SYSTEM IDENTIFICATION AND DEREVERBERATION OF SPEECH SIGNALS IN THE SINGLE-SIDE-BAND TRANSFORM DOMAIN

Anna Oyzerman and Israel Cohen

Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

ABSTRACT

Single-Side-Band transform (SSB) is an important real-valued time-frequency representation, often preferred in applications involving speech signals. In this paper, the problems of system identification and dereverberation are addressed using the SSB transform. First, an analytical relation between the input and the output signals is derived in the SSB domain. Then, a system identification routine is formulated for a band-to-band approximation of that relation. Second, the dereverberation problem is addressed, using a statistical model for the acoustic impulse response (AIR) function. Exact and approximate representations of the AIR and the reverberant signal are derived directly in the SSB domain. The performance of the dereverberation algorithm is evaluated as a function of the representation complexity. Finally, the SSB and the short-time Fourier transform (STFT) representations are compared for the application of dereverberation.

Index Terms— Dereverberation, Single-Side-Band Transform, System Identification

1. INTRODUCTION

The Single-Side-Band (SSB) transform is an important time-frequency representation. Unlike the short-time Fourier transform (STFT), the SSB representation has real-valued channel signals instead of complex valued signals, and therefore it is often the choice in real-time low-cost applications involving communication, coding systems and speech processing. The SSB can be realized in an efficient manner by sharing computations among channels, employing efficient methods for decimation and interpolation, and by using fast algorithms for modulation and demodulation.

In this work, we employ the SSB transform in two related subjects: system identification and dereverberation. System identification is of major importance in many applications, including acoustic echo cancellation [1], beamforming [2], and dereverberation [3, 4]. As a first step in identification we derive an analytical expression for the impulse response of a linear time invariant (LTI) system in the SSB domain, and propose a possible approximation for that expression. We then present an offline system identification procedure for the approximation using a least squares (LS) criterion and investigate the performance of the identification for different signal-to-noise (SNR) conditions.

The second problem addressed in this work is dereverberation via a spectral enhancement method, that assumes a statistical model for the AIR [5, 6]. Based on one of the statistical models proposed in [7, 8], the algorithm estimates the late reverberant spectral variance (LRSV) component, which is the main contributor to the degradation of the signal. The clean speech signal is then estimated using one of the methods presented in [9–11].

In many existing dereverberation methods, the AIR model is defined in the time domain, and suppression of late reverberation is performed in the STFT domain [5, 6, 12]. Alternatively, defining the AIR in the STFT domain requires to incorporate cross-band filters, in order to achieve a sufficiently accurate representation [13], which complicates the algorithm's implementation. Therefore, we apply a formulation of the AIR model and the reverberated signal directly in the SSB domain, using approximate representations. Then we study how the dereverberation performance depends on the number of cross-bands. Finally, we compare the performance using the SSB transform to the one obtained using the STFT representation.

This paper is organized as follows. Section 2 describes an LTI system representation in the SSB domain. Section 3 addresses the problem of system identification. Section 4 presents the dereverberation in the SSB domain. Experimental results are demonstrated in Section 5.

2. REPRESENTATION OF LTI SYSTEMS IN THE SSB DOMAIN

In this section, we derive an analytical relation between the input and the output signals of an LTI system in the SSB domain. Throughout this paper, unless explicitly noted, the summation indexes range from $-\infty$ to ∞. The SSB repre-

This research was supported by the Israel Science Foundation (grant no. 1130/11).

sentation of a signal $x(n)$ is given by

$$X_{m,k} = Re\left[\sum_n \tilde{\psi}(mM-n)x(n)e^{\frac{j\pi m}{2}}W_K^{-kn}\right] \quad (1)$$

where $\tilde{\psi}$ denotes the analysis window, m the frame index, k the frequency-band index, M the decimation factor and K represents the number of frequency bands used in the transform. W_K is defined as

$$W_K = e^{\frac{2\pi j}{K}}. \quad (2)$$

The inverse SSB transform is given by

$$x(n) = \frac{1}{K}\sum_{k=0}^{K-1}\sum_m Re\left[\psi(mM-n)X_{m,k}e^{-\frac{j\pi m}{2}}W_K^{kn}\right] \quad (3)$$

where ψ denotes the synthesis window. Let $h(n)$ denote an impulse response of an LTI system of length Q. The output signal in the SSB domain is given by

$$Y_{m,k} = Re\left[\sum_n \tilde{\psi}(n-mM)\sum_{l=0}^{Q-1}h(l)x(n-l)e^{\frac{j\pi m}{2}}W_K^{-kn}\right]. \quad (4)$$

After some manipulations $Y_{m,k}$ can be written as

$$Y_{m,k} = \frac{1}{K}\sum_{k'=0}^{K-1}\sum_{m'}H_{m,m',k,k'}X_{m',k'} \quad (5)$$

where

$$H_{m,m',k,k'} = \sum_n \vartheta_{1_{m,k,n}}\sum_{l=0}^{Q-1}h(l)\vartheta_{2_{m',k',n-l}} \quad (6)$$

with

$$\vartheta_{1_{m,k,n}} = \tilde{\psi}(mM-n)\cos\left(\frac{\pi m}{2}-\frac{2\pi kn}{K}\right)$$
$$\vartheta_{2_{m',k',n}} = \psi(n-m'M)\cos\left(\frac{\pi m'}{2}-\frac{2\pi k'n}{K}\right) \quad (7)$$

We refer to $H_{m,m',k,k'}$ for $k = k'$ as a band-to-band filter and for $k \neq k'$ as a cross-band filter. In order to simplify the expression in (6) we propose approximate representations which employ only part or none of the cross-band filters. For an approximation which uses $2K_{max}$ cross-bands, the output signal is given by

$$Y_{m,k} = \frac{1}{K}\sum_{k'=k-K_{max}}^{k+K_{max}}\sum_{m'}H_{m,m',k,k'}X_{m',k'}. \quad (8)$$

For $K_{max} = 0$ the approximate representation uses only the band-to-band filter.

3. SYSTEM IDENTIFICATION IN THE SSB DOMAIN

In this section, we consider system identification in the SSB domain using the band-to-band approximation and an LS optimization criterion. The input signal $x(n)$ passes through an unknown system characterized by its impulse response $h(n)$, resulting in the desired signal $d(n)$. Together with the background white noise $v(n)$, the output signal is given by

$$y(n) = d(n) + v(n) = h(n) * x(n) + v(n). \quad (9)$$

From (9) and (5), the SSB representation of $y(n)$ may be written as

$$Y_{m,k} = D_{m,k}+V_{m,k} = \frac{1}{K}\sum_{k'=0}^{K-1}\sum_{m'}H_{m,m',k,k'}X_{m',k'}+V_{m,k} \quad (10)$$

where $V_{m,k}$ is the SSB transform of $v(n)$.

Let us define N_{xh} as the number of time samples of the filter $H_{m,m',k,k'}$, with the index m. Similarly, N_x is defined as the number of cross-time samples of that filter, with the index m'.

Let $\mathbf{H}^{\mathbf{bb}}_{m,k}$ be the band-to-band filter for time sample m and frequency band k:

$$\mathbf{H}^{\mathbf{bb}}_{m,k} = \begin{bmatrix} H^{\mathbf{bb}}_{m,0,k} & H^{\mathbf{bb}}_{m,1,k} & \cdots & H^{\mathbf{bb}}_{m,N_x-1,k}\end{bmatrix}^T \quad (11)$$

and let $\mathbf{H}^{\mathbf{bb}}_k$ denote a column-stack concatenation of the above band-to-band filter $\left\{\mathbf{H}^{\mathbf{bb}}_{m,k}\right\}_{m=0}^{N_{xh}-1}$ for all the time samples m:

$$\mathbf{H}^{\mathbf{bb}}_k = \begin{bmatrix} \mathbf{H}^{\mathbf{bb}}_{0,k}{}^T & \mathbf{H}^{\mathbf{bb}}_{1,k}{}^T & \cdots & \cdots & \mathbf{H}^{\mathbf{bb}}_{N_{xh}-1,k}{}^T\end{bmatrix}^T. \quad (12)$$

The dimensions of $\mathbf{H}^{\mathbf{bb}}_k$ are $N_{xh} \times N_x$. Let $\mathbf{X}_k$ be the signal X at band k and let

$$\mathbf{\Delta}_k = \begin{bmatrix} \mathbf{X}_k & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{X}_k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & \mathbf{X}_k \end{bmatrix} \quad (13)$$

represent a sparse matrix constructed from the input signal SSB coefficients of the k-th frequency-band, replicated N_{xh} times, where each replication is shifted by N_x columns with respect to the previous line. Now we can write the band-to-band estimate of the desired signal $\mathbf{D}_k$ in a vector form as

$$\mathbf{D}^{\mathbf{bb}}_k = \mathbf{\Delta}_k\mathbf{H}^{\mathbf{bb}}_k. \quad (14)$$

This represents the SSB coefficients of the output signal at the k-th frequency-band, resulting from only the band-to-band filter $\mathbf{H}_k$.

Using the above notations, the LS optimization problem can be expressed as

$$\hat{\mathbf{H}}^{\mathbf{bb}}_k = \arg\min_{\mathbf{H}^{\mathbf{bb}}_k}\left\|\mathbf{Y}_k - \mathbf{\Delta}_k\mathbf{H}^{\mathbf{bb}}_k\right\|^2. \quad (15)$$

The solution to (15) is given by

$$\hat{\mathbf{H}}_k^{\mathbf{bb}} = \left(\boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k\right)^{-1} \boldsymbol{\Delta}_k^H \mathbf{Y}_k \tag{16}$$

where we assumed that $\boldsymbol{\Delta}_k^H \boldsymbol{\Delta}_k$ is not singular (otherwise, some regularization is included). Substituting (16) into (14), we obtain

$$\hat{\mathbf{D}}_k^{\mathbf{bb}} = \boldsymbol{\Delta}_k \hat{\mathbf{H}}_k^{\mathbf{bb}} \tag{17}$$

which is the estimate of the desired signal in the SSB domain at the k-th frequency-band using a band-to-band filter.

3.1. MSE computation

After calculating the estimated signal, we can analyse the mean-squared error (MSE) from two aspects:

1. An estimated error - derived by calculating the MSE between the estimated signal, $\hat{\mathbf{D}}_k^{\mathbf{bb}}$, and the real signal $D_{m,k}$ as defined in (10):

$$\epsilon_{\mathbf{estimate}} = \frac{E\left\{\left\|\mathbf{D}_k - \hat{\mathbf{D}}_k^{\mathbf{bb}}\right\|^2\right\}}{E\left\{\left\|\mathbf{D}_k\right\|^2\right\}}. \tag{18}$$

2. A theoretical error - derived by calculating the MSE between the estimated signal, $\hat{\mathbf{D}}_k$, and the signal $\mathbf{D}_k^{\mathbf{bb}}$ as defined in (14):

$$\epsilon_{\mathbf{theory}} = \frac{E\left\{\left\|\mathbf{D}_k^{\mathbf{bb}} - \hat{\mathbf{D}}_k^{\mathbf{bb}}\right\|^2\right\}}{E\left\{\left\|\mathbf{D}_k^{\mathbf{bb}}\right\|^2\right\}}. \tag{19}$$

4. DEREVERBERATION IN THE SSB DOMAIN

In a reverberant environment, the AIR model in the time domain is given by [6]

$$h(n) = \begin{cases} b_d(n) & \text{if } 0 \le n < T_s \\ b_r(n) e^{-\delta(k)n} & \text{if } n \ge T_s \end{cases} \tag{20}$$

where $\delta(k)$ denotes the decay rate related to the reverberation time, $b_d(n)$ and $b_r(n)$ are zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random variables, and T_s is the time when the early reflections end.

Assuming that the path from the source to the microphone can be treated as an LTI system, and using (5) and (20), we can express the reverberant signal $y(n)$ in the SSB domain as:

$$Y_{m,k} = \frac{1}{K}\sum_{k'=0}^{K-1}\sum_{m'} H_{m,m',k,k'} X_{m',k'} =$$

$$= \begin{cases} \frac{1}{K}\sum\limits_{k'=0}^{K-1}\sum\limits_{m'}\left(\sum\limits_n \vartheta_{1_{m,k,n}} \sum\limits_{l=0}^{Q-1} b_d(l)\vartheta_{2_{m',k',n-l}} X_{m',k'}\right), \\ \qquad\qquad \text{if } 0 \le m < N_e, \\ \frac{1}{K}\sum\limits_{k'=0}^{K-1}\sum\limits_{m'}\left(\sum\limits_n \vartheta_{1_{m,k,n}} \sum\limits_{l=0}^{Q-1} b_r(l)e^{-\delta(k)l}\vartheta_{2_{m',k',n-l}} X_{m',k'}\right), \\ \qquad\qquad \text{if } m \ge N_e. \end{cases} \tag{21}$$

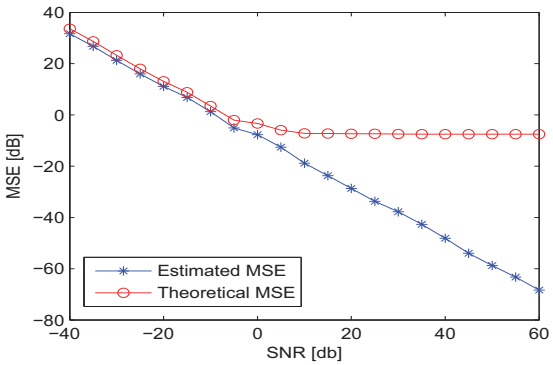

Fig. 1. Theoretical and estimated MSE curves for the band-to-band identification system, as a function of SNR for a white Gaussian noise input signal.

The parameter N_e specifies the portion of the AIR that is considered as late reverberations, and is related to T_s in the time domain.

Assuming that the SSB coefficients of the speech signal can be modelled as zero-mean i.i.d real random variables with a certain distribution and variance $\lambda_x(m,k)$, the expression for the reverberant component as presented in [6] is:

$$\begin{aligned}\lambda_r(m,k) = [1-\kappa(k)]\, e^{-2\delta(k)R}\lambda_r(m-1,k) + \\ +\kappa(k)e^{-2\delta(k)R}\lambda_y(m-1,k)\end{aligned} \tag{22}$$

where $\lambda_y(m,k) = E\left\{|Y(m,k)|^2\right\}$ and $\kappa(k)$ denotes the ratio between the energy of the reverberant and the direct path. The LRSV [5] is then given by

$$\lambda_l(m,k) = e^{-2\delta(k)R(N_e-1)}\lambda_r(m-N_e+1,k). \tag{23}$$

5. EXPERIMENTAL RESULTS

The signals used in the simulations include synthetic white Gaussian noise as well as real speech signals. Throughout this section, the AIR was simulated according to the method proposed in [14], with room dimensions of $6 \times 8 \times 5$ m, and a reverberation time of 500 ms. The SSB was implemented using $K = 32$ frequency bands, Kaiser synthesis window of $4N + 1 = 129$ samples, and the related bi-orthogonal analysis window. The overlap between two successive frames was 50%.

5.1. System Identification

System identification results are shown under the assumption of band-to-band filtering, for SNR conditions ranging from -40 to 60 dB. Both the signal and the noise were white Gaussian noise of 2000 samples. In this subsection the source-microphone distance was 1 m, the length of the AIR was truncated to $Q = 700$, and the sampling rate was 8 kHz.

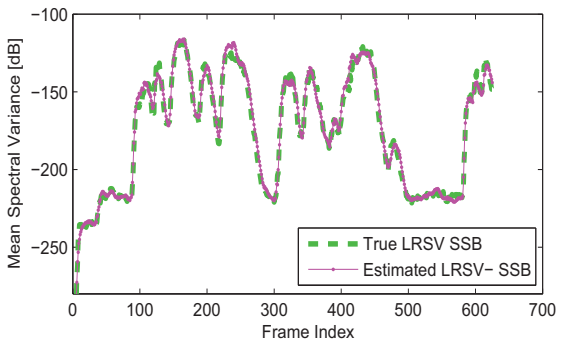

Fig. 2. Mean Spectral Variance of true and estimated LRSVs of speech signal in the SSB Domain.

Figure 1 shows the graph of theoretical and estimated MSE for different SNR conditions. The estimated-MSE, is getting smaller as the SNR increases in spite of the fact that the model neglects all the cross-band filters. On the other hand, the theoretical-MSE remains almost constant after a certain SNR. This is due to the fact that the LS optimization was performed using the real output full-band signal. In other words, the identified model is closer to the representation of the full system, even though it lacks one dimension.

5.2. Dereverberation

In this subsection, we present and discuss results of dereverberation obtained using the SSB representation. The simulated AIR was of length $Q = 4096$ and the source-microphone distance varied between 0.5m and 3m. The parameter T_s was set to 48ms.

For qualitative evaluation of the LRSV estimation we used the mean spectral variance of the LRSV over all the frequency bins, which is given by

$$\text{Mean Spectral Variance [dB]} = 10\log\left(\text{mean}_k\left\{\lambda_l(m,k)\right\}\right) \quad (24)$$

The Mean Spectral Variance of the estimated LRSV was compared to the "true" LRSV, known from the AIR simulation [14]. The quantitative evaluation of the LRSV estimator was determined by the Log Spectral Distance measure. The dereverberation performance was evaluated using the mean segmental Signal to Reverberation Ratio (SRR) and the mean Log Spectral Distortion (LSD). Figure 2 shows the resulting true and estimated mean LRSVs of speech signals in the SSB domain, for a source-microphone distance of 1.3 m.

Figure 3 shows the dereverberation evaluation curves for a speech signal as a function of source microphone distance for the SSB and STFT representations. Clearly, the performance using the STFT representation is higher, which implies that real-valued representations are less suitable for dereverberation. This is associated with the fact that real-valued representations combine the phase information into the amplitude representation. Consequently, in estimating the LRSV we have to use a larger smoothing factor to compensate for multiple reflections with different delays, and this degrades the performance.

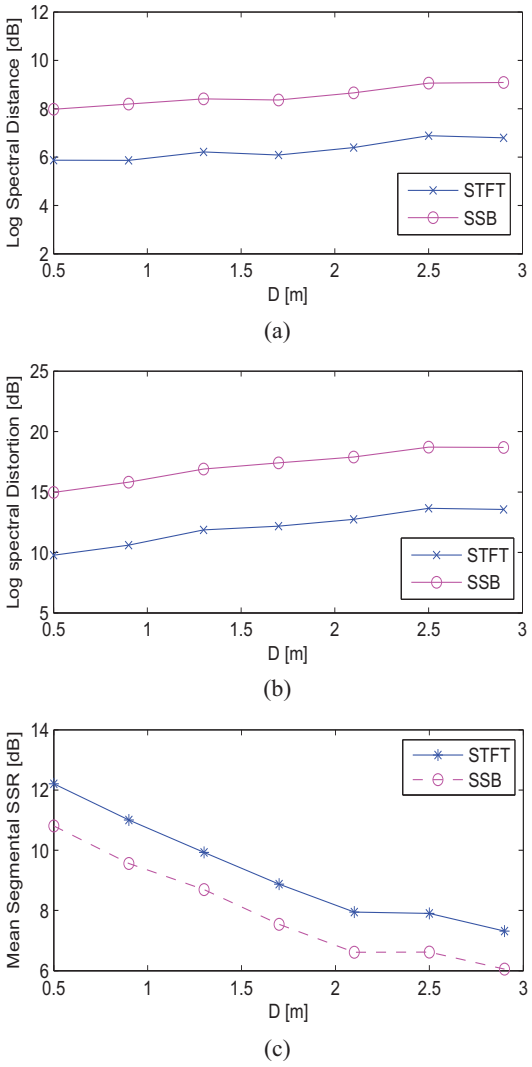

Fig. 3. Dereverberation evaluation in the SSB domain in comparison to the STFT domain. (a) Log Spectral Distance; (b) Mean LSD; (c) Mean SRR.

5.3. Cross-band analysis

Here, we analyse the dereverberation performance when using an increasing number of cross-bands such that $0 \leq K_{max} \leq 15$. The input signal is white Gaussian noise of 2000 samples. The sampling rate is 4 kHz, and the length of the AIR is 1000 taps.

As can be seen from Figure 4, unlike the STFT case [13], the contribution of the cross-band filters is distributed almost equally along all the cross bands. Nevertheless, as was shown in the system identification procedure, the band-to-band rep-

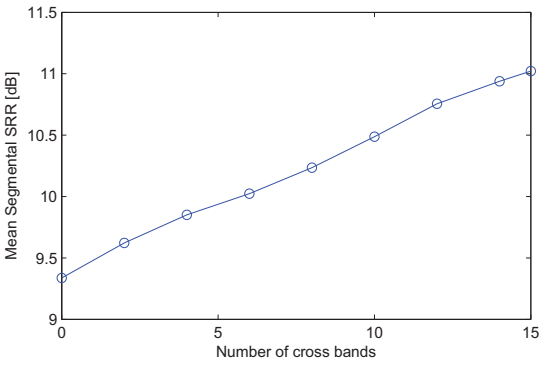

Fig. 4. Mean SRR of the derverberation in the SSB Domain using various numbers of cross-bands.

resentation sufficiently describes the system and thus yields satisfying results with a low computational complexity.

6. CONCLUSIONS

We have investigated the SSB transform as a time-frequency domain representation for speech signal processing. First, we developed a formulation of LTI systems in the SSB domain. Then we proposed system identification using a band-to-band filter approximation. We showed that as SNR improves, the identified band-to-band system becomes closer to the real system, even though it lacks the cross-band dimension. This implies that the band-to-band approximation can sufficiently describe the system. Hence, band-to-band approximation in the SSB domain is suitable, e.g., for acoustic echo cancellation.

We also investigated the performance of dereverberation in the SSB transform domain, compared to dereverberation in the STFT domain. The evaluation measures show that the STFT enables better results due to the fact it separates the spectral magnitude and phase representations, and thus facilitates the LRSV estimation. Finally, we examined the relationship between the AIR model complexity and the dereverberation performance, and showed that although the band-to-band representation gives sufficient results, each additional cross-band contributes to further improvement.

7. REFERENCES

[1] J. Benesty, T. Gnsler, D. R. Morgan, M. M. Sondhi, and Gay S. L., *Advances in Network and Acoustic Echo Cancellation*, Springer, New York, 2001.

[2] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[3] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept. 2005.

[4] Mingyang Wu and DeLiang Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[5] 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.

[6] E. A. P. Habets, *Single and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, 2007.

[7] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[8] J.D. Polack, *La transmission de lenergie sonore dans les salles*, Ph.D. thesis, 1988.

[9] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[11] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, pp. 113–116, Apr. 2002.

[12] E.A.P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sept. 2009.

[13] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[14] E.A.P. Habets, "Room impulse response (RIR) generator," May 2008.