although not very suitable for the evaluation of speech enhancement algorithms it is often used.

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Table: Mean opinion score scale

4.1 Subjective Listening Test

ITU-T P.835 was recommended specifically for the evaluation of speech enhancement algorithms.

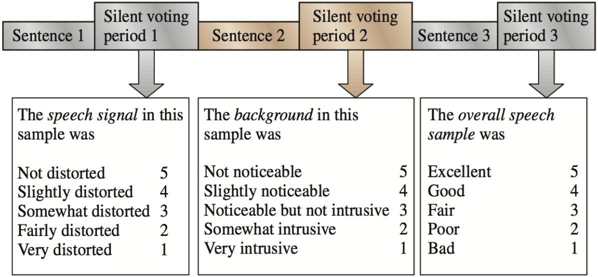


Figure: The ITU-T P-835's scheme for evaluating the subjective quality of speech enhancement algorithms. Each test sample is comprised of three subsamples, where each subsample is followed by a silent voting period.

4.2 Intrusive Objective Quality Measures

For noise reduction:

- Mean-squared error (MSE)
- Signal to noise ratio (SNR)
- Segmental SNR - average SNR in dB across frames (geometric mean)
- Log-spectral distance (LSD)
- Itakura-Saito (IS) - based on linear prediction coefficients
- Noise reduction factor
- Speech reduction factor
- ...

In the context of array processing (subband/fullband):

- Array gain
- Directivity factor
- Directivity index
- White noise gain (be careful - higher values are better!)
- ...

These measures are often computed over short-time frames and subsequently averaged across frames.

4.2 Intrusive Objective Quality Measures

Perceptually motivated quality measures:

- Weighted segmental SNR (computed in the frequency-domain)
- Weighted spectral slope (WSS)
- Bark spectral distortion (BSD)
- Perceptual evaluation of speech quality (PESQ)
  ITU Recommendation ITU-T P.862
- Perceptual speech quality measure (PSQM)
  ITU Recommendation ITU-T P.861
- Perceptual evaluation of audio quality (PEAQ)
  ITU Recommendation BS.1387
- Low complexity speech quality assessment (LCQA)
- ...

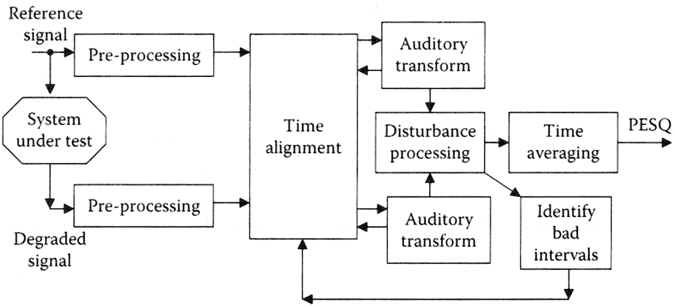


Figure: Block diagram PESQ measure [Loizou, 2007]

4.2 Intrusive Objective Quality Measures

Designed to evaluate how much reverberation is present / reduced.

Signal-based:

- Signal to reverberation ratio
- Segmental signal to reverberation ratio [Naylor et al., 2010]
- Reverberation decay tail (RDT) [Wen and Naylor, 2006]
- Speech to reverberation modulation energy ratio (SRMER) [Falk et al., 2010]
- ...

Channel-based:

- Reverberation time
- Direct to reverberation ratio
- Early decay time (RT_{60} measured over the first 10 dB of the decay)
- Early to late reverberation ratio (a.k.a. Clarity or Klarheitsmass)
- Early to total energy ratio (a.k.a. Definition or Deutlichkeit)
- ...

References I

Allen, J. B. and Berkley, D. A. (1979).
Image method for efficiently simulating small-room acoustics.
J. Acoust. Soc. Am., 65(4):943–950.

Benesty, J., Chen, J., and Habets, E. A. P. (2011).
Speech Enhancement in the STFT Domain.
SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag.

Benesty, J., Chen, J., and Huang, Y. (2008).
Microphone Array Signal Processing.
Springer-Verlag, Berlin, Germany.

Falk, T., Zheng, C., and Chan, W.-Y. (2010).
A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech.
IEEE Trans. Audio, Speech, Lang. Process., 18(7):1766–1774.

Gannot, S., Burshtein, D., and Weinstein, E. (2001).
Signal enhancement using beamforming and nonstationarity with applications to speech.
IEEE Trans. Signal Process., 49(8):1614–1626.

References II

Habets, E. A. P. and Benesty, J. (2013).
A two-stage beamforming approach for noise reduction and dereverberation.
IEEE Trans. Audio, Speech, Lang. Process., 21(5):945–958.

Habets, E. A. P., Benesty, J., Cohen, I., Gannot, S., and Dmochowski, J. (2010).
New insights into the MVDR beamformer in room acoustics.
IEEE Trans. Audio, Speech, Lang. Process., 18:158–170.

Habets, E. A. P., Cohen, I., and Gannot, S. (2008).
Generating nonstationary multisensor signals under a spatial coherence constraint.
J. Acoust. Soc. Am., 124(5):2911–2917.

Habets, E. A. P. and Gannot, S. (2007).
Generating sensor signals in isotropic noise fields.
J. Acoust. Soc. Am., 122(6):3464–3470.

Jarrett, D. P., Habets, E. A. P., Thomas, M. R. P., and Naylor, P. A. (2012).
Rigid sphere room impulse response simulation: algorithm and applications.
J. Acoust. Soc. Am., 132(3):1462–1472.

References III

Jot, J.-M., Cerveau, L., and Warusfel, O. (1997).
Analysis and synthesis of room reverberation based on a statistical time-frequency model.
In *Proc. Audio Eng. Soc. Convention*.

Kuttruff, H. (2000).
Room Acoustics.
Taylor & Francis, London, fourth edition.

Loizou, P. C. (2007).
Speech Enhancement Theory and Practice.
Taylor & Francis.

Naylor, P. A., Gaubitch, N. D., and Habets, E. A. P. (2010).
Signal-based performance evaluation of dereverberation algorithms.
Journal of Electrical and Computer Engineering, 2010:1–5.

Polack, J. D. (1988).
La transmission de l'énergie sonore dans les salles.
PhD thesis, Université du Maine, Le Mans, France.

References IV

Schroeder, M. R. (1962).
Frequency correlation functions of frequency responses in rooms.
J. Acoust. Soc. Am., 34(12):1819–1823.

Schroeder, M. R. (1987).
Statistical parameters of the frequency response curves of large rooms.
Journal Audio Eng. Soc., 35:299–306.

Wen, J. Y. C. and Naylor, P. (2006).
An evaluation measure for reverberant speech using tail decay modelling.
In *Proc. European Signal Processing Conf. (EUSIPCO)*, pages 1–4, Florence, Italy.