INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Coherence-based dual-channel noise reduction algorithm in a complex noisy environment

Youna Ji, Jun Byun, and Young-cheol Park

Computer and Telecomm. Eng. Division
Yonsei University, Wonju, Korea
jyn282@yonsei.ac.kr

Abstract

In this paper, a coherence-based noise reduction algorithm is proposed for a dual-channel speech enhancement system operating in a complex noise environment. The spatial coherence between two omnidirectional microphones is one of the crucial information for the dual-channel speech enhancement system. In this paper, we introduce a new model of coherence function for the complex noise environment in which a target speech coexists with a coherent interference and diffuse noise around. From the coherence model, three numerical methods of computing the normalized signal to interference plus diffuse noise ratio (SINR), which is related to the Wiener filter gain, are derived. Objective parameters measured from the enhanced speech demonstrate superior performance of the proposed algorithm in terms of speech quality and intelligibility, over the conventional coherence-based noise reduction algorithm.

Index Terms: speech enhancement, noise reduction, diffuse noise, interference noise, spatial coherence

1. Introduction

Speech quality and intelligibility can be severely degraded in noisy or highly reverberant environments. In such condition, the noise reduction and dereverberation techniques become an important factor for building a successful speech-related applications. Since the dual-channel speech enhancement system utilizes spatial information as well as input spectra, it is possible to expect better noise reduction performance than mono systems.

The spatial coherence is one of the crucial information in many dual-channel noise reduction algorithms. In [1], a real-valued noise coherence model was introduced in order to design a post-filter of a microphone array in a diffuse noise environment. It was shown that a more accurate estimate of the Wiener filter could be obtained by using the real-valued noise coherence function. More recently, a technique of estimating the SNR of the input signal using real and imaginary parts of a complex spatial coherence function was suggested [2, 3]. In [2], a coherence-based noise reduction technique was proposed in a situation that a frontal target was present together with an undesired interference. It achieved a significant improvement of speech intelligibility and quality over the conventional coherence-based algorithm.

However, the previous algorithm in [2] consider only one type of noise; directional interference without reverberation. In practical environments, listeners often experience more complex acoustic noise field, such as the presence of directional interferences together with diffusive and reverberant ambient noise. While the interference is highly correlated over the two microphone signals, the diffuse noise shows little coherence over the microphone signals except for low frequencies. Thus the performance of the conventional algorithms in [1, 2] will be degraded in the complex noise where a target coexist with a coherent interference and diffusive ambient noise.

In [3], a hybrid coherence function was introduced to simultaneously model both the coherent and reverberant noises. Using the hybrid coherence model, better performance than the previous method in [2] was obtained under a reverberant environment. Still, since ambient noise surrounding listener was not considered, the performance is likely to degrade as the level of ambient noise increases. In [4], we proposed *a priori* speech absence probability (SAP) estimator in a complex acoustic noise field in which a frontal target exists together with a directional interference and diffuse noise. Since the signal model in [4] considers both the coherent and diffusive types of noises, it exhibits superior performance under such complex noise environment than the previous algorithms.

In this paper, we propose a dual-channel noise reduction algorithm based on a new coherence model for the complex noise environment. Starting with the new coherence model, we develop three different methods of calculating the noise reduction gain. Through computer simulations, the performance of the proposed algorithm is evaluated, and it is compared with that of the previous algorithm.

2. Signal modeling for a complex noise environment

We assume that noisy input signals are captured by two omnidirectional microphones with an interval d. In an environment with complex noise, the dual channel input signals can be represented in the frequency domain as

$$Y_i(k,l) = S_i(k,l) + V_i(k,l) + N_i(k,l), i = 1, 2, \quad (1)$$

where $V_i(k,l)$ is a directional interference and $N_i(k,l)$ is the diffuse noise signal. The interference is a dominant directional noise, that is, the dominant noise incident from a specific direction. Thus, it is highly correlated with the two microphones signals. On the other hand, diffuse noise is uncorrelated with each other except at low frequency [1, 5].

The coherence between the two observation signals can be calculated as

$$\Gamma_Y(k,l) = \frac{\Phi_{YY}^{12}(k,l)}{\sqrt{\Phi_{YY}^{11}(k,l)\Phi_{YY}^{22}(k,l)}}, \quad (2)$$

where $\Phi_{YY}^{ij}(k,l) = E\{Y_i(k,l)Y_j^*(k,l)\}, i,j = 1, 2$ are cross- and auto-PSDs of the microphone signals. For the algorithm development, we assume that, $E\{S_i(k,l)V_i^*(k,l)\} = 0$, and $E\{S_i(k,l)N_i^*(k,l)\} = 0$. For sake of simplicity, we will omit the frequency and frame indices whenever necessary.

http://dx.doi.org/10.21437/Interspeech.2017-1464

2.1. Coherence modeling for complex noise environment

As shown in [6], the coherence in the complex noise environment can be represented as an weighted sum of the directional signal and the diffuse noise coherences:

$$\Gamma_Y(k,l) = \Gamma_D(k,l) \cdot \frac{DDR}{DDR+1} + \Gamma_N(k,l) \cdot \frac{1}{DDR+1}, \quad (3)$$

where $\Gamma_D(k,l)$ and $\Gamma_N(k,l)$ are the coherences of the directional signals and the diffuse noise, respectively. DDR represent the true local direct to diffuse ratio (DDR) of i-th channel microphone signal in a linear scale, that is, $DDR = (\Phi_{SS}^{ii} + \Phi_{VV}^{ii})/\Phi_{NN}^{ii}$. According to the result in [2], the DDR of first and second microphones are nearly identical if the distance between two is close enough, so any of the DDR values of the two microphones can be used.

Meanwhile, the input signal in (1) contains two directional signal components: target speech and interference. Thus, the coherence of the directional signal components can be represented as

$$\Gamma_D(k,l) = \Gamma_S(k,l) \cdot \frac{SIR}{1+SIR} + \Gamma_V(k,l) \cdot \frac{1}{SIR+1}, \quad (4)$$

where $\Gamma_S(k,l)$ and $\Gamma_V(k,l)$ are the coherences of the target speech and the directional interference, respectively. The $SIR = \Phi_{SS}^{ii}/\Phi_{VV}^{ii}$ represents the target speech to the interference power ratio (SIR) at the i-channel microphone. In [2], it was shown that the ratio of SIR has almost the same value in both channels as in the case of DDR.

By substituting (4) into (3), the coherence of the noisy observation can be rewritten as

$$\Gamma_Y = \Gamma_S G + \Gamma_V(K-G) + \Gamma_N(1-K), \quad (5)$$

where $K = DDR/(DDR+1)$ is the normalized DDR and G represents normalized signal to interference plus diffuse noise ratio (SINR) which is bounded as $0 \leq G \leq 1$ and related to the Wiener filter noise reduction gain [7]:

$$\begin{aligned} G &= SIR/(SIR+1) \cdot DDR/(DDR+1) \\ &= \frac{SINR}{SINR+1}. \end{aligned} \quad (6)$$

The spatial coherence between the signals from two omnidirectional microphones in the diffuse noise field is often modeled as a real-valued analytic function [1], as given by

$$\hat{\Gamma}_N(k) = sinc\left(\frac{2\pi k f_s d}{N \cdot c}\right), \quad (7)$$

where d is the microphone spacing, N is the maximum frequency bin index, f_s and $c \approx 340m/s$ represent the sampling frequency and the speed of sound, respectively. On the other hand, the target speech and interference are assumed to be generated from a single well-defined directional sound source, and thus the signals received by the two microphones are perfectly coherent except for a time delay. Thus, the coherence function of the directional signal, $\hat{\Gamma}_C$, can be expressed as [8, 9],

$$\hat{\Gamma}_C(k) = e^{j2\pi k f_s (d/(N\cdot c))\sin\theta}, \quad (8)$$

where θ is the angle of incidence. In a complex noise environment where the target speech and directional interference can occur in any direction around the listener. If we further assume that the target direction is known a priori, the noisy coherence function in (5) can be rewritten as

$$\begin{aligned} \Gamma_Y &= (\cos\alpha + j\sin\alpha)G + \\ &\quad (\cos\beta + j\sin\beta)(K-G) + \hat{\Gamma}_N(1-K), \end{aligned} \quad (9)$$

where $\alpha = 2\pi k f_s (d/(N\cdot c))\sin\theta_s$, $\beta = 2\pi k f_s (d/(N \cdot c))\sin\theta_i$ and θ_s and θ_i are the angle of target and interference respectively.

The DDR or so called coherent to diffuse ratio (CDR) estimation from the measured coherence between two omnidirectional microphones have been widely investigated [10, 11, 12]. In [10], the heuristically motivated DDR estimator was introduced. The DDR estimate is computed using coherence of observations and the analytic model of diffuse noise field, as given by

$$DDR = \frac{|\hat{\Gamma}_N|^2 - |\Gamma_Y|^2}{|\Gamma_Y|^2 - 1}. \quad (10)$$

Using (10), DDR can be computed without any prior knowledge such as target direction and noise PSD. In this paper, we utilize (10) to compute the normalized DDR, K.

3. Proposed noise reduction methods

In this section, three different methods of estimating the denoising gain in complex noisy environment are developed. All algorithms are based on the coherence model in (5) and (9), and it is assumed that the target direction is known a priori.

3.1. Method 1

The first method utilizes the real and imaginary parts of the observation as introduced in [3]. The real and imaginary part of the observation coherence (9) can be represented as:

$$\begin{aligned} \Re_Y &= (\cos\alpha - \cos\beta)G + \cos\beta K + \Gamma_N(1-K), \\ \Im_Y &= (\sin\alpha - \sin\beta)G + \sin\beta K. \end{aligned} \quad (11)$$

After a few steps of rearrangements, we can combine the real and imaginary terms into a single equation,

$$G = \frac{\Re_Y - \cos\beta K - \Gamma_N(1-K)}{\cos\alpha - \cos\beta} = \frac{\Im_Y - \sin\beta K}{\sin\alpha - \sin\beta}. \quad (12)$$

Since K can be computed using (10), G is obtained only if $\sin\beta$ or $\cos\beta$ is available. To compute these two values, the method in the previous studies [2, 3, 4] can be used, and they will be presented at the end of this section.

3.2. Method 2

Since the directional target speech signal is fully coherent, normalized SINR function G can be derived from the fact that $|\Gamma_S| = 1$. Then, (5) can be rewritten as

$$|\Gamma_s| = |(\Gamma_Y - \Gamma_V(K-G) - \Gamma_N(1-K)) \cdot \frac{1}{G}| = 1. \quad (13)$$

After squaring both side of equation and rearranging the terms for G, the following result can be obtained:

$$G = \frac{-|\Gamma_Y - \Gamma_V K - \Gamma_N(1-K)|^2}{2\Re\{\Gamma_V(\Gamma_Y - \Gamma_V K - \Gamma_N(1-K))^*\}}, \quad (14)$$

where $*$ and $\Re$ denote conjugate and real part respectively. As like Method 1, knowledge of interference coherence is required.

3.3. Method 3

The last method utilizes the magnitude squared coherence (MSC) of the noisy observation. Since the MSC can be obtained by sum of squared real and imaginary part, it is formulated using (9) as

$$|\Gamma_Y|^2 = (A'G + B')^2 + (C'G + D')^2, \tag{15}$$

where

$$\begin{aligned} A' &= \cos\alpha - \cos\beta \\ B' &= \cos\beta + \Gamma_N(1-K) \\ C' &= \sin\alpha - \sin\beta \\ D' &= \sin\beta K. \end{aligned} \tag{16}$$

After rearranging the terms in (15), we can get two G according to the sign of root by solving the quadratic equation.

As noted in our previous study in [4], the normalized SINR can be either positive or negative root, depending on the dominant power of the signal. In a region where the power of the target speech is dominant, positive root represents the normalized SINR. On the other hand, in a region where the power of the interference is dominant, negative root tracks the normalized SINR. To account for this power dependency, we selectively choose between two roots based on the angle difference. Since we assume that diffuse noise has real-valued coherence, imaginary part of Γ_Y is only affected by directional target and interference. It also tends to approach the imaginary part of the signal with dominant power in the time-frequency bin. Therefore, when the imaginary part difference between observation and target speech is larger than the difference between observation and interference, the target speech is considered to be dominant in the corresponding time-frequency bin, so we take the plus sign and vice versa. Then G can be computed as

$$G = \frac{-(A'B' + C'D') + \gamma \cdot \sqrt{T'}}{A'^2 + C'^2}, \tag{17}$$

where

$$\gamma = \begin{cases} 1 & \text{if } |\Im_Y - \sin\alpha| < |\Im_Y - \sin\beta| \\ -1 & \text{otherwise ,} \end{cases} \tag{18}$$

and $T' = |\Gamma_Y|^2(A'^2 + C'^2) - (A'D' - B'C')^2$.

3.4. Estimates of interference coherence

All of the above de-noising gain calculations require interference coherence information. To compute the interference coherence, we can use the methods introduced in [2, 4]. First, (12) can be rewritten as,

$$A\cos\beta = B\sin\beta + C \tag{19}$$

where

$$\begin{aligned} A &= \sin\alpha K - \Im_Y \\ B &= \cos\alpha K - \Re_Y + \Gamma_N(1-K) \\ C &= (\Re_Y - \Gamma_N(1-K))\sin\alpha - \Im_Y\cos\alpha \end{aligned} \tag{20}$$

After squaring both sides of (19), and solving the quadratic equation using $\cos^2\beta + \sin^2\beta = 1$, two solutions can be obtained according to the sign of root. Through mathematical and experimental analysis, we found that the negative root of the quadratic equation solution tracks the correct answer. Thus $\sin\beta$ is given as:

$$\sin\beta = \frac{-BC - AT}{A^2 + B^2}, \tag{21}$$

where $T = K - \cos\alpha(\Re_Y - \Gamma_N(1-K)) - \Im_Y \sin\alpha$.

Similarly to $\sin\beta$, the real part of coherence, $\cos\beta$, is also obtained:

$$\cos\beta = \frac{AC - BT}{A^2 + B^2}. \tag{22}$$

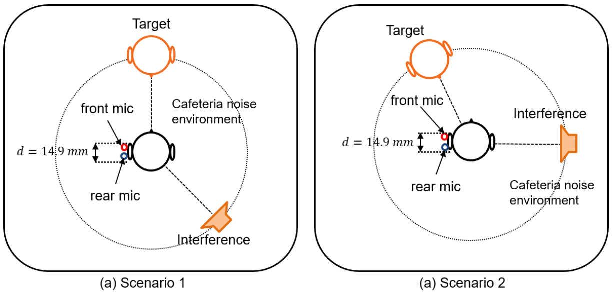
Also similar as $\sin\beta$, negative root is taken to yield correct solution. Now, interference coherence is readily obtained by substituting (21) and (22) into (8).

4. Simulations and Results

Through computer simulations, we analyzed the performance of the three proposed methods and compared it with that of the conventional coherence-based noise reduction algorithm in [3].

10 speech sentences from TIMIT databases were extracted for the target signals and speech-like AR random processing for the interferences were binaurally convolved with HRIR. Binaural Behind-The-Ear Impulse Response (BTE-IR) measured in a cafeteria from [13] have been used to generate dual-channel input signals. The room reverberation time was 1250 ms. For computer simulations, the front and rear microphones signal of the left channel were selected from database. The cafeteria babble noise recorded in the same environment was added to the target and interference signals according to the SINR. The sampling frequency was set to 16 kHz, and noisy input signals were segmented into subframes of 512 samples with a 50% overlap using a sine window. We used the 1st-order recursive averaging to estimate the PSD and set the smoothing factor to 0.8 for all algorithms. The target direction was assumed to be known a priori, and equal minimum gain floor, $G_{min} = 0.1$, was applied for all algorithms. The simulated spatial configurations including locations of the microphones are shown in Fig.1 and details of the source position can be found in [13].

First, to assess the adequacy of the noisy coherence model in (5), the magnitude difference between the true and estimated coherences was computed according to the DDR for the scenario 1 in Fig.1 (a). In this simulation, the recorded cafeteria noise and modeled ambient noise were used as diffuse noise and the magnitude error was computed and averaged over the frequency band. The results are shown in Fig.2. As expected, the proposed model shows lower error (Fig.2 (a)) than the model in [3] for both noise types. The difference is greater in low DDR environments. The snapshots of the true and estimated noisy



(a) Scenario 1 (a) Scenario 2

Figure 1: *The simulated spatial configurations*

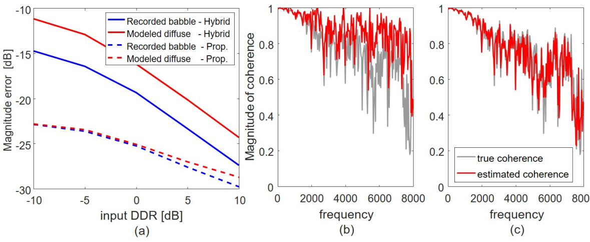


Figure 2: *(a) Magnitude error of DDR, and comparison of true and estimated coherence obtained using (b) the method in [3] and (c) the proposed (5) at 0 dB DDR*

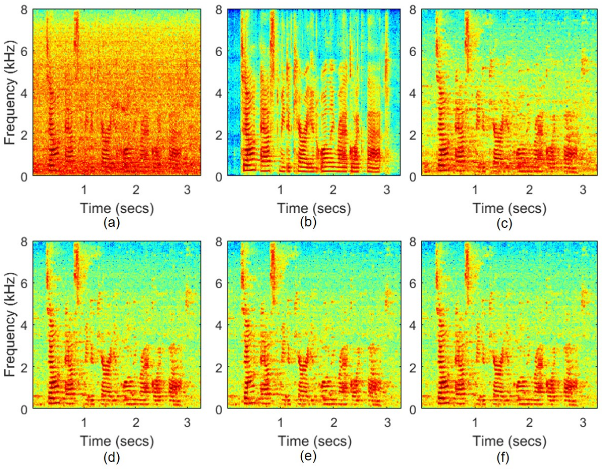


Figure 3: *The spectra of (a) noisy and (b) clean speech signals; enhanced speech output obtained using (c) the Hybrid method [3], (d) the Method 1, (e) the Method 2 and (f) the Method 3*

coherences are compared in Fig.2 (b) and (c). Superiority of the proposed method in the noisy coherence model is clearly visible.

Spectrograms of the noisy input and clean signals are shown in Fig. 3 (a) and (b), respectively. Also the spectrograms of enhance output obtained using the hybrid, and three proposed methods are shown in Fig. 3 (c)-(f). Both SIR and DDR of the noisy input signal were set to 5 dB. It can be seen from the spectrograms that the conventional method (c) contains more residual noise than the proposed methods (d)-(f) especially at low frequencies.

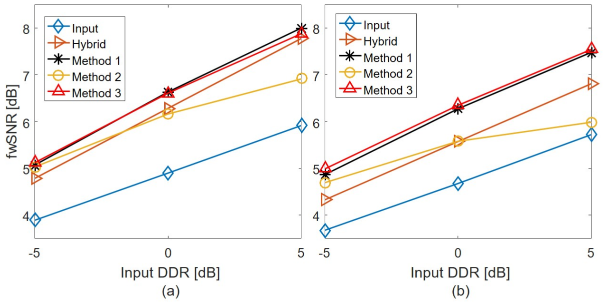


Figure 4: *fwSNR results in (a) Scenario 1 and (b) Scenario 2*

Finally, some objective parameters were measured; the frequency weighted SNR (fwSNR), a short-time objective intelligibility measure (STOI), and the perceptual evaluation of speech quality (PESQ) [14, 15]. It should be noted that the fwSNR is optimized for signals sampled at 8 kHz. Therefore, the signals were downsampled from 16 kHz to 8 kHz before measuring fwSNR. However, STOI and PESQ were measured at 16 kHz. Graphical comparisons of the results are shown in Figs. 4-6. First, Fig. 4 shows fwSNR results, under the two noise scenarios shown in Fig. 1. It can be noted that the three proposed methods yield higher fwSNR than the conventional method for most of cases, except for Method 2 at 5 dB DDR. Fig. 6 shows the STOI results. The STOI has a value between 0 and 1, and the closer to 1, the higher speech intelligibility. The Method 3 shows the highest score in both PESQ and STOI, followed by Method 1.

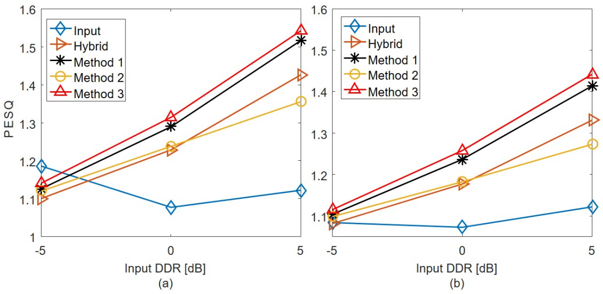


Figure 5: *PESQ results in (a) Scenario 1 and (b) Scenario 2*

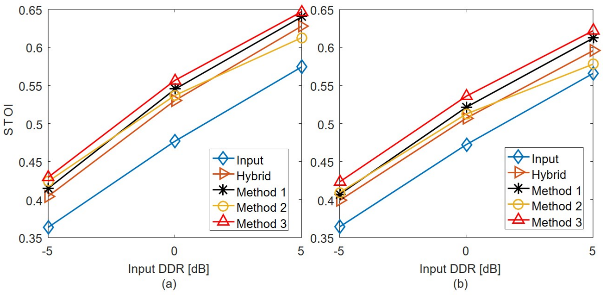


Figure 6: *STOI results in (a) Scenario 1 and (b) Scenario 2*

5. Conclusions

In this paper, we proposed a novel dual-channel noise reduction algorithm for a complex noise environment where a target speech is present together with a coherent interference as well as diffusive ambient noise including reverberation. Utilizing a new coherence model, three different methods of calculating the de-noising gain were developed. Through computer simulations, it was shown that the proposed algorithm could effectively suppress both the interference as well as the diffusive ambient noise. Objective parameters also demonstrated that the proposed algorithm has superior performance in terms of speech quality and intelligibility, over the conventional coherence-based noise reduction algorithm. Among the three proposed algorithms, the MSC-based method achieved the best result.

6. Acknowledgements

This research was supported by the Ministry of Trade, Industry & Energy(MOTIE), Korea Institute for Advancement of Technology(KIAT) through the Encouragement Program for The Industries of Economic Cooperation Region.

7. References

[1] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 709 – 716, nov. 2003.

[2] N. Yousefian and P. C. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 599–609, 2012.

[3] N. Yousefian, J. H. Hansen, and P. C. Loizou, "A hybrid coherence model for noise reduction in reverberant environments," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 279–282, 2015.

[4] Y. Ji and Y.-c. Park, "Improved a priori sap estimator in complex noisy environment for dual channel microphone system," *Interspeech 2016*, pp. 2567–2571, 2016.

[5] H. Abutalebi, H. Sheikhzadeh, R. Brennan, and G. Freeman, "A hybrid subband adaptive system for speech enhancement in diffuse noise fields," *Signal Processing Letters, IEEE*, vol. 11, no. 1, pp. 44 – 47, jan. 2004.

[6] Y. Ji, Y. Baek, and Y.-C. Park, "A priori SAP estimator based on the magnitude square coherence for dual-channel microphone system," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4415–4419.

[7] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[8] N. Yousefian, P. C. Loizou, and J. H. Hansen, "A coherence-based noise reduction algorithm for binaural hearing aids," *Speech Communication*, vol. 58, pp. 101–110, 2014.

[9] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[10] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 1347–1351.

[11] ——, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 1347–1351.

[12] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 6, pp. 1006–1018, 2015.

[13] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 6, 2009.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.