HEAD-ORIENTATION COMPENSATION WITH VIDEO-INFORMED SINGLE CHANNEL SPEECH ENHANCEMENT

Soumitro Chakrabarty, Deepth Pilakeezhu and Emanuël A. P. Habets

International Audio Laboratories Erlangen*
Am Wolfsmantel 33, 91058 Erlangen, Germany
{soumitro.chakrabarty, emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

It has been shown that human speakers do not radiate voice sound uniformly in all directions and that the radiation pattern is frequency dependent. As a consequence, the quality of the speech signal acquired by distant microphones depends on the relative orientation of the head with respect to the microphone. In this paper, a single channel speech enhancement framework is proposed that incorporates the head orientation information to compensate for the reduction in sound energy due to the relative orientation of the speaker with respect to the microphone, while attenuating the noise. In the proposed framework, the head orientation at each time instance, which can potentially be estimated using computer vision techniques, is used to compute the frequency dependent gain factor that needs to be applied to compensate for the head orientation. The computed gain is then incorporated in a single channel filter which simultaneously suppresses the noise. Based on experimental evaluations, with both simulated and measured data, we demonstrate the ability of the proposed system to improve the quality of the acquired speech signal.

Index Terms— head-orientation, single channel, speech enhancement, video-informed

1. INTRODUCTION

Different researches [1–3] have shown that a human speaker radiates more sound energy in the forward direction than towards the sides or rear direction. Also, at the rear of the speaker high frequency components suffer a higher attenuation compared to low speech frequencies, therefore implying the frequency dependence of the radiation pattern. Due to this non-uniformity, the quality of the speech signals acquired by distant microphones is influenced by the relative orientation of the head of the speaker, in addition to being degraded by acoustic noise and reverberation. As a result, the performance of speech based applications such teleconferencing, speaker diarisation, speech based human computer interfaces etc. are also affected by the head orientation of the speaker.

Generally, the knowledge about the head-orientation of a speaker is unavailable and needs to be estimated. This problem has been predominantly tackled using computer vision techniques. A detailed survey of video-data based head-orientation estimation techniques can be found in [4]. However, in recent years, several methods have been proposed to estimate the head orientation of a speaker based on multichannel speech signals [5, 6]. The audio data based methods generally require a large number of microphones deployed in a distributed manner to obtain an accurate estimate of the head-orientation. Other methods, such as [7, 8], employ a multimodal approach by using both audio and visual cues.

In speech related literature, the head-orientation information is generally incorporated within an acoustic source localization technique to account for the degrading effect of the head-orientation and develop a robust source localization method [9, 10]. Some works (c.f. [11, 12]) have tried to incorporate the head orientation information within a speech recognition system to improve its performance. However, to the best knowledge of the authors, no previous work explores the significance of the head orientation information for speech enhancement and its incorporation into a speech enhancement framework to improve the quality of the acquired speech signals.

We propose a single channel speech enhancement method that incorporates the head-orientation information to compensate for the attenuation of sound energy due to relative orientation of the speaker with respect to the microphone. In the proposed system, we opt for a video-data based method to estimate the head-orientation of the speaker at each time instant. The current orientation estimate is then used to compute the appropriate frequency dependent gain factor. In this paper, a single channel minimum mean square error (MMSE) filter is proposed to perform the speech enhancement. The derived filter allows us to perform noise reduction as well as compensate for the reduction of sound energy due to the head-orientation of the speaker. Through experimental evaluation, with both simulated and measured data, we demonstrate the ability of the system to improve the quality of the acquired signal when the head orientation of the speaker is known, as well as provide motivation for incorporating head-orientation information within speech enhancement methods.

2. PROBLEM FORMULATION

Let us consider a simple setup with an omnidirectional microphone and the head of the speaker placed at fixed locations, as shown in Fig. 1. Then, the received signal at the microphone at time-frequency instant (n, k), where n is the time index and k is the frequency index, $Y(n, k)$ is given by

$$Y(n, k) = H(\theta, k)S(n, k) + V(n, k), \tag{1}$$

where $S(n, k)$ is the speaker signal, and $V(n, k)$ represents the spatially white noise component. The acoustic transfer function (ATF) between the microphone and the speaker, with head-orientation θ, is given by $H(\theta, k)$.

As an illustration of the dependence of the ATF on the head-orientation as well as the frequency, the energy of the ATF between the mouth and a microphone as a function of microphone position, for a simulated scenario in an anechoic environment, at frequencies

[*] A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

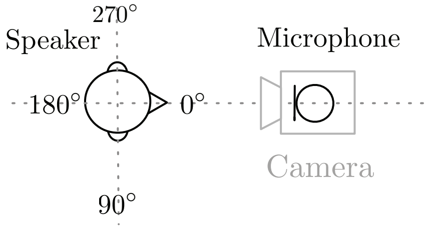


Fig. 1. Illustrative diagram for a simple speaker-microphone setup along with the camera position.

of 100 Hz and 3 kHz is shown in Fig. 2. In this work, we only consider the propagation of sound waves in the horizontal xy plane. It can be observed that at 100 Hz the radiation pattern is omnidirectional with the head having little effect. However, at 3 kHz, the effect of scattering due to the head becomes more significant, as the energy at the sides and the back of the head is reduced.

We assume $\theta = 0^\circ$, when the speaker is directly facing the microphone (as illustrated in Fig. 1). With this assumption, the received microphone signal can be reformulated as

$$Y(n,k) = A(\theta,k)X(n,k) + V(n,k), \tag{2}$$

where $X(n,k)$ is the signal proportional to the sound pressure at the microphone for $\theta = 0^\circ$, i.e., $X(n,k) = H(0,k)S(n,k)$. The attenuation of the speaker signal due to the head-orientation of the speaker is represented by the relative attenuation factor $A(\theta,k)$, given by $A(\theta,k) = H(\theta,k)/H(0,k)$. Assuming the signal and noise components in (2) to be mutually uncorrelated, the expected power of the microphone signal $\phi_Y(n,k) = \mathrm{E}\{|Y(n,k)|^2\}$ is given by

$$\phi_Y(n,k) = |A(\theta,k)|^2\phi_X(n,k) + \phi_V(n,k), \tag{3}$$

where the expected power of the speaker signal for $\theta = 0^\circ$, $\phi_X(n,k)$, and the expected power of the noise, $\phi_V(n,k)$, are defined similarly.

In this work, the aim is to compensate for the reduction of sound energy due to head-orientation of the speaker as well as attenuate the noise. Then, the desired signal is given by $X(n,k)$.

3. PROPOSED FRAMEWORK

In this section we present the proposed framework to obtain an estimate of the desired signal.

3.1. Orientation compensation filter

An estimate of the desired signal can be given by

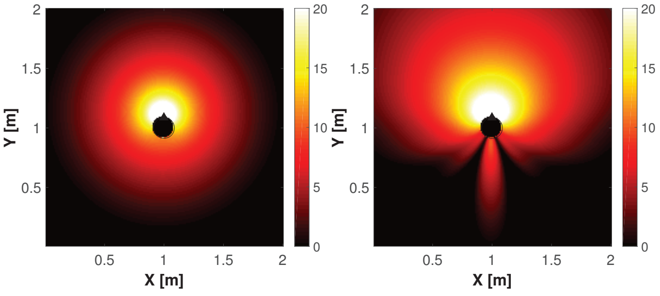
$$\widehat{X}(n,k) = W(\theta,k)Y(n,k), \tag{4}$$

where $W(\theta,k)$ is the head-orientation dependent weighting factor. Using the MMSE criterion [13], the weighting factor can be found by minimizing the mean square error between the desired and the estimated signal, i.e.,

$$W(\theta,k) = \underset{W}{\arg\min}\ \mathrm{E}\{|WY(n,k) - X(n,k)|^2\}. \tag{5}$$

Considering the signal model in Sec. 2, and setting the derivative of the cost function to be minimized to zero, the solution is given by

$$W(\theta,k) = \frac{|A(\theta,k)|\phi_X}{|A(\theta,k)|^2\phi_X + \phi_V}, \tag{6}$$



(a) 100 Hz

(b) 3 kHz

Fig. 2. Sound energy radiation pattern (dB) for (a) 100 Hz and (b) 3 kHz. The mouth position is denoted by a triangle.

where ϕ_X and ϕ_V denote the power spectral density (PSD) of the desired signal $X(n,k)$ and the noise $V(n,k)$, respectively. To simplify and make the obtained solution more intuitive, we introduce the orientation dependent gain $G(\theta,k)$, which is defined as $G(\theta,k) = |A(\theta,k)|^{-1}$. Substituting the defined gain, $G(\theta,k)$, in (6), the solution can be rewritten as

$$W(\theta,k) = G(\theta,k) \cdot \frac{\phi_X}{\phi_X + G^2(\theta,k)\phi_V}. \tag{7}$$

The filter given in (7) can be interpreted as a two-stage filter, where in the first stage the attenuation in sound energy due the head-orientation of the speaker is compensated by the gain $G(\theta,k)$, followed by a noise reduction filter that suppresses the noise in the compensated signal. It should be noted that a simple application of the head-orientation dependent gain to the input signal, without the noise reduction filter, can result in noise amplification at the output. To compute the filter given by (7), several parameters need to be estimated, namely, the noise power ϕ_V, the desired signal power ϕ_X and the orientation dependent gain $G(\theta,k)$. It should be noted that to compute the gain $G(\theta,k)$, an estimate of the head orientation θ is required. The estimation of these parameters is presented in the following.

3.2. Head orientation estimation

Since we consider sound propagation in the horizontal plane, we are only interested in the head orientation known as *yaw* [4]. The orientation can be estimated using both audio and visual data.

Using audio based methods, to obtain an accurate estimate of the head orientation, distributed microphone arrays are generally required. Also, audio related methods are susceptible to degradation in performance in reverberant and noisy acoustic environments. Video based methods have been more popular for tackling this problem.

In our framework, we opt for video-based estimation of the head-orientation of the speaker. The choice of video-based head-orientation estimation allows us to utilize spatial information about the sound scene in the framework while using a single microphone for speech acquisition. A proprietary software known as Sophisticated High-speed Object Recognition Engine (SHORE$^{\text{TM}}$) [14], developed at Fraunhofer IIS, is used in this paper to estimate the head-orientation. SHORE$^{\text{TM}}$ is mainly built to detect multiple facial features for face detection and facial analysis. The head-orientation estimation is a built-in feature which is used to get the estimates. The estimator provides a single orientation angle estimate for each time frame n.

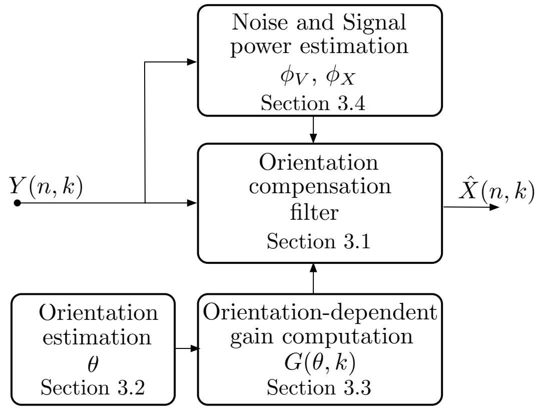


Fig. 3. Block diagram of the proposed system.

3.3. Orientation dependent gain computation

Given the head-orientation estimate θ, the next task is to compute the orientation dependent gain $G(\theta, k)$. At each time frame n, based on the current estimate of θ, the frequency dependent gain is selected from a pre-computed gain table.

To compute the gain table, we use the spherical microphone array impulse response generator (SMIRgen) [15] as a mouth simulator. The head of the speaker is modeled as a rigid sphere of radius r_h, with the mouth as an omnidirectional point source on this sphere. Then, by sampling the complete orientation angle space ($[0^\circ, 360^\circ]$, see Fig. 1) at I discrete points, we compute the ATF for each of these points on the sphere, using SMIRgen, assuming the speaker and microphone positions to be known. The orientation dependent gain for each θ_i $\forall\, i \in \{1,\ldots,I\}$, is then given by

$$G(\theta_i, k) = \left| \frac{\widehat{H}(\theta_i, k)}{\widehat{H}(0, k)} \right|^{-1}, \tag{8}$$

where $\widehat{H}(0,k)$ and $\widehat{H}(\theta_i,k)$ are the simulated ATFs for the orientation angle of 0° and θ_i, respectively. The computation of the gain table presented in (8) follows from the definition of the orientation dependent gain $G(\theta,k) = |A(\theta,k)|^{-1}$. It can be seen that the precomputed gain table is a matrix of size $I \times K$, where K is the total number of frequency bins. The gain $G(\theta,k)$ applied in the filter (7) is selected from this gain table as the element corresponding to the current estimate of $\theta \approx \theta_i$ and frequency bin k.

Based on the amount of information available regarding the room, with SMIRgen, it is possible to compute the gain table for both anechoic and reverberant environments. The gain table can also be computed from measured ATFs for a specific room with known speaker-microphone locations. As an example, a computed gain table with SMIRgen considering an anechoic environment is shown in Fig. 4. It can be observed that in the range of $[90^\circ, 270^\circ]$, i.e., when the speaker is facing away from the microphone, a gain of more than 3 dB is required to compensate for the orientation. As expected, higher gain is required at higher frequencies.

3.4. Noise and desired signal power estimation

There exist several methods for single channel noise power estimation [16]. In this work, we assume that the noise power ϕ_V is stationary and is estimated during speech absence. To estimate the power of the desired signal, ϕ_X, we use the decision directed approach presented in [13].

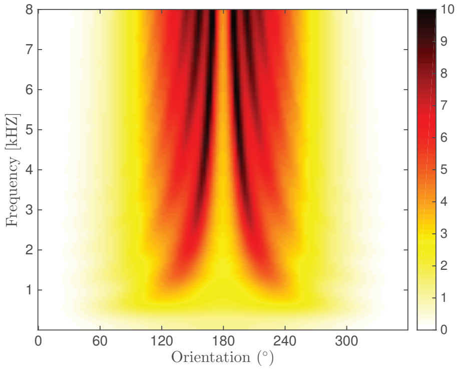


Fig. 4. Gain table computed with SMIRgen for an anechoic environment.

A block diagram of the complete framework is presented in Fig. 3. It can be seen that the complete system consists of a single channel orientation compensation filter that provides an estimate of the desired signal based on the estimate of the orientation dependent gain, and the noise and desired signal powers.

4. PERFORMANCE ANALYSIS

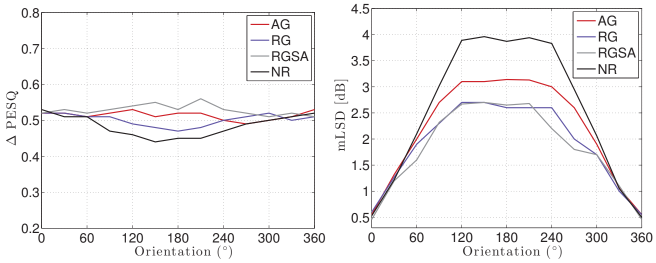
To analyse the ability of the proposed system to improve the quality of the acquired speech signal by incorporating the head orientation information, we present the results with both simulated and measured room impulse responses (RIRs). The signal quality at the output of the proposed filter is evaluated in terms of two objective measures: perceptual evaluation of speech quality (PESQ) score improvement [17] and mean log spectral distance (mLSD) [18]. The PESQ score improvement, denoted by ΔPESQ, is computed as the difference of the PESQ score of the inverse STFT of the estimate of the desired signal $\widehat{X}(n,k)$ and the PESQ score of the inverse STFT of the input signal $Y(n,k)$. The mLSD is obtained by averaging the LSD between the output signal $\widehat{X}(n,k)$ and the desired signal $X(n,k)$ over all active speech frames.

With both simulated and measured RIRs, the proposed filter was applied to input signals obtained by convolving a dry speech signal of 7 s duration with RIRs for 13 different head orientations, distributed uniformly over the whole orientation angle space, i.e., $[0^\circ, 360^\circ]$. White noise was added to the input signals to obtain the noisy microphone signals $Y(n,k)$, such that segmental signal-to-noise ratio for the 0° orientation signal was 20 dB. For the experiments, both the source and the microphone were placed at the same height. The sampling frequency for all experiments was 16 kHz and the STFT frame length was 1024 samples with 50% overlap. For the experimental results presented here, we assume that the head orientation of the speaker is known for all cases.

4.1. Simulation experiment

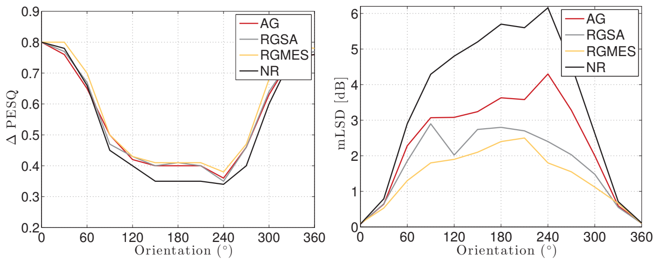
For the simulation experiments, we considered a room with dimensions $4.55\,\mathrm{m} \times 4.45\,\mathrm{m} \times 2.55\,\mathrm{m}$, with a reverberation time of $T_{60} = 0.17$ s. With the head modeled a sphere and the mouth as a point source on it, the RIRs for the 13 different head orientations were simulated using SMIRgen. The distance between the source and the microphone was kept constant at 1.5 m.

The experimental results for the simulated scenario is presented in Fig. 5. The experimental results are presented for three differ-



(a) PESQ (b) mLSD

Fig. 5. Experimental results with simulated RIRs.



(a) PESQ (b) mLSD

Fig. 6. Experimental results with measured RIRs.

ent gain table computations as well as with application of the proposed filter without any orientation related gain. First, the gain table was computed assuming an anechoic environment (denoted by AG in Fig. 5). Second, we considered the knowledge of the reverberation time and the room dimensions to be available, and computed the gain table for the known reverberant environment (denoted by RG). Third, we computed the gain table, for the known reverberant environment, with spatial averaging (denoted by RGSA), to compensate for the position changes of the speaker and microphone. For the spatial averaging, we placed the speaker and the microphone at different sampled points in the room while maintaining the speaker-microphone distance of 1.5 m, and the gain table was obtained by averaging over all the gain tables obtained for the different locations. The spatial averaging was performed since it was found that the gain table computed for the known reverberant environment (RG), did not posses the same smooth characteristic as the anechoic gain table (AG), shown in Fig. 4. Finally, to demonstrate the improvement in speech quality due to the incorporation of head orientation information, we also present the results when the proposed filter given by (7) is applied with the $G(\theta, k) = 1$ (denoted by NR). Please note that this results in a simple MMSE noise reduction filter.

From the presented results in Fig. 5, it can be seen that a better overall improvement in speech quality is obtained when the head-orientation dependent gain is incorporated. The improvement is more prominent when the head-orientation lies in the range of $[90°, 270°]$, i.e., when the speaker is facing away from the microphone. It can also be seen that the overall best improvement, both in terms of ΔPESQ and mLSD, is obtained when the spatially averaged gain table (RGSA) is used.

4.2. Experiment with measured RIRs

For this case, the RIRs for the 13 different head orientations were measured, where the signal was emitted from a KEMAR head and torso simulator, placed on a turntable. The distance between the source and the microphone was kept constant at 1 m. The measurements were done in an acoustic lab at Fraunhofer IIS, with the same room dimensions and T_{60} as above.

The experimental results with the measured RIRs is presented in Fig. 6. For this case, the results are presented for three of the same variants of the proposed filter as above, namely, AG, RGSA and NR. In addition, we also provide the results when the gain table is computed using the measured ATFs (denoted by RGMES).

From the results presented in Fig. 6, it can be seen that with measured RIRs there is a greater reduction in ΔPESQ for all applied filters as the speaker starts to face away from the microphone. However, when the speaker is facing away from the microphone, the incorporation of the head orientation gain leads to a better PESQ score improvement, similar to the results with simulated RIRs. It can be seen that the incorporation of the orientation dependent gain leads to a significantly better performance in terms of mLSD. The best overall improvement of speech quality is obtained when the gain table computed with measured ATFs (RGMES) is used, however, it can be seen that the results with the spatially averaged simulated gain table (RGSA) are comparable to RGMES, for both ΔPESQ and mLSD.

4.3. With estimated head orientations

The current implementation of our video-based head orientation method only provides head-orientation estimates for the frontal view of the speaker with a single camera, i.e. we do not obtain estimates when the orientation lies in the range of $[90°, 270°]$. Using a webcam, placed at the same position as the microphone (as shown in Fig.1), the performance of the complete system (Fig. 3), was analyzed with measured RIRs. The results were found to be similar to the ones presented in the previous section. Since the head orientation estimation was restricted up to $90°$, and the orientation dependent gains in this range have a small variation, there was no significant influence of estimation errors. An extension of our orientation estimation method to cover the complete head orientation range and a detailed analysis of the influence of head orientation estimation errors are topics for future research.

5. CONCLUSION AND FUTURE WORK

A single channel speech enhancement framework that incorporates the head orientation information to improve the quality of the acquired speech signal is presented. Through experimental results with both simulated and measured RIRs, it was shown that the incorporation of head orientation information in a speech enhancement method can help in improving the speech quality, thereby providing motivation for further exploring the significance of head orientation information for speech enhancement.

Though we showed improvement in speech quality with orientation information integrated into the system, the results also demonstrated the need for a better method to obtain significant improvement in performance compared to simple noise reduction. Another avenue for future work would be to gradually relax the constraints of the presented system and develop a modified method to improve the speech quality in a more practical setting.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the help and guidance they received from the Intelligent Systems group at Fraunhofer IIS and in particular Jens Garbas, regarding the SHORE^{TM} library.

7. REFERENCES

[1] H. K. Dunn and D. W. Farnsworth, "Exploration of pressure field around the human head during speech," *Journal Acoust. Soc. of America*, vol. 10, no. 1, pp. 83–83, 1938.

[2] W. T. Chu and A. C. C. Warnock, *Detailed directivity of sound fields around human talkers*, Institute for Research in Construction, Ottawa, 2002.

[3] G. A. Studebaker, "Directivity of the human vocal source in the horizontal plane.," *Ear and Hearing*, vol. 6, no. 6, 1985.

[4] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, April 2009.

[5] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches for head orientation estimation in a smart room," in *European Conference on Speech Communication and Technology*, August 2007, pp. 590–593.

[6] C. Segura, C. Canton-Ferrer, A. Abad, J.R Casas, and J. Hernando, "Multimodal head orientation towards attention tracking in smart rooms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007, pp. 681–684.

[7] C. Canton-Ferrer, C. Segura, J.R Casas, M. Pards, and J. Hernando, "Audiovisual head orientation with particle filtering in multisensor scenarios," *Journal on Advances in Signal Processing*, vol. 2008, pp. 12, September 2008.

[8] J. Odobez and O. Lanz, "Sampling techniques for audio-visual tracking and head pose estimation," in *Multimodal Signal Processing: Human Interactions in Meetings*, chapter 6, pp. 84–102. Cambridge University Press, June 2012.

[9] B Mungamuru and P Aarabi, "Enhanced Sound Localization," in *IEEE Trans. Syst., Man, Cybern. B*, June 2004, vol. 34, pp. 1526–1540.

[10] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, Germany, 2001.

[11] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Role of head pose estimation in speech acquisition from distant microphones," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 3557–3560.

[12] A. Sasou, "Head-orientation-estimation-integrated speech recognition for the smart-chair," in *Second International Symposium on Universal Communication,*, Dec 2008, pp. 482–489.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[14] T. Ruf, A. Ernst, and C. Küblbeck, *Microelectronic Systems: Circuits, Systems and Applications*, chapter Face Detection with the Sophisticated High-speed Object Recognition Engine (SHORE), pp. 243–252, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[15] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *Journal Acoust. Soc. of America*, vol. 132, pp. 1462, 2012.

[16] R.C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Synthesis lectures on speech and audio processing. Morgan & Claypool, 2013.

[17] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 749–752.

[18] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.