IMPROVED NOISE POWER SPECTRAL DENSITY TRACKING BY A MAP-BASED POSTPROCESSOR

Aleksej Chinaev, Alexander Krueger, Dang Hai Tran Vu, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
{chinaev,krueger,tran,haeb}@nt.uni-paderborn.de

ABSTRACT

In this paper we present a novel noise power spectral density tracking algorithm and its use in single-channel speech enhancement. It has the unique feature that it is able to track the noise statistics even if speech is dominant in a given time-frequency bin. As a consequence it can follow non-stationary noise superposed by speech, even in the critical case of rising noise power. The algorithm requires an initial estimate of the power spectrum of speech and is thus meant to be used as a postprocessor to a first speech enhancement stage. An experimental comparison with a state-of-the-art noise tracking algorithm demonstrates lower estimation errors under low SNR conditions and smaller fluctuations of the estimated values, resulting in improved speech quality as measured by PESQ scores.

***Index Terms*—** Noise power estimation, MAP parameter estimation, speech enhancement

1. INTRODUCTION

The noise power spectral density (PSD) estimation algorithm is a key component of any speech enhancement system, as the achievable quality of the enhanced speech critically depends on it. When the noise is non-stationary, it is not sufficient to update the noise PSD in speech absence periods only — the noise PSD needs to be tracked even during speech activity. Several algorithms have been proposed for this, such as the minimum statistics (MS) algorithm, the minima controlled recursive averaging (MCRA) algorithm and its improved version IMCRA, a subspace noise tracking algorithm and the minimum mean squared error estimation of the noise periodogram. These and other algorithms have been recently compared on a common evaluation database in [1].

All of the aforementioned noise PSD estimation algorithms make two basic assumptions: First, the noise is assumed to be “more stationary” than speech and second, time-frequency bins can be found, which allow for the observation of solely the noise even in a speech-noise mixture. While still requiring the first, the algorithm presented here, does away with the second requirement: The noise PSD can be estimated even if speech is dominant in the time-frequency bin under consideration. This is made possible by assuming that an initial estimate of the speech power is given. The estimation of the parameters of the noise process in the short time Fourier transform domain is then cast into the problem of estimating the variance of a complex-valued zero-mean white Gaussian random process (WGP) in the presence of ”noisy” observations (the target process being “corrupted” by the superposed speech). This is solved by the Maximum a Posteriori (MAP)-based estimator recently proposed in [2].

The remainder of this paper is organized as follows: in Section 2 we derive the MAP estimator for the variance of the short-time Fourier transform (STFT) coefficients of the noise process given noisy speech. Section 3 illustrates, how the proposed estimator is integrated in a speech enhancement system. Next we describe the experimental framework and the results of the performance evaluation in Section 4, before we draw some conclusions in Section 5.

2. MAP-BASED NOISE VARIANCE ESTIMATION

In the following we derive an estimator of the noise PSD given the STFT of the noisy speech $Y_{k,l} = X_{k,l} + N_{k,l}$, where k represents the frequency bin index and l the frame index. $X_{k,l}$ and $N_{k,l}$ denote the STFTs of the clean speech and the noise signals, respectively. Since the PSD estimator treats each frequency component identically and independently of the others, we will drop the frequency bin index.

The STFTs are modeled as complex-valued zero-mean WGPs. Adjoining the real and imaginary parts to two-dimensional column vectors $\mathbf{Y}_l$, $\mathbf{X}_l$ and $\mathbf{N}_l$, respectively, and denoting the time-variant variances of $\mathbf{X}_l$ and $\mathbf{N}_l$ by $\sigma^2_{\mathbf{X},l}$ and $\sigma^2_{\mathbf{N},l}$, the probability density function (PDF) of $\mathbf{Y}_l$ is a zero-mean Gaussian with variance $\sigma^2_{\mathbf{X},l} + \sigma^2_{\mathbf{N},l}$, if we assume that speech and noise are independent.

Our goal is to estimate the noise PSD, i.e., the variance $\sigma^2_{\mathbf{N},l}$, from the noisy observation $\mathbf{y}_l$, which is a realization of $\mathbf{Y}_l$. To achieve this we are going to extend a special case of the method that was proposed in [2] to the two-dimensional case. There, an approximate MAP estimate of the parameters

The work was in part supported by Deutsche Forschungsgemeinschaft under contract no. Ha3455/8-1.

of a Gaussian has been derived, if the observation is drawn from the superposition of the Gaussian with another Gaussian of known variance. Applying this to the problem at hand, we are going to derive a MAP estimate of $\sigma^2_{\mathbf{N},l+1}$, at frame $l+1$, given a prior distribution of the variance and a new observation $\mathbf{y}_{l+1}$ and further assuming knowledge of $\sigma^2_{\mathbf{X},l+1}$.

Let us assume for the moment that the noise is stationary, i.e., $\sigma^2_{\mathbf{N},l+1} = \sigma^2_{\mathbf{N}}$. If $\sigma^2_{\mathbf{X},l+1} = 0$, the scaled inverse chi-square distribution

$$p_{\sigma^2_{\mathbf{N}}}(\sigma^2) \propto (\sigma^2)^{-\frac{\nu_l+2}{2}} \cdot e^{-\frac{\nu_l \lambda_l^2}{2\sigma^2}} \tag{1}$$

with the hyper parameters ν_l and λ_l^2, indicating the degrees of freedom and the scale, respectively, is a conjugate prior for the observation PDF

$$p_{\mathbf{Y}_{l+1}|\sigma^2_{\mathbf{N}}}(\mathbf{y}_{l+1}|\sigma^2) = \frac{1}{\pi\sigma^2} e^{-\frac{|\mathbf{y}_{l+1}|^2}{\sigma^2}}. \tag{2}$$

The posterior can then be computed as [3]

$$\begin{aligned} p_{\sigma^2_{\mathbf{N}}|\mathbf{Y}_{l+1}}(\sigma^2|\mathbf{y}_{l+1}) &\propto p_{\mathbf{Y}_{l+1}|\sigma^2_{\mathbf{N}}}(\mathbf{y}_{l+1}|\sigma^2) \cdot p_{\sigma^2_{\mathbf{N}}}(\sigma^2) && (3)\\ &\propto (\sigma^2)^{-\frac{\nu_l+4}{2}} \cdot e^{-\frac{2|\mathbf{y}_{l+1}|^2 + \nu_l \cdot \lambda_l^2}{2\sigma^2}} && (4)\\ &= (\sigma^2)^{-\frac{\nu_{l+1}+2}{2}} \cdot e^{-\frac{\nu_{l+1}\lambda_{l+1}^2}{2\sigma^2}}, && (5) \end{aligned}$$

where the parameters ν_l and λ_l^2 of (1) have been replaced by

$$\nu_{l+1} := \nu_l + 2 \quad \text{and} \quad \lambda_{l+1}^2 := \frac{2|\mathbf{y}_{l+1}|^2 + \nu_l \lambda_l^2}{\nu_l + 2}. \tag{6}$$

With these update rules for the hyper parameters, the posterior (3) has the same form as the prior (1). The MAP estimate at frame $l+1$ of the variance $\sigma^2_{\mathbf{N}}$ is then given by

$$\begin{aligned} \hat{\sigma}^2_{\mathbf{N},l+1} &= \underset{\sigma^2}{\operatorname{argmax}} \left[p_{\sigma^2_{\mathbf{N}}|\mathbf{Y}_{l+1}}(\sigma^2|\mathbf{y}_{l+1}) \right] && (7)\\ = \frac{\nu_{l+1}}{\nu_{l+1}+2} \cdot \lambda_{l+1}^2 &= \frac{\nu_{l+1}}{\nu_{l+1}+2} \cdot \left(\frac{2|\mathbf{y}_{l+1}|^2}{\nu_{l+1}} + \hat{\sigma}^2_{N,l} \right). && (8) \end{aligned}$$

Thus, ν_{l+1} determines the weight, by which a new observation $\mathbf{y}_{l+1}$ is taken into account for the parameter update.

Now let's turn to the case where $\sigma^2_{\mathbf{X},l+1} \neq 0$. Then, the posterior PDF has the form [2]

$$\begin{aligned} &p_{\sigma^2_{\mathbf{N}}|\mathbf{Y}_{l+1}}(\sigma^2|\mathbf{y}_{l+1}) && (9)\\ &\propto (\sigma^2_{\mathbf{X},l+1} + \sigma^2)^{-1} \cdot (\sigma^2)^{-\frac{\nu_l+2}{2}} \cdot e^{-\left(\frac{|\mathbf{y}_{l+1}|^2}{\sigma^2_{\mathbf{X},l+1} + \sigma^2} + \frac{\nu_l \lambda_l^2}{2\sigma^2} \right)}, \end{aligned}$$

which is different from (1), i.e., (1) is no longer a conjugate prior. Before coming back to this issue we first show how the maximum of (9) can be found, i.e., the MAP estimate $\hat{\sigma}^2_{\mathbf{N},l+1}$.

With $\psi := \sigma^2 > 0$, searching for the maximum of (9) corresponds to finding the minimum of

$$\begin{aligned} &f(\psi) := -2\ln(p_{\sigma^2_{\mathbf{N}}|\mathbf{Y}_{l+1}}(\psi|\mathbf{y}_{l+1}))\\ &\propto \underbrace{2\ln(\sigma^2_{\mathbf{X},l+1} + \psi) + (\nu_l + 2)\ln(\psi)}_{f_1(\psi)} + \underbrace{\frac{2|\mathbf{y}_{l+1}|^2}{\sigma^2_{\mathbf{X},l+1} + \psi} + \frac{\nu_l \lambda_l^2}{\psi}}_{f_2(\psi)}. \end{aligned}$$

Note that $f_1(\psi)$ is strictly monotonically increasing and $f_2(\psi)$ is strictly monotonically decreasing for $\psi > 0$. And since $\lim_{\psi \to 0} f(\psi) = \lim_{\psi \to \infty} f(\psi) = \infty$, $f(\psi)$ has exactly one local positive minimum, which can be found as a positive root of

$$\begin{aligned} f'(\psi) &= \frac{2}{\sigma^2_{X,l+1} + \psi} + \frac{\nu_l + 2}{\psi} - \frac{2|\mathbf{y}_{l+1}|^2}{(\sigma^2_{X,l+1} + \psi)^2} - \frac{\nu_l \lambda_l^2}{\psi^2}\\ &= \frac{2[\psi - \psi_a]\psi^2 + (\nu_l + 2)[\psi - \psi_b](\sigma^2_{X,l+1} + \psi)^2}{(\sigma^2_{X,l+1} + \psi)^2 \cdot \psi^2} \end{aligned} \tag{10}$$

$$\text{with} \quad \psi_a = |\mathbf{y}_{l+1}|^2 - \sigma^2_{\mathbf{X},l+1} \quad \text{and} \quad \psi_b = \frac{\nu_l}{\nu_l + 2}\lambda_l^2. \tag{11}$$

Since the denominator is always positive, it suffices to consider the numerator, which will be denoted by $g(\psi)$. It can be verified that $g(b_D) < 0$ and $g(b_U) > 0$, where

$$b_D = \min(\max(0, \psi_a), \psi_b) \quad \text{and} \quad b_U = \max(\psi_a, \psi_b).$$

The desired positive root ψ_{l+1} of $g(\psi)$ can then be determined efficiently using a combination of a bisection and Newton approach [2] and delivers the MAP estimate $\hat{\sigma}^2_{\mathbf{N},l+1} = \psi_{l+1}$. In order to obtain an efficient MAP estimation procedure on successive observations $\mathbf{y}_{l+1}, \mathbf{y}_{l+2}, \ldots$ we need to establish a conjugate prior. This is done by approximating (9) by a scaled inverse chi-squared distribution according to (1) with maximum at ψ_{l+1}. As stated by (8), this is achieved by setting

$$\lambda_{l+1}^2 := \frac{\nu_{l+1} + 2}{\nu_{l+1}} \cdot \psi_{l+1} \quad \text{and} \quad \nu_{l+1} = \nu_l + 2. \tag{12}$$

If $\mathbf{N}_l$ is a non-stationary process, the estimation of the time-variant variance $\sigma^2_{\mathbf{N},l+1}$ can be accomplished by a simple modification: The parameter ν_l from (1) is kept at some constant value $\nu_{l+1} = \nu_l = \nu_0$. In this way we introduce a forgetting mechanism, since the weight of the new observation is kept constant, irrespective of the number of observations used so far, see Eq. (8). The parameter ν_0 thus acts as a smoothing parameter: The larger ν_0 the smoother is the time trajectory of variance estimates. The choice of ν_0 depends on the desired trade-off between the estimator's variance in stationary noise and the ability to track the time-variant $\sigma^2_{\mathbf{N},l}$.

The proposed algorithm has very low computational complexity. Another important property is that it has only one parameter, ν_0, which needs to be chosen according to the degree of non-stationarity of the noise.

3. INTEGRATION INTO SPEECH ENHANCEMENT SYSTEM

In practice, the speech variance $\sigma^2_{\mathbf{X},l}$ is not known. We therefore propose to use the introduced noise PSD estimator, denoted by MAP-B in the following, as a postprocessor of a first speech enhancement system, which provides an estimate $\hat{\sigma}^2_{\mathbf{X},k,l}$ of the clean speech variance for all frequency bins k.

To do so it requires a first noise PSD estimator, for which any of the known algorithms can be taken. In our experiments we used the IMCRA algorithm [4] for this purpose.

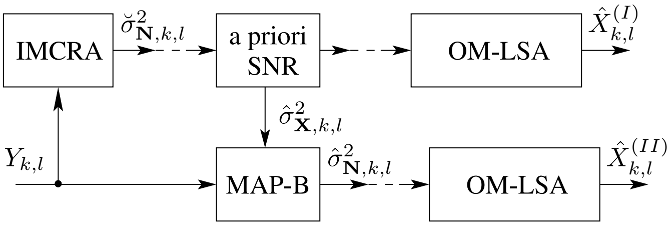

IMCRA
$\breve{\sigma}^2_{\mathbf{N},k,l}$
a priori SNR
OM-LSA
$\hat{X}^{(I)}_{k,l}$
$\hat{\sigma}^2_{\mathbf{X},k,l}$
$Y_{k,l}$
MAP-B
$\hat{\sigma}^2_{\mathbf{N},k,l}$
OM-LSA
$\hat{X}^{(II)}_{k,l}$

Fig. 1. Integration of MAP-B estimator into a single-channel speech enhancement system.

Fig.1 illustrates the setup. Given the noisy speech STFT $Y_{k,l}$ at its input the IMCRA algorithm delivers a first estimate of the noise variance $\breve{\sigma}^2_{\mathbf{N},k,l}$. With this the a priori SNR is estimated by a decision-directed approach, from which the desired estimate of the speech variance $\hat{\sigma}^2_{\mathbf{X},k,l}$ is obtained. With this estimate the MAP-B algorithm will deliver an updated noise PSD estimate $\hat{\sigma}^2_{\mathbf{N},k,l}$, which can be used in the optimally-modified log-spectral amplitude (OM-LSA) estimator $\hat{X}^{(II)}_{k,l}$ of the clean speech signal [4].

4. EXPERIMENTAL FRAMEWORK AND PERFORMANCE EVALUATION

In our experiments the clean speech signals were taken from the TIMIT database [5]. By concatenating sentences and removing beginning and trailing silences, a male speaker and a female speaker test sample were created, each consisting of speech of seven different speakers and having a total length of 3 minutes. The speech signals were sampled at 16 kHz and the STFT spectral analysis used a Hamming window of 512 samples length with a frame overlap of 75%.

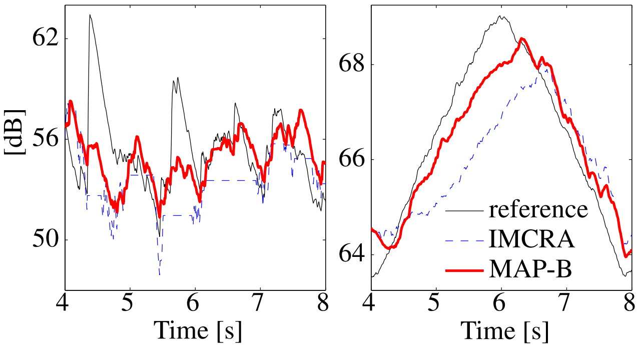
The clean speech signals were artificially degraded by adding noise. Four different noise types were considered. Stationary white Gaussian noise (WGN), 'Babble' noise and so called 'Factory-1' noise were taken from the Noisex92 database [6]. Moreover, to examine the performance of the algorithm in highly non-stationary noise, we generated a modulated version of 'Stationary WGN' named 'Triangular WGN' according to the modulation signal used: The level of the noise was increased at a rate of 2 dB/s for a period of 3 seconds and then reduced again to the original level at the same rate. We conducted experiments at different SNR levels. The global SNR was varied from −5 dB to 15 dB in steps of 5 dB.

For the reference noise PSD, against which the estimates are to be compared, we adopted the approach taken in [1], i.e. a recursive temporal smoothing was applied to the known noise periodogram:

$$\sigma^2_{\mathbf{N},k,l} = 0.95 \cdot \sigma^2_{\mathbf{N},k,l-1} + 0.05 \cdot |\mathbf{N}_{k,l}|^2, \qquad (13)$$

with start value $\sigma^2_{\mathbf{N},k,0} = |\mathbf{N}_{k,0}|^2$ for $l = 1$.

To verify the claim stated in the introduction that the MAP-B estimator is able to estimate the noise even if the speech is dominant, consider the trajectories depicted in Fig.2(a). It shows the reference noise PSD $\sigma^2_{\mathbf{N},l}$ and the time-variant noise PSDs $\breve{\sigma}^2_{\mathbf{N},l}$ and $\hat{\sigma}^2_{\mathbf{N},l}$ estimated by IMCRA and MAP-B, respectively, for 'Babble' noise at an SNR of 0 dB and a frequency bin $k = 97$ (center frequency 3 kHz). In the experiments we set $\nu_0 = 40$ corresponding to a time constant of 0.164 s. Unlike IMCRA, MAP-B continuously updates its estimate and thus follows the reference noise PSD more closely. Fig.2(b) displays an extract of the noise PSD estimates for 'Triangular WGN' averaged over frequency. It shows that the response of the MAP-B estimator to rising noise power is much faster than that of the IMCRA estimator.



(a) (b)

Fig. 2. Noise PSD estimations for a noisy speech signal at an SNR of 0 dB: (a) degraded by 'Babble' noise for a single frequency bin $k = 97$ (center frequency 3 kHz); (b) degraded by 'Triangular WGN' averaged over frequency.

For a quantitative evaluation we adopted the performance measures proposed in [1], however with a slight modification. The first measure is the minimum averaged log distance LE_m between the estimated and reference noise PSD

$$LE_m = \underset{\tau}{min} \left[LE_m(\tau) \right] = \underset{\tau}{min} \left[\frac{1}{LK} \sum_{l=1}^{L} \sum_{k=1}^{K} \Delta_{k,l}(\tau) \right] \quad (14)$$

$$\text{with} \quad \Delta_{k,l}(\tau) = \left| 10 \log_{10} \frac{\sigma^2_{\mathbf{N},k,l-\tau}}{\tilde{\sigma}^2_{\mathbf{N},k,l}} \right| \quad \text{and} \quad \tilde{} \in \{\breve{},\hat{}\}.$$

LE_m is the mean of the logarithmic difference between the 'true' and estimated noise variances, averaged over frequency bins and frames. In contrast to [1], the time-variant true noise variances were first aligned to the temporal sequence of the estimates and the reported value is the smallest value obtained by varying the lag τ. This optimization was done because the computation of the 'true' noise variance according to (13) and the estimation procedures can induce different latencies, which should be eliminated.

The second performance measure is the variance of the logarithmic difference

$$LE_v = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} LE_{v,n,m} \quad \text{with} \tag{15}$$

$$LE_{v,n,m} = \frac{1}{L_s K_s}\sum_{l=mL_s+1}^{(m+1)L_s}\ \sum_{k=nK_s+1}^{(n+1)K_s} (\Delta_{k,l}(\tau_{min}) - \mu_l^n)^2$$

$$\text{and} \quad \mu_l^n = \frac{1}{K_s}\sum_{k=nK_s+1}^{(n+1)K_s} \Delta_{k,l}(\tau_{min}),$$

where $\tau_{\text{min}} = \text{argmin}_\tau[LE_m(\tau)]$. N and M are the number of frequency and time blocks over which the variance estimates are averaged. $LE_{v,n,m}$ is the value computed from the time-frequency block starting at time frame mL_s and frequency bin nK_s and having the length of $L_s = 2K_s$ frames and $K_s = 16$ frequency bins. LE_v measures the amount of fluctuations in the estimated noise PSD. The stronger these fluctuations the more likely will the speech enhancement system produce musical tones [1].

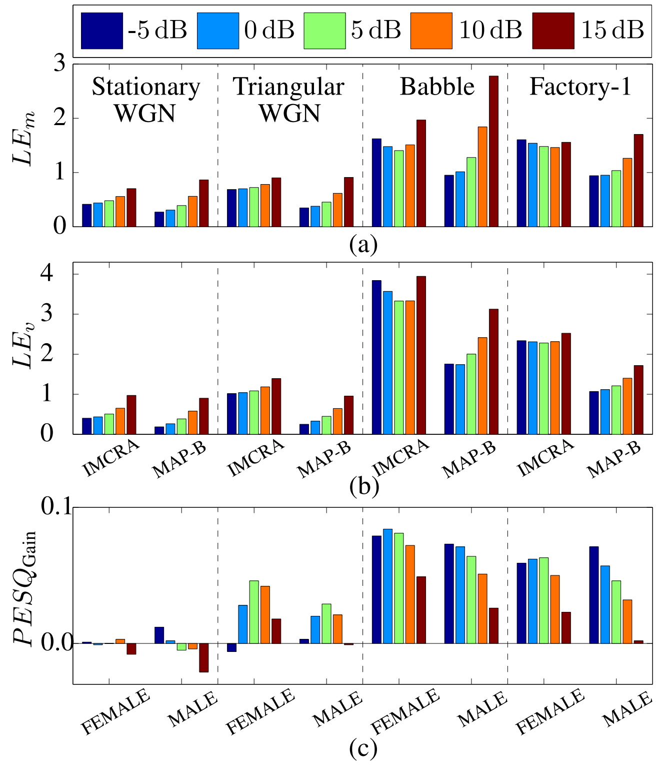


Fig. 3. Performance measures for various noise types and SNRs: (a) and (b) LE_m and LE_v, respectively, obtained by IMCRA and MAP-B; (c) $PESQ_{\text{Gain}}$ compared to IMCRA for female and male speakers.

Fig.3 (a) and (b) compares IMCRA and MAP-B with respect to the performance measures LE_m and LE_v for various noise types and noise levels. Depicted are the averages over performance measures for male and female speakers. These results demonstrate that the proposed MAP-B method obtains lower LE_m values for all noise types and SNRs less than or equal to 5 dB. For 'Triangular WGN' and 'Factory-1' the LE_m of the MAP-B estimator is better as well for an SNR of 10 dB. Further the MAP-B estimator yields lower variance LE_v of the logarithmic difference for all noise types and SNRs than the IMCRA estimator. Fig.3 (c) shows the gains $PESQ_{\text{Gain}} = PESQ_{\text{MAP-B}} - PESQ_{\text{IMCRA}}$ of scores for the perceptual speech quality obtained by PESQ [7] calculated using the enhanced signals $\hat{X}_l^{(II)}$ and $\hat{X}_l^{(I)}$ for female and male speakers separately. As expected, the improved noise tracking has a favourable, though small, effect on speech quality for non-stationary noise types.

5. CONCLUSION

We have proposed a new approach for the noise PSD estimation in a speech enhancement system. It is based on the maximum a posteriori estimation of the noise variance of a non-stationary complex white Gaussian process in the presence of an additive Gaussian interference of known variance. In contrast to most known noise PSD estimators it is able to track the noise statistics even if the speech is dominant in noisy speech. The method has low computational complexity and has only one parameter which has to be adjusted according to the degree of non-stationarity of the noise. The experimental evaluation has shown that MAP-B obtains a lower estimation error under low SNR conditions and a lower fluctuation of the estimated values under all tested environments, resulting in an improved speech quality for non-stationary noise types as measured by PESQ scores.

6. REFERENCES

[1] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *ICASSP 2011*, pp. 4640–4643.

[2] A. Krueger and R. Haeb-Umbach, "MAP-based estimation of the parameters of non-stationary Gaussian processes from noisy observations," in *ICASSP 2011*, pp. 3596–3599.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Capman & Hall, second edition, 2003.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[5] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[6] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[7] "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, Geneva, February 2001.