A HYBRID METHOD FOR BLIND ESTIMATION OF FREQUENCY DEPENDENT REVERBERATION TIME USING SPEECH SIGNALS

Song Li, Roman Schlieper, and Jürgen Peissig

Gottfried Wilhelm Leibniz Universität Hannover
Institute of Communications Technology
Hannover, Germany
song.li@ikt.uni-hannover.de

ABSTRACT

Reverberation time is an important room acoustical parameter that can be used to identify the acoustic environment, predict speech intelligibility and model the late reverberation for binaural rendering, etc. Several blind estimation algorithms of reverberation time have been proposed by analyzing recorded speech signals. Unfortunately, the estimation accuracy for the frequency dependent reverberation time is lower than for the full-band reverberation time due to the lower signal energy in sub-band filters. This study presents a novel approach for the blind estimation of reverberation time in the full frequency range. The maximum likelihood method is applied for the estimation of the reverberation time from low- to mid-frequencies, and the reverberation time from mid- to high-frequencies is predicted by our proposed model based on the analysis of the reverberation time calculated from room impulse responses in different rooms. The proposed method is validated by two experiments and shows a good performance.

***Index Terms*—** reverberation time, full frequency range, blind estimation, room acoustics, speech signals.

1. INTRODUCTION

Reverberation time (RT_{60}) is defined as the time it takes for the acoustical energy density in an enclosure to attenuate by $60\,\mathrm{dB}$ after switching off the sound source. The conventional methods for estimating RT_{60} are based on recorded white noises (interrupted noise method [1]), measured room impulse responses (Schroeder method [2]), or the Sabine or Eyring equations [3, 4]. However, these methods are not always suitable for calculating the RT_{60}, especially in common consumer scenarios and in noisy environments. It is desirable to obtain the RT_{60} from a recorded reverberant signal without knowing the excitation signal, the geometry and the surface material of the room. Recently, several algorithms [5, 6, 7, 8, 9, 10] have been proposed to blindly estimate the RT_{60} based on recorded speech signals. Some of these methods, such as [7], [9] and [10], have been shown to give good performances of the obtained RT_{60} even in noisy environments. However, the estimation accuracy for the frequency dependent RT_{60} is lower than for the full-band RT_{60} due to the smaller bandwidth of the sub-band filters and thus the lower signal energy.

Frequency dependent RT_{60} can be used to model late reverberation for auralization purposes. The simulated virtual acoustical environment (VAE) should be highly consistent with the acoustics of an actual real room, especially for augmented reality (AR) applications [11]. Therefore, a high estimation accuracy of RT_{60} in the full frequency range is required. In the present study a hybrid method for blind estimation of the frequency dependent RT_{60} has been developed. The maximum likelihood (ML) method [7] is applied to estimate the RT_{60} from low- to mid-frequencies since it has demonstrated a good performance in a noiseless or noisy environment in the detailed benchmarking shown in [12]. The RT_{60} from mid- to high-frequencies is predicted by applying our proposed model, which is based on the analysis of the RT_{60} calculated from six room impulse responses (RIRs) from the Aachen Impulse Response (AIR) database [13].

2. BLIND ESTIMATION OF THE FREQUENCY DEPENDENT REVERBERATION TIME

2.1. Short Overview of the Maximum Likelihood Method

Löllmann and Vary [14] have modeled the sound decay $d(k)$ within a speech pause caused by the room reverberation using a discrete random process:

$$d(k) = A_r\, m(k)\, e^{-\rho\, k/f_s}\, u(k), \tag{1}$$

where A_r denotes the amplitude of the sound decay and $u(k)$ represents the unit step function. f_s is the sampling frequency and $m(k)$ represents a random sequence with zero mean and normal distribution. ρ is the decay rate, which is related to the RT_{60} by:

$$\rho = \frac{3}{\log_{10}(e)\,\mathrm{RT}_{60}}\,. \tag{2}$$

The sound decay $d(k)$ is further represented by a random variable with Gaussian probability density function (PDF) based

on the sound decay model (cf. Eq. 1):

$$P_{d(k)}(x) = \frac{1}{\sqrt{2\pi}\,\xi(k)}\, e^{\left(-\frac{x^2}{2\,\xi^2(k)}\right)}\,, \tag{3}$$

with

$$\xi(k) = A_r\, e^{-\rho\; k/f_s}\; u(k)\,. \tag{4}$$

Then the decay rate ρ can be obtained by finding the maximum estimated ρ of the log-likelihood function (natural logarithm of the likelihood function) from a given sequence $d(k)$ of length N [7, 14]

$$\rho = \underset{\rho}{\operatorname{argmax}} \left\{ -\frac{N}{2}\left(\frac{-\rho\,(N-1)}{f_s} + \quad ... \right.\right. \\ \left.\left. ... \quad + ln\left(\frac{2\pi}{N}\sum_{i=0}^{N-1} d^2(i)\, e^{\frac{2\rho i}{f_s}}\right) + 1\right)\right\}. \tag{5}$$

Finally, the RT_{60} can be calculated based on Eq. 2. However, the estimation accuracy for the RT_{60} in noisy environments is frequency restricted by using the ML method.

2.2. Model of The Reverberation Time from Mid- to High-Frequencies

In the present study, we have calculated and analyzed the frequency dependent RT_{60} in 6 different rooms based on the RIRs from the AIR database [13]. The RIR for each room is filtered through a gammatone filter bank [15] with a bandwidth of one equivalent rectangular bandwidth (ERB) [16], which is widely used for modeling the peripheral filtering in the cochlea. The frequency dependent RT_{60} is then determined using the Schroeder method [2] with a least squares fitting of the energy decay curve (EDC) between -5 and -25 dB for the RIR in each frequency channel $h(f_c, t)$. The equation for calculating the EDC is expressed as:

$$\mathrm{EDC}(f_c, t) = 10\log_{10}\left(\int_t^\infty h^2(f_c, t)\, dt\right) \approx \alpha_{f_c} t + \beta_{f_c}, \tag{6}$$

where $h(f_c, t)$ is the filtered RIR at the center frequency f_c of the gammatone filter bank. Fig. 1 shows the RT_{60} in 6 different rooms (studio booth, office room, meeting room, lecture room, stairway and aula carolina) as a function of the center frequencies of the gammatone filter bank. The calculated RT_{60} (solid black lines) is in accordance with the results shown in [13]. It can be observed that the RT_{60} decreases monotonically from 4 to 20 kHz due to the material and air absorption. This effect can be utilized to build a model for the RT_{60} from mid- to high-frequencies. Therefore, in this study, a 2^{nd}-order polynomial function is applied as a model to predict the RT_{60} from mid- to high-frequencies, which can be expressed as:

$$\widehat{\mathrm{RT}}_{60}(f_c) = a\,(f_c - 4\,\mathrm{kHz})^2 + b\,(f_c - 4\,\mathrm{kHz}) + \mathrm{RT}_{60,4\,\mathrm{kHz}} \\ \text{for } 4\,\mathrm{kHz} \le f_c \le 20\,\mathrm{kHz} \tag{7}$$

where $\widehat{\mathrm{RT}}_{60}(f_c)$ is the modeled reverberation time, $\mathrm{RT}_{60,4\,\mathrm{kHz}}$ is the reverberation time measured at 4 kHz and f_c is the center frequency of the gammatone filter bank. The two model

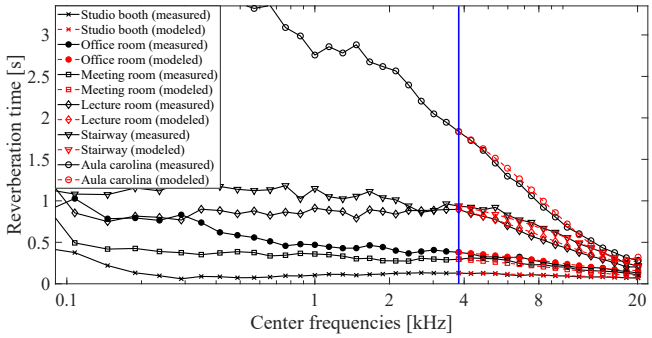

Fig. 1. Frequency dependent RT_{60} calculated by a gammatone filter bank and the Schroeder method (black solid lines). The red dashed lines denote the modeled RT_{60} from 4 to 20 kHz with our proposed method.

parameters a and b for the given six rooms can be estimated by solving the nonlinear least squares optimization problem. For that, we used the trust-region algorithm [17] provided by the MATLAB Curve Fitting Toolbox. The $\widehat{\mathrm{RT}}_{60}(f_c)$ from 4 to 20 kHz is displayed in Fig. 1 (red dashed lines). Note that other models, such as exponential function, etc., can also be used to model the reverberation time from mid- to high-frequencies. In the case of an unknown room, these two parameters can be estimated from the more reliable estimates of the RT_{60} in mid-frequency channels. For this purpose, the averaged reverberation time from 1 to 4 kHz ($\overline{\mathrm{RT}}_{60,1\text{-}4\,\mathrm{kHz}}$) is mapped to the model parameters a and b by a 2^{nd}-order polynomial function, since the RT_{60} from 1 to 4 kHz can be well estimated using the ML method from a recorded speech signal (cf. Fig. 4). The mapping functions can be written as follows:

$$a = c_1\,\overline{\mathrm{RT}}^2_{60,1\text{-}4\,\mathrm{kHz}} + d_1\,\overline{\mathrm{RT}}_{60,1\text{-}4\,\mathrm{kHz}} + e_1\,, \tag{8}$$

$$b = c_2\,\overline{\mathrm{RT}}^2_{60,1\text{-}4\,\mathrm{kHz}} + d_2\,\overline{\mathrm{RT}}_{60,1\text{-}4\,\mathrm{kHz}} + e_2\,. \tag{9}$$

The parameters c_1, c_2, d_1, d_2, e_1 and e_2 are fit to the estimated model parameters a and b, and the $\overline{\mathrm{RT}}_{60,1\text{-}4\,\mathrm{kHz}}$ for these six different rooms also by using the trust-region algorithm.

2.3. Hybrid Method for Blind Estimation of Frequency Dependent Reverberation Time

Fig. 2 shows the block diagram of our proposed algorithm for blind estimation of the RT_{60} from 100 Hz to 20 kHz. The recorded speech signal is first filtered through a gammatone filter bank with a bandwidth of one ERB from 100 Hz to 4 kHz, and the RT_{60} in each frequency channel is blindly estimated by the ML method (cf. sec. 2.1). In the case of a noisy environment, the accuracy of the blindly estimated RT_{60} might be reduced. Therefore, a median filter is applied as a smoothing filter for the estimated RT_{60} from 1 to 4 kHz. Then the averaged reverberation time from 1 to 4 kHz ($\overline{\mathrm{RT}}_{60,1\text{-}4\,\mathrm{kHz}}$) is calculated to obtain the model parameters a and b according to Eqs. 8 and 9. After that, the RT_{60} from mid- to high-frequencies (4-20 kHz) is predicted based on the estimated

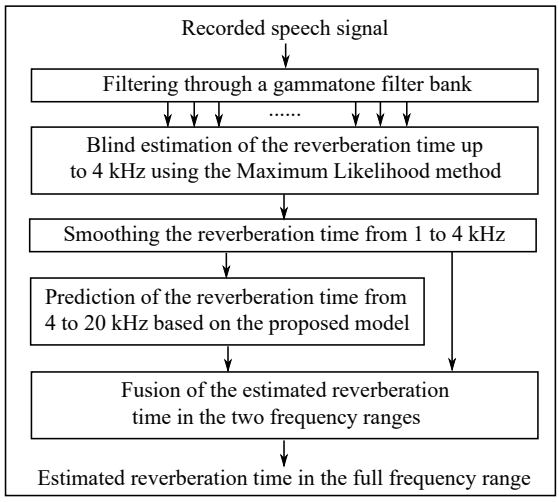

Fig. 2. Structure of the method for blind estimation of the frequency dependent reverberation time.

reverberation time at 4 kHz ($\mathrm{RT}_{60,4\,\mathrm{kHz}}$) and the obtained model parameters a and b by using Eq. 7. Finally, the RT_{60} from 100 Hz to 20 kHz can be determined.

3. PERFORMANCE EVALUATION

Two experiments were performed to evaluate (a) the accuracy of the prediction model for the RT_{60} from mid- to high-frequencies, and (b) the performance of the proposed method for the estimation of the RT_{60} from low- to high-frequencies.

3.1. Performance Evaluation of the Prediction Model from Mid- to High-Frequency Reverberation Time

The first experiment is to evaluate the performance of the model for predicting the RT_{60} from mid- to high-frequencies using four sets of RIRs from the Open AIR Library [18] measured in the Cripta di San Sebastiano (Room A), Chiesa di San Biagio (Room B), York Guildhall Council Chamber (Room C), and Stairway in the University of York (Room D). The frequency dependent RT_{60} are calculated by the Schroeder Method (Eq. 6) and the averaged RT_{60} from 1 to 4 kHz ($\overline{\mathrm{RT}}_{60,1\text{-}4\,\mathrm{kHz}}$) is used to predict the reverberation time from mid- to high-frequencies. Fig. 3 shows the comparison between the actual RT_{60} (black solid line) and the predicted RT_{60} (red dashed line) from 4 to 20 kHz for these four rooms. It is clear to see that the predicted RT_{60} is well matched with the actual RT_{60} over frequencies. To quantify the estimation accuracy of this prediction model, the average of estimation errors (AE) of the predicted RT_{60} over frequencies is calculated for each room according to:

$$\mathrm{AE} = \frac{\sum_{i=1}^{N_{f_c}} |\widehat{\mathrm{RT}}_{60}\,(f_{c,i}) - \mathrm{RT}_{60}\,(f_{c,i})|}{N_{f_c}}\,, \qquad (10)$$

where N_{f_c} denotes the number of center frequencies of the gammatone filter bank from 4 to 20 kHz, $\widehat{\mathrm{RT}}60\ (f_{c,i})$ and $\mathrm{RT}_{60}\ (f_{c,i})$ denote the predicted and actual RT_{60}, respectively. The results in Table 1 illustrate the high accuracy of the predicted RT_{60}. The largest AE is about 0.03 s, suggesting that the proposed model can be well used to predict the RT_{60} from mid- to high-frequencies.

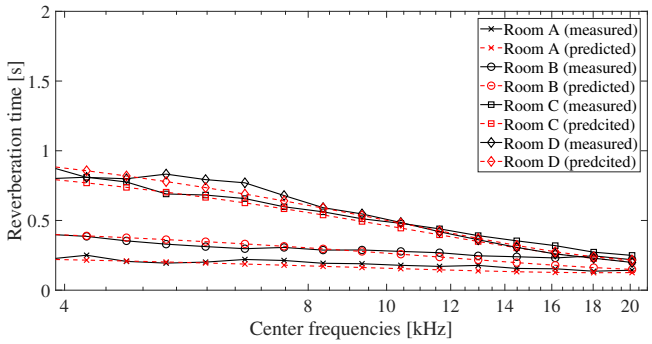

Fig. 3. Comparison between measured and predicted reverberation time from mid- to high-frequencies.

Room	A	B	C	D
AE	0.02 s	0.03 s	0.03 s	0.03 s

Table 1. Averaged estimation error of the predicted reverberation time from mid- to high-frequencies.

3.2. Performance Evaluation of the Hybrid Method for the Estimation of the Frequency Dependent Reverberation Time

In the second experiment, the four RIRs used in the first experiment are applied to evaluate the performance of the hybrid method for estimating the RT_{60} over the full frequency range. An anechoic speech signal from [19] is convolved with these RIRs to simulate reverberant speech signals. Then a white Gaussian noise (WGN) is added to the reverberant speech signals with Signal-to-Noise Ratios (SNRs) of 60, 30, 20 and 10 dB to simulate the additive noise at the microphone. These simulated reverberant speech signals are used to blindly estimate the RT_{60} from 100 Hz to 20 kHz using our proposed method (cf. sec. 2). To simplify the simulation, the frequency distribution of noises in real indoor environments was not considered, i.e., noises should have a strong power in low-frequencies.

Fig. 4 shows the estimated RT_{60} using the ML and our proposed method (hybrid method) from a reverberant speech signal with different SNR levels. The black solid lines represent the reference RT_{60} calculated from the RIRs using the Schroeder method (Eq. 6). It is clearly visible that the RT_{60} can be well estimated up to approx. 2 kHz using the ML method[1] with different SNR levels. However, the accuracy of

[1] An upper limit of the $\widehat{\mathrm{RT}}_{60}(f_{c,i})$ is set to 2.5 s for the ML method

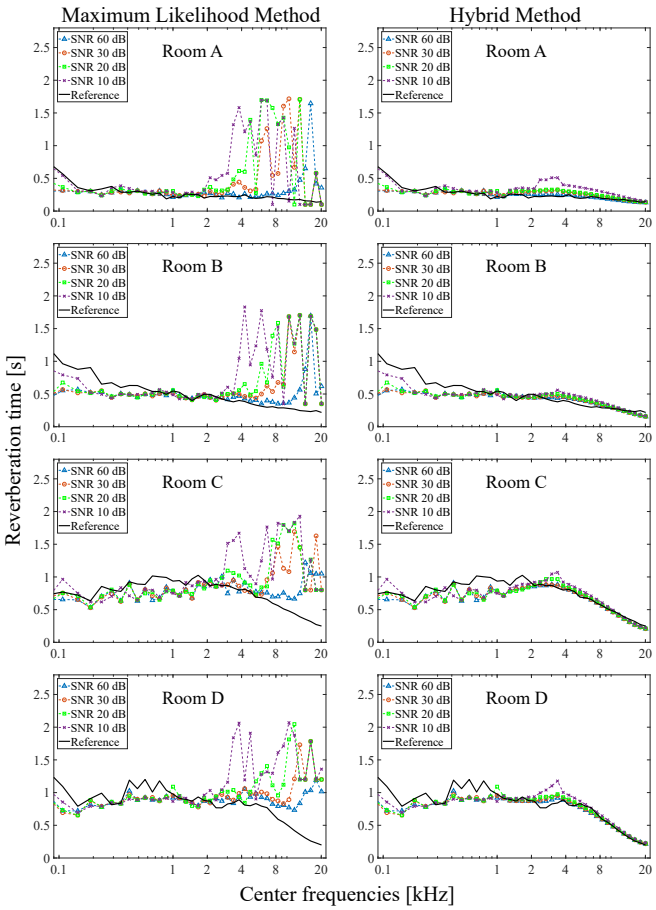

Fig. 4. Blind estimation of reverberation time in full frequency range from a reverberant speech signal with different signal-to-noise ratios (SNRs). The left and right columns show the estimated reverberation time using the Maximum Likelihood method and the hybrid method, respectively.

the blindly estimated RT_{60} from 2 to 8 kHz is reduced by decreasing the SNR levels. The reason for this may be that the speech signal has a lower energy at high frequencies compared to low- or mid-frequencies. The high estimation errors for the RT_{60} in high frequency channels are consistent with the results presented in [20].

For the hybrid method, the RT_{60} is smoothed over frequencies from 1 to 4 kHz to reduce the estimation errors caused by additive noises. Furthermore, the RT_{60} from 4 to 20 kHz is predicted by our proposed model. It should be noted that the smoothing process is only to ensure reliable values fo the estimated RT_{60} between 1 and 4 kHz by using the ML method, since the reverberation time from 1 and 4 kHz is important to predict the reverberation time from mid- to high-frequencies. It can be seen that the estimated RT_{60} is well in agreement with the reference RT_{60} over the frequencies. The AE (Eq. 10) of the estimated RT_{60} over frequencies is calculated for each room and SNR level by using the ML and hybrid method (from 100 Hz to 20 kHz). As shown in Table 2, the hybrid method achieves a good estimation accuracy for all cases (the highest AE is 0.11 s), which was consistent with the visual inspection in Fig. 4. The AE of estimated RT_{60} is clearly lower for the hybrid method in comparison with the ML method, especially for the low SNRs. In the present study, the ML method is used directly to determine the RT_{60} from low- to mid-frequencies. It can be seen that the estimated RT_{60} in low sub-bands (from 100 to 200 Hz) shows some deviations from the reference. These estimation errors can be minimized by using a modified Rayleigh distribution model [21] for the RT_{60} at low-frequencies. Since our study focuses on the prediction of RT_{60} from mid- to high-frequencies, this model is not included in our experiment.

SNR	Room A	Room B	Room C	Room D
		ML method		
10 dB	0.37 s	0.43 s	0.43 s	0.49 s
20 dB	0.34 s	0.35 s	0.34 s	0.37 s
30 dB	0.26 s	0.28 s	0.28 s	0.30 s
60 dB	0.11 s	0.16 s	0.21 s	0.21 s
		Hybrid method		
10 dB	0.09 s	0.09 s	0.10 s	0.11 s
20 dB	0.06 s	0.09 s	0.09 s	0.08 s
30 dB	0.05 s	0.09 s	0.08 s	0.08 s
60 dB	0.04 s	0.08 s	0.08 s	0.07 s

Table 2. Averaged estimation error of the predicted reverberation time using the maximum likelihood method and the hybrid method with different SNRs.

4. CONCLUSIONS

A hybrid method for blind estimation of the frequency dependent RT_{60} is proposed. The ML method is used to determine the RT_{60} up to 4 kHz, and the RT_{60} from 4 to 20 kHz is predicted by applying a model based on the analysis of the RT_{60} calculated from 6 RIRs from the AIR database. In addition, a smoothing filter is used for the RT_{60} between 1 and 4 kHz to reduce the inaccuracies caused by additional noises at the microphone. Two experimental results show the good performance of our proposed method. The blindly estimated RT_{60} from low- to high-frequencies in an unknown room is useful to rapidly adapt virtual 3D-Audio to local environment acoustics [11]. The further work is to use more RIR sets to build a more accurate model. In addition, the modified Rayleigh distribution model will be applied to predict the RT_{60} in low-frequencies.

Acknowledgment

This work is supported by Huawei Innovation Research Program FLAGSHIP (HIRP FLAGSHIP) project.

Literatur

[1] ISO Norm 3382, *Acoustics: measurement of the reverberation time of rooms with reference to other acoustical parameters*, International Organization for Standardization, 1997.

[2] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 2, pp. 409–412, 1965.

[3] W. C. Sabine, *Collected papers on acoustics*, Cambridge, MA: Harvard University Press, 1927.

[4] C. F. Eyring, "Reverberation time in "dead" rooms," *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 217–241, 1930.

[5] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[6] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[7] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[8] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[9] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[10] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[11] J.-M. Jot and K. S. Lee, "Augmented reality headphone environment rendering," in *Proc. AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[12] N. D. Gaubitch, H. W. Loellmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2012.

[13] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE 16th International Conference on Digital Signal Processing*, 2009.

[14] H. W. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[15] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[16] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[17] T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on optimization*, vol. 6, no. 2, pp. 418–445, 1996.

[18] D. T. Murphy and S. Shelley, "Openair: An interactive auralization web resource and database," in *Proc. 129th Audio Engineering Society Convention*, 2010.

[19] P. Kabal, "TSP speech database," Tech. Rep., Department of Electrical and Computer Engineering, McGill University, 2002.

[20] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Acoustic characterization of environments (ACE) challenge results," Tech. Rep., Imperial College London, 2017.

[21] H. W. Löllmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood T60 estimation exploiting subband information," in *Proc. ACE Challenge Workshop, Satellite Event of IEEE-WASPAA*, 2015.