NOISE POWER SPECTRAL DENSITY ESTIMATION FOR PUBLIC ADDRESS SYSTEMS IN NOISY REVERBERANT ENVIRONMENTS

Neda Faraji *

Amirkabir University of Technology, EED
424,Hafez Ave., Tehran, Iran
Email:nfaraji@aut.ac.ir

Richard. C. Hendriks

Delft University of Technology, SIPlab,
2628 CD Delft, The Netherlands
Email: r.c.hendriks@tudelft.nl

ABSTRACT

In many public areas, public address systems play a central role in broadcasting important messages. Consider e.g., airports or train stations. The intelligibility of messages broadcast by such public address systems gets reduced due to environmental noise sources. Usually, these systems therefore employ a pre-processing module to modify the original message such that its intelligibility in the noisy environment is increased. Most of the pre-processing methods for intelligibility improvement require an estimate of the environmental noise PSD. However, existing noise PSD estimators overestimate the noise PSD in reverberant environments, leading to an overamplification of the speech by the speech intelligibility pre-processing module. In this paper, we propose a new MMSE-based noise PSD estimator that can be utilized in public address systems considering reverberation.

Index Terms— Noise PSD tracking, Public address system, Reverberation

1. INTRODUCTION

Nowadays, public address (PA) systems are involved in many daily situations like train stations, airports, school auditoriums and many other large public spaces. In this work we consider a PA system as depicted in Fig. 1. A PA system constitutes of an amplifier and loudspeakers to broadcast a prerecorded message or live speech throughout the public area. It is of crucial importance that the broadcast message is well understood, i.e., has high intelligibility. Unfortunately, environmental noise sources, which are also perceived beside the announcement, decrease the intelligibility of the played message at the receiver side. As the noise signals are part of the environment and can not be removed, the only practical way to increase the intelligibility is to manipulate the clean speech in a pre-processing module before playing.

Most intelligibility improvement algorithms, e.g., [1] [2] [3], depend on the noise PSD, which is assumed to be available. However, in most applications there is no or little a priori information on the noise sources. Assumming a microphone is located somewhere in the vicinity of the human receiver, the signal recorded by the microphone consists of the environmental noise, the processed announcement and its reverberation. Hence, a noise PSD estimation algorithm is necessary to extract the noise PSD from the noisy reverberated signal. Some recent works for speech intelligibility improvement relevant for this application, e.g. [4] and [5], use a minimum mean-squared error (MMSE) based noise PSD estimator [6], but do not take reverberation into account. In this approach, a bias-compensated weighting factor, that is a function of the a priori and a posteriori SNRs, is applied to the magnitude-squared noisy DFT coefficients. The resulting noise PSD estimate is optimal in MMSE sense. Directly using this approach in a reverberated environment, might lead to leakage of the reverberation into the noise PSD estimate. This will erroneous influence the intelligibility processing performed by methods like [1-4]. Because of the low complexity as well as the good performance of the approach presented in [6], we propose in this paper a modification, such that it can be applied for speech intelligibility improvement of public address systems in reverberant environments.

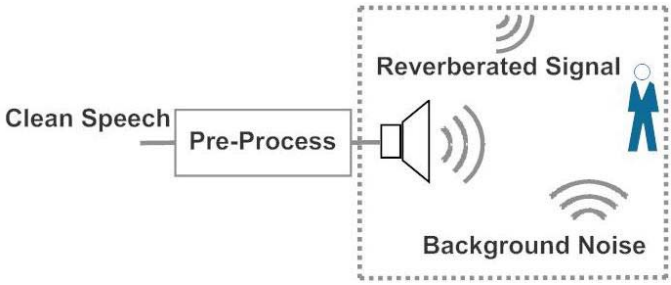

Fig. 1. An outline of a public address system.

The remainder of this paper is organized as follows. First, in Sec. 2 the notation and assumptions used in the baseline MMSE-based noise PSD estimator are given. Then, we derive a new noise PSD estimator considering reverberation in Sec. 3. The experimental results and conclusion are explained in

*The work has been performed during a research visit to Delft University of Technology, financially supported by the Iranian Telecommunication Research Center (ITRC).

Sec. 4 and 5, respectively.

2. NOTATIONS AND ASSUMPTIONS IN MMSE-BASED NOISE PSD ESTIMATION

We briefly summarize in this section the main results on MMSE based noise PSD estimation from [6]. In the sequel, time-sample, time-frame and frequency-bin indices are denoted respectively by l, i and k. In [6], it is assumed that the observed data x can be presented by the following additive model:

$$x(l) = s(l) + d(l), \tag{1}$$

in which $s(l)$ is the clean speech and $d(l)$ denotes the noise signal. In frame-based processing, for frame index i we can write (1) in the Fourier domain as:

$$X(i,k) = S(i,k) + D(i,k). \tag{2}$$

Hereafter, the frame and frequency indices are removed for notational simplicity, unless stated otherwise. The goal of the noise PSD estimator in [6] is to estimate the noise PSD in MMSE sense, i.e., by

$$E\{|D|^2|X\} \tag{3}$$

Assuming a complex-Gaussian distribution for both S and D, it follows that [6]

$$E\{|D|^2|X\} = \left(\frac{1}{(1+\xi)^2} + \frac{\xi}{(1+\xi)\zeta}\right)|X|^2, \tag{4}$$

where σ_D^2 and σ_S^2 represent the variances of the corresponding DFT coefficients D and S, respectively, and $\xi = \frac{\sigma_S^2}{\sigma_D^2}$ and $\zeta = \frac{|X|^2}{\sigma_D^2}$ are a priori and a posteriori SNRs, respectively.

To compute (4), it was proposed in [6] to use the following maximum likelihood (ML) estimator for ξ

$$\xi_{ML} = \max\left(\frac{X^2(i,k)}{\hat{\sigma}_D^2(i-1,k)} - 1, 0\right). \tag{5}$$

Utilizing this estimate leads to a bias in the noise PSD estimator. Hence, a bias compensation factor B has been derived in [6] to compensate for this, that is,

$$B^{-1}(\xi) = \left((1+\xi)\gamma\left(2, \frac{1}{1+\xi}\right) + e^{\frac{-1}{1+\xi}}\right) \tag{6}$$

in which $\gamma(\mu_1, \mu_2)$ is the incomplete Gamma function. As it is clear from (6), the bias compensation factor depends on the a priori SNR ξ as well. The more accurate ξ is, the more accurate the bias compensation factor and the estimated noise will be. In [6], a Decision-Directed a priori SNR estimate [7], i.e. ξ_{DD}, has been used to compute (6). Finally, the noise PSD is estimated as $\tilde{\sigma}_D^2 = E\{|D|^2|x; \hat{\xi}_{ML}\}B(\hat{\xi}_{DD})$.

3. PROPOSED METHOD

3.1. Observation model in a PA system

Consider a PA system in which an amplified message s_{Amp} is being announced. The amplified clean speech signal, s_{Amp}, travels via a direct path from the source (speaker) to the receiver (in this case a human listener with a microphone close by), which is modeled by the attenuation factor α. Moreover, the reflections from the closed environment introduce some reverberation besides the direct-path signal. The reverberated signal is the convolution of the playing amplified clean signal s_{Amp} with the Room Impulse Response (RIR) denoted by h. Finally, the signal recorded by the microphone can be represented as

$$x(l) = g(l) * s_{Amp}(l) + d(l), \tag{7}$$

where $g(l)$ models both the reverberation and the direct-path and is defined by the following equation:

$$g(l) = \begin{cases} \alpha & l = 0 \\ h(l-1) & l \geq 1, \end{cases} \tag{8}$$

where h is the room impulse response excluding the direct path. Here, we use Polack's statistical model [8] in which a specific RIR is generated as one realization of the following stochastic process

$$h(l) = b(l) \times e^{-\eta l} \text{ for } l \geq 0, \tag{9}$$

where $b(l)$ is a zero-mean Normal stochastic process with variance ν^2, which defines the fine structure of the RIR modulated with an exponential function with decay rate η. The decay rate is defined as $\eta = \frac{3 \times ln(10)}{T_r \times f_s}$ in which T_r and f_s are reverberation time and sampling frequency, respectively.

3.2. Derivations

As the clean speech s_{Amp} is available and the attenuation factor α could be simply determined considering the speaker-microphone distance, without loss of generality we can then rewrite the observed signal from (7) using (8) as

$$\begin{aligned} y(l) &= g(l) * s_{Amp}(l) + d(l) - \alpha s_{Amp}(l) \\ &= z(l) + d(l), \end{aligned} \tag{10}$$

where $z(l)$ is the amplified signal excluding the direct path given by $z(l) = \sum_{j=0}^{\infty} h_l(j) s_{Amp}(l-j-1)$.

Assuming that the RIR does not change much during half a frame length L, it has been shown in [9] that

$$Z(i,k) \approx \sum_{j=0}^{\infty} h_{i+\frac{L}{2}}(j) S_{Amp}(i-j-1,k), \tag{11}$$

where $S_{Amp}(i-j-1,k)$ is the Short Time Discrete Fourier Transform (STDFT) coefficient of a frame of amplified clean

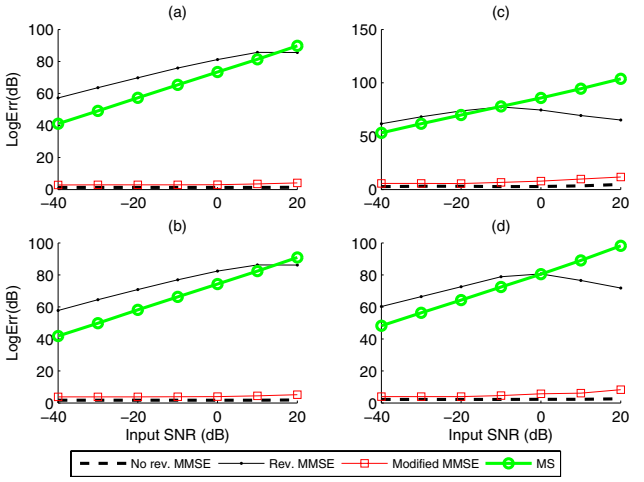

Fig. 2. LogErr (dB) (a) white noise, (b) modulated white noise, (c) non-stationary train noise, (d) babble noise.

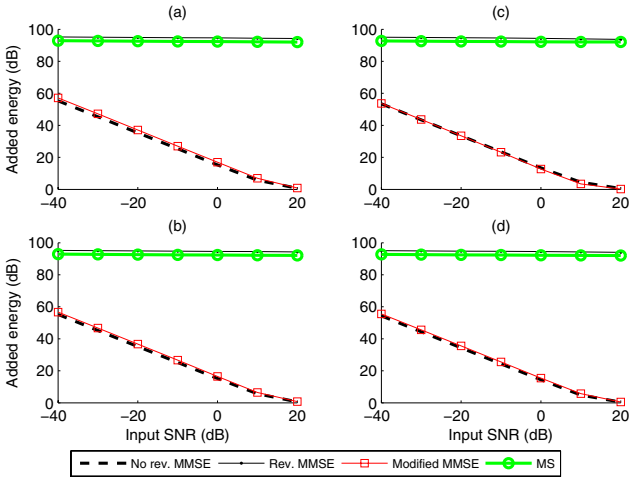

Fig. 3. Added energy (dB) (a) white noise, (b) modulated white noise, (c) non-stationary train noise, (d) babble noise.

speech starting at sample point $i-j-1$. Using (10) and (11), we can then write

$$Y(i,k) = Z(i,k) + D(i,k).$$

Instead of (3), which is conditioned on the noisy DFT coefficient X only, we estimate the noise power conditioned on both Y and S_{Amp} and assume (estimates of) ν^2 and η to be given, that is, $E\{|D|^2|Y, S_{Amp}, \nu^2, \eta\}$. Along similar lines as in [6] we can then derive

$$E\{|D|^2|Y\} = \left(\frac{1}{(1+\xi_{ML})^2} + \frac{\xi_{ML}}{(1+\xi_{ML})\zeta}\right)|Y|^2 B(\xi), \tag{12}$$

where the parameters in (12) are now defined as

$$Y = X - \alpha S_{Amp} = Z + D, \tag{13a}$$

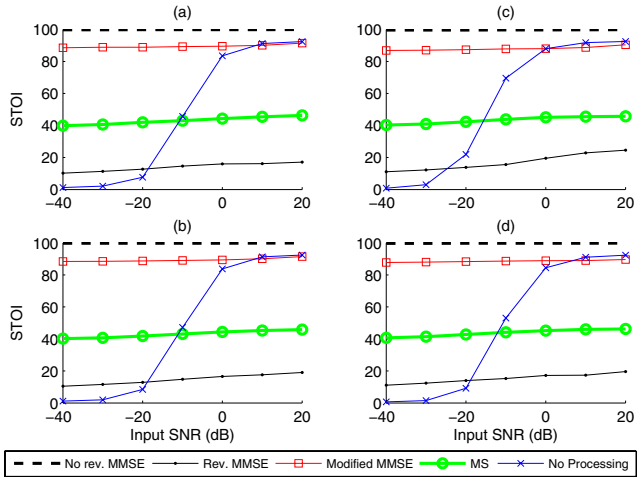

Fig. 4. STOI (a) white noise, (b) modulated white noise, (c) non-stationary train noise, (d) babble noise.

$$\xi_{ML} = \max\left(\frac{Y^2(i,k)}{\hat{\sigma}_D^2(i-1,k)} - 1, 0\right), \tag{13b}$$

$$\xi = \frac{\sigma_Z^2}{\sigma_D^2(i-1,k)}, \zeta = \frac{|Y(i,k)|^2}{\sigma_D^2(i-1,k)}, \tag{13c}$$

and

$$\sigma_Z^2 = \sum_{p=0}^{\infty} \nu_{i+\frac{L}{2}}^2 \times e^{-2\eta_{i+\frac{L}{2}} \times p} \times S_{Amp}^2(i-p-1). \tag{13d}$$

4. SIMULATION AND EXPERIMENTAL RESULTS

4.1. simulation setup

The evaluation is carried out on 6 minutes of speech taken from the TIMIT database [10]. First, this test set is amplified to reach the sound pressure level of 62.35 dB SPL and is then convolved with the RIR to generate the reverberated signal. The noise sources are white noise, modulated white noise with the setup as in [6], babble noise and non-stationary train noise. Assumming the noise power is compared with the original clean signal of 62.35 dB SPL, the SNR ranges from -40 to 20 dB. Setting the sampling frequency at 8 kHz, the frame-based processing is done on square-root-Hann-windowed frames with the length of 256 samples and 50 % overlap. Finally, the proposed noise PSD estimation algorithm is employed in the intelligibility improvement algorithm presented in [3]. Moreover, the enclosed space is simulated by generating a RIR based on Polack's model with $\nu^2 = 0.25$ and $T_{60} = 0.3$ sec. Here, we assume the room characteristics, i.e. ν^2 and T_{60}, have been measured through an offline method and are given. The summation in (13d) is upper-bounded to 400.

4.2. performance measures

The true noise PSD is estimated by smoothing noise periodgrams by a recursive averaging with time-constant 0.9, as below

$$\sigma_D^2(i,k) = 0.9\sigma_D^2(i,k-1) + 0.1|D(i,k)|^2. \quad (14)$$

The estimated noise PSDs obtained from the different methods are compared with the true noise PSD by symmetric log-error distortion measure (Fig. 2) defined as

$$\text{LogErr} = \frac{1}{IK}\sum_{k=1}^{K}\sum_{i=1}^{I}\left|10\log_{10}\left[\frac{\sigma_D^2(i,k)}{\hat{\sigma}_D^2(i,k)}\right]\right| \quad (15)$$

where I and K denote the total number of frames and frequency bins, respectively. As the employed method for intelligibility improvement [3] increases the SNR by ampliying the clean speech, the added energy to the original clean speech is also quantified (Fig. 3). Further to instrumentally evaluate the intelligibility improvement, the short-time objective intelligibility measure STOI [11] measure is used (Fig. 4).

4.3. performance evaluation

We evaluated the baseline MMSE noise PSD estimator as presented in [6] in both reverberant and non-reverberant environments, to show the loss in performance when the environment turns out to be reverberant and compare this to minimum statistics (MS) [12] and our proposed method. MS and our proposed method are evaluated only in the reverberant case. Although in the non-reverberant case, the baseline MMSE achieves the expected good results as in line with [3], both the baseline MMSE and MS are unable to accurately track the noise PSD in the reverberant condition. This is shown by the large LogErr values in Fig. 2. Moreover, Fig. 3 demonstrates that the baseline MMSE method as well as MS leads to overamplified clean speech in reverberant conditions. This results as the noise PSD is over-estimated leading to an overamplification by the speech intelligibility improvement algorithm. It can be inferred that the original noise PSD estimators from [6] and [12] that are not adjusted to reverberant conditions consider parts of the reverberated signal as noise, resulting in an overestimated noise PSD. Our proposed method, which is a modification of the baseline MMSE algorithm, is able to deal with the reverberation, as rather low LogErr and low added energy values validate this. Predicted intelligibility by STOI measure values shown in Fig. 4, demonstrates how the more accurate noise PSD estimate improves the final predicted intelligibility over existing noise PSD estimators.

5. CONCLUSION

In this paper, we derived an MMSE noise estimator that can be employed in an intelligibility improvement module of a public address system. Existing noise PSD estimators wrongly classify parts of the reverberant energy as noise and unnecessarily amplify the clean speech. Experimental results show superior performance of the proposed algorithm in reverberant noisy conditions.

6. REFERENCES

[1] B. Sauert and P. Vary, "Near end listening enhancement: speech intelligibility improvement in noisy environments," in *ICASSP*, 2006, vol. 1, pp. 493–496.

[2] B. Sauert, G. Enzner, and P. Vary, "Near end listening enhancement with strict loudspeaker output power constraining," in *IWAENC*, 2006.

[3] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index," in *EUSIPCO*, 2009, pp. 1844–1848.

[4] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *ICASSP*, 2012.

[5] B. Sauert and P. Vary, "Near end listening enhancement considering thermal limit of mobile phone loudspeakers," in *ESSV*, 2011.

[6] R. C. Hendriks, R. Heusdens, and J. Jensen, " MMSE based noise PSD tracking with low complexity," in *ICASSP*, 2010, pp. 4266–4269.

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

[8] J. D. Polack, *La transmission de l' energie sonore dans les salles*, Ph.D. thesis, Universite du Maine, 1988.

[9] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1746 – 1765, 2010.

[10] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, 2010.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. SAP*, vol. 9, no. 5, pp. 504–512, 2001.