Quality Analysis and Optimization of the MAP-based Noise Power Spectral Density Tracker

Aleksej Chinaev, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098, Paderborn, Germany
Email: {chinaev, haeb}@nt.uni-paderborn.de
Web: www.nt.uni-paderborn.de

Abstract

It has been lately shown that noise tracking and speech denoising can be improved by a postprocessor established on a maximum a-posteriori based (MAP-B) noise power spectral density (PSD) estimation algorithm. In the current contribution we investigate the MAP-B estimator by carrying out a quality analysis comprising the following three steps. First, we analyse the estimator with respect to unbiasedness and consistency, second, the tracking ability in non-stationary noise is investigated, and finally, the sensitivity of the MAP-B noise tracker with respect to estimation errors in the preprocessing stage is considered. The findings are used to develop an optimized MAP-B postprocessor. The performance comparison with the original MAP-B tracker indeed reveals improved performance at high signal-to-noise ratios.

1 Introduction

Speech spectral enhancement systems include a noise PSD estimation algorithm which aims at tracking the noise statistics from the noisy observations. This is especially difficult to do at time-frequency (TF) bins with a strong speech power, where many sophisticated algorithms for noise PSD estimation hold their estimates at some constant value. In [1] we introduced the MAP-B postprocessor, which is able to track the statistics of non-stationary noise continuously, even if speech is dominant in the TF bin. To do so, it requires an initial estimate of the current clean speech power. This can be provided by a first enhancement stage (ES), upon which the MAP-B estimator acts as a postprocessor.

In [1] it has been shown that the performance of the improved minima controlled recursive averaging (IMCRA) algorithm [2], which was used as noise tracker in the first ES, can be improved by the MAP-B postprocessor. However, some important performance aspects of the MAP-B noise tracker have not yet been investigated. The present contribution is going to fill this gap by analysing the quality of MAP-B estimator with respect to unbiasedness and consistency. Furthermore we inspect estimator's ability to track non-stationary noise. To investigate its susceptibility to speech power estimation errors produced in the first ES, we also carry out a sensitivity analysis. Since a closed-form analytical estimation rule of the MAP-B algorithm does not exist, we carry out the quality analysis numerically with the aid of the Monte Carlo method [3].

2 Simulation Framework

A block diagram of the simulation framework is depicted in Fig.1. The short-time Fourier transform (STFT) of a noisy speech signal at a given frequency bin is modelled as a complex-valued zero-mean Gaussian process Y_l, which is composed of the random processes N_l and X_l with given time-variant variances $\sigma^2_{N,l}$ and $\sigma^2_{X,l}$, where $l \in [1,L]$ is a time frame index with a total signal length of $L = 10000$ samples in all simulations.

The MAP-B algorithm, introduced in [1], calculates a noise power estimate $\hat{\sigma}^2_{N,l}$ from the observation Y_l and an estimate of the current clean speech power $\tilde{\sigma}^2_{X,l}$. To this end the prior probability density function (PDF) of the noise power $p_{\sigma^2_N}$ is modelled by a scaled inverse chi-square (SICS) distribution, which is a conjugate prior of the observation PDF p_Y in case of speech absence. In order to maintain an efficient MAP estimation procedure for speech presence the posterior PDF $p_{\sigma^2_N|Y}$ is approximated by a SICS distribution, whose mode can be effectively calculated by using a combination of a bisection and Newton approach.

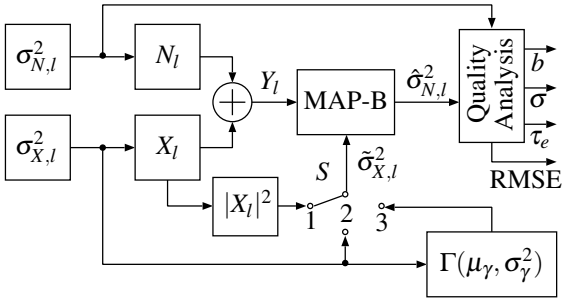

Figure 1: Simulation framework.

In our simulation framework $\tilde{\sigma}^2_{X,l}$ is modelled in three different ways as illustrated by the three positions of the switch S in Fig.1. The instantaneous true speech power $|X_l|^2$ is passed to the MAP-B postprocessor, if S is connected to the node 1. For the switch position 2 the true speech variance $\sigma^2_{X,l}$ is forwarded to the noise tracker. In the third position, randomly distributed values of $\tilde{\sigma}^2_{X,l}$ are generated by drawing from the gamma distribution

$$\tilde{\sigma}^2_{X,l} \sim \Gamma(\mu_\gamma, \sigma^2_\gamma) \tag{1}$$

with mean $\mu_\gamma = \sigma^2_{X,l}$ and variance $\sigma^2_\gamma = var(\tilde{\sigma}^2_{X,l})$ for different constant values of σ^2_γ.

In all simulations we set the true speech power at some constant value $\sigma^2_{X,l} = \sigma^2_X$ according to the desired signal-to-noise ratio (SNR) defined as $\mathrm{SNR} = 10 \cdot \log_{10}(P_X/P_N)$, where P_X and P_N are the average power of X_l and N_l respectively. For each SNR value we run K experiments resulting in estimates $\hat{\sigma}^2_{N,k,l}$ with $k \in [1,K]$. The dependence of $\hat{\sigma}^2_{N,k,l}$ on the initial value is reduced by setting $\hat{\sigma}^2_{N,k,0}$ close to the true value. The quality analysis comprises the calculation of an average scaled bias b, an average standard deviation σ, an estimator's latency τ_e and the root mean square error (RMSE). These are defined in the following.

3 Quality Analysis

In [1] we presented two different estimation rules of the MAP-B estimator: The first, which is intended for stationary noise, employs an increasing degrees of freedom parameter $\nu_{l+1} = \nu_l + 2$, and the second, meant for non-stationary noise, uses a constant $\nu_{l+1} = \nu_l = \nu_0$. ν_0 determines the bandwidth of the noise tracker, where large values of ν_0 correspond to small bandwidth. In the following we investigate unbiasedness and consistency for both rules separately. Analysis of tracking ability and sensitivity have been done only for the second rule, which is more relevant for practical applications.

3.1 Unbiasedness and Consistency

In order to investigate the MAP-B estimator with regard to its unbiasedness and consistency we generate the complex-valued random process N_l with the constant true noise power $\sigma^2_{N,l} = \sigma^2_N = 1$ and set the switch S to the position 1 as illustrated in Fig.1.

3.1.1 Increasing Degrees of Freedom

Fig. 2 shows histograms of $\hat{\sigma}^2_{N,k,l}$ at $l = \{250, 1000, 2500, 5000, 10000\}$ samples after the start of the estimation process for the three different SNR values $\{-5, 0, 5\}$dB and the total number of simulations $K = 10000$. The recursion on the degrees of freedom was initialized with $\nu_0 = 40$.

Defining the scaled bias b_l observed after processing of l samples by

$$b_l = \frac{\bar{\sigma}^2_{N,l} - \sigma^2_N}{\sigma^2_N}, \quad \text{where} \quad \bar{\sigma}^2_{N,l} = \frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}^2_{N,k,l}, \tag{2}$$

it can be observed in Fig. 2(a) that the MAP-B estimator seems to be asymptotically bias-free ($\lim_{l\to\infty} b_l = 0$) for small SNR. However, the bias of the MAP-B postprocessor does not vanish in Fig. 2(c). Furthermore it grows slightly with increasing SNR as further simulations have shown.

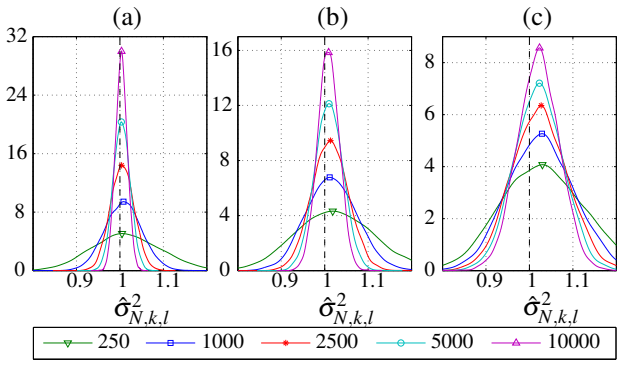

Figure 2: Histograms of MAP-B estimates $\hat{\sigma}^2_{N,k,l}$ for different $l = \{250, 1000, 2500, 5000, 10000\}$ at an SNR of (a) -5 dB, (b) 0 dB and (c) 5 dB.

The figures further suggest, that, at least for low SNR, the MAP-B estimator is consistent, since the variance of the estimates (i.e. variance of the histograms) tends to zero if the number of observations l tends to infinity. Again, the larger the SNR, the more observations are required to reach this asymptotic behaviour.

The shapes of the histogram in Fig. 2 further seem to indicate that the lower the SNR the more the estimation error tends to have a normal distribution. This is to be expected, because for $\sigma^2_{X,l} = 0$ and thus SNR $\to -\infty$ the estimate is a simple linear combination of the observations [1]

$$\hat{\sigma}^2_{\mathrm{N},l+1} = \frac{\nu_{l+1}}{\nu_{l+1}+2} \cdot \hat{\sigma}^2_{\mathrm{N},l} + \frac{2}{\nu_{l+1}+2} \cdot |Y_{l+1}|^2. \tag{3}$$

3.1.2 Constant Degrees of Freedom

The case of a constant value ν_0 is more relevant for practical applications, as it allows for tracking the statistics of non-stationary noise. Fig.3 displays the resulting scaled bias

$$b = \frac{1}{L/2}\sum_{l=L/2}^{L} b_l \tag{4}$$

averaged over the last $L/2$ samples and the average standard deviation σ defined as

$$\sigma = \frac{1}{L/2}\sum_{l=L/2}^{L} \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\sigma}^2_{N,k,l} - \bar{\sigma}^2_{N,l}\right)^2}. \tag{5}$$

Results of $K = 1000$ simulations are presented as a function of SNR in dB for different degrees of freedom ν_0.

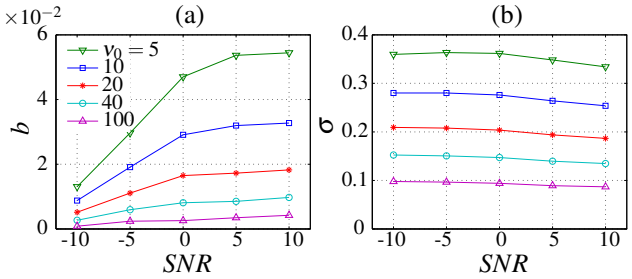

Figure 3: (a) Average scaled bias b and (b) average standard deviation σ of MAP-B estimator over SNR values in dB for degrees of freedom $\nu_0 = \{5, 10, 20, 40, 100\}$.

In Fig. 3(a) it can be observed that the average scaled bias b of the MAP-B estimator is always positive. It increases with growing SNR values and decreasing ν_0 values. According to Fig. 3(b) the average standard deviation σ rises with decreasing ν_0 and seems to be independent of SNR values. As a conclusion, the parameter ν_0 should be chosen large enough to ensure a low value of the mean squared error $\mathrm{MSE} = b^2 \cdot \sigma^4_N + \sigma^2$.

3.2 Tracking Ability

An analysis of estimator's tracking ability is performed by setting the switch S to the position 2, see Fig. 1, and generating a time-variant true noise power

$$\sigma^2_{N,l} = 1 + \sin^2\left(\frac{2\pi \cdot l}{L}\right). \tag{6}$$

First we calculated the RMSE as a function of the delay τ defined as

$$\mathrm{RMSE}(\tau) = \frac{1}{L-l_0}\sum_{l=l_0+1}^{L} \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\sigma}^2_{N,k,l} - \sigma^2_{N,l-\tau}\right)^2}. \tag{7}$$

for $K = 1000$ experiments. For computation of $\mathrm{RMSE}(\tau)$ samples $\hat{\sigma}^2_{N,k,l}$ are gathered after $l_0 = 50$, in order to avoid

the impact of the acquisition phase and solely concentrate on the tracking performance. The optimal latency of the tracker is defined as

$$\tau_e = \underset{\tau}{\operatorname{argmin}}[\mathrm{RMSE}(\tau)]. \tag{8}$$

However, this often results in unacceptably large delays in a speech communication application. We therefore consider in the following the performance for zero latency, i.e., $\mathrm{RMSE}(\tau = 0)$, see Fig. 4(a). As expected, estimators performance deteriorates for growing SNR values. Moreover $\mathrm{RMSE}(\tau = 0)$ increases both for little values of ν_0 because of growing estimator's bias b and for large values of ν_0 because of rising estimator's latency τ_e. The figure further suggests that $\nu_0 = 40$ is a good choice.

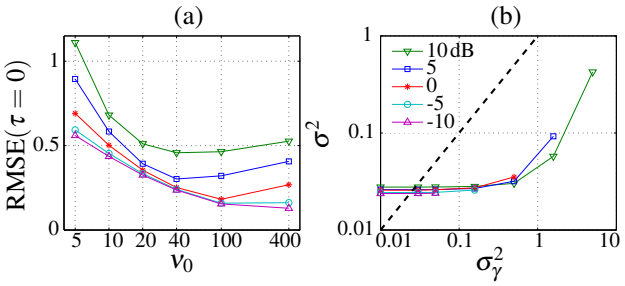

Figure 4: (a) $\mathrm{RMSE}(\tau = 0)$ over degrees of freedom ν_0 and (b) estimator variance σ^2 over the variance of the gamma distribution σ_γ^2 for SNR values $\{-10,-5,0,5,10\}$dB.

3.3 Sensitivity to Erroneous Speech Power Estimates

Here we consider the behaviour of the MAP-B postprocessor in case of randomly distributed errors $\tilde{\sigma}^2_{X,l}$ according to the equation (1), which corresponds to the position 3 of the switch S in Fig.1. Using a constant noise power $\sigma^2_{N,l} = \sigma^2_N = 1$ and the constant degrees of freedom $\nu_0 = 40$ we run $K = 1000$ simulations for different SNR values of $\{-10,-5,0,5,10\}$dB. The resulting average estimator variance σ^2 is shown in Fig. 4(b) as a function of the variance σ_γ^2 of the speech power estimation error for the range $\sigma_\gamma^2 \in (10^{-2}, 10)$. The black dashed line corresponds to the variance of the estimator based on the trivial estimation rule: $\sigma_Y^2 = \tilde{\sigma}_X^2 + \hat{\sigma}_N^2$, which results in $\mathrm{var}(\hat{\sigma}_N^2) = \sigma_\gamma^2$.

It is striking that the MAP-B estimator is quite insensitive to estimation errors in σ_X^2: it holds its estimation variance at a low constant value for a wide range of σ_γ^2 values. However, from a certain value of σ_γ^2 on, the estimator variance increases rapidly.

4 Optimization of MAP-B estimator

According to the results of the quality analysis we suggest two improvements of the MAP-B postprocessor: In order to obtain an unbiased estimate we subtract an estimate of the bias: $\hat{\sigma}^2_{N,UN} = (1 - b(\mathrm{SNR})) \cdot \hat{\sigma}^2_N$. As the bias depends on the SNR, see Fig. 3(a), it is approximated by

$$b(\mathrm{SNR}) = \beta_{max} \cdot \left(\frac{\arctan(\mathrm{SNR})}{\pi} + \frac{1}{2} \right) \tag{9}$$

with a bias compensation factor $\beta_{max} = 0.01$.

In [1] we noticed, that MAP-B postprocessor could improve the noise tracking predominantly for low SNR values. This can be attributed to the fact that the speech power estimation of the first ES is very unreliable during time periods with the strong speech power. An effective way to improve the MAP-B performance for high SNR values is to decrease its bandwidth with increasing SNR. This can be accomplished by updating rule

$$\nu_0(\mathrm{SNR}) = \nu_0 + \frac{\Delta\nu_0}{\pi} \cdot \arctan(\mathrm{SNR}), \tag{10}$$

where $\nu_0 = 40$ and $\Delta\nu_0 = 10$.

5 Experimental Evaluation

The MAP-B postprocessor is used as a noise PSD tracker in the 2nd ES of a single-channel speech enhancement system, which is depicted in Fig. 5.

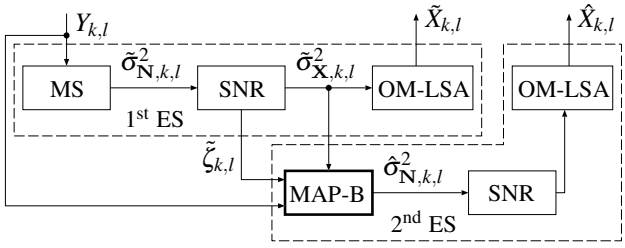

Figure 5: Integration of MAP-B postprocessor into a single-channel speech enhancement system.

$Y_{k,l}$ represents the STFT of the noisy speech, where k and l are the frequency bin and the frame index respectively. For noise tracking in the first ES we used the Minimum Statistics (MS) approach [4] with the data window length $D_{MS} = 120$, the smoothing parameter $\alpha_{MS} = 0.9$ and the bias correction factor $B_{min,MS} = 1.66$, which are fixed for all frequency bins and frames. Notice that in [4] the values of α_{MS} and $B_{min,MS}$ are estimated for each frequency bin and each frame. The 'SNR'-block represents a decision-directed approach resulting in an estimation of a-priori SNR $\tilde{\xi}_{k,l}$ [5], from which we compute a current speech power estimate $\tilde{\sigma}^2_{\mathrm{X},k,l}$ and a recursive average of the a-priori SNR

$$\tilde{\zeta}_{k,l} = \alpha_\zeta \cdot \tilde{\zeta}_{k,l-1} + (1 - \alpha_\zeta) \cdot \tilde{\xi}_{k,l}$$

with $\alpha_\zeta = 0.7$. While the original MAP-B uses only $\tilde{\sigma}^2_{\mathrm{X},k,l}$ for the calculation of noise power estimates $\hat{\sigma}^2_{\mathrm{N},k,l}$, the improved MAP-B algorithm utilizes additionally $\tilde{\zeta}_{k,l}$ according to (9) and (10). Notice that the bias $b_{k,l}$ and the degrees of freedom $\nu_{0,k,l}$ are updated for each frequency bin and each frame index. Enhanced speech signals $\tilde{X}_{k,l}$ and $\hat{X}_{k,l}$ are calculated by the optimally-modified log-spectral amplitude (OM-LSA) estimator [5].

For experimental evaluation we used the clean speech signals from the TIMIT database [6]. 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 clean speech signals were artificially degraded by the white Gaussian

noise (WGN), 'F16', 'Babble' and 'Factory 1' noise signals from the Noisex92 database [7]. All signals were sampled at 16 kHz and the STFT spectral analysis used a Hamming window of $K = 512$ samples length with a frame overlap of 75 %. The global SNR was varied from -10 dB to 20 dB in steps of 5 dB. The reference noise PSD $\sigma^2_{N,k,l}$ was generated by applying a zero-phase filter to the known noise periodogram $|N_{k,l}|^2$. For this we used the MATLAB function $filtfilt(0.1, [1\ -0.9], |N_{k,l}|^2)$.

The noise tracking performance is evaluated by calculating the reduction of the log-spectral distance between the true and the estimated noise power from the first to the second enhancement stage, defined as $R_{SD} = \tilde{S}_D - \hat{S}_D$ with

$$\breve{S}_D = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{2}{K}\sum_{k=1}^{K/2} 10\log\frac{\sigma^2_{N,k,l}}{\breve{\sigma}^2_{N,k,l}}\right]^{-\frac{1}{2}} \quad \text{for} \quad \breve{\cdot}\in\{\tilde{\cdot},\hat{\cdot}\}.$$

R_{SD}, as depicted in Fig. 6, was calculated by using the original [1] and the improved MAP-B postprocessor in the second ES of the system in Fig. 5 for different noise types and SNR values. Furthermore the reduction of log-spectral distance averaged over all considered SNR values is printed in Fig. 6 for each noise type.

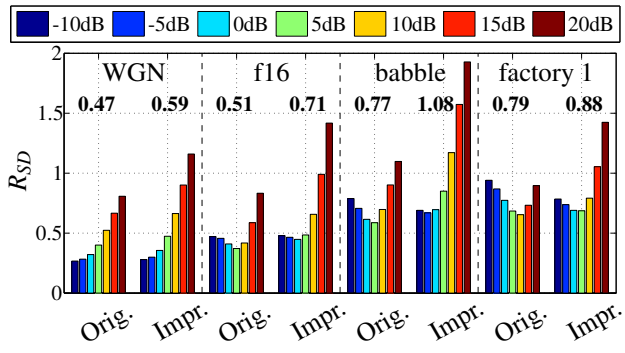

Figure 6: (a) Reduction R_{SD} of log-spectral distance calculated by using the original and the improved MAP-B postprocessor for different noise types and SNR values.

The figure shows that the improved MAP-B postprocessor reduces the estimation error of the MS noise tracker better than the original MAP-B estimator for all noise types. On average, the R_{SD} was improved by the optimized tracker in comparison to the original one by about 28 %. However, the original MAP-B estimator reduced already the estimation error of the MS approach by 15 %, so that the averaged reduction of the estimation error of the MS approach achieved by the improved MAP-B estimator now totals 19 %.

Looking at the experimental results for different SNR values we notice that the optimized MAP-B postprocessor improves the noise tracking particularly well for high SNR values. This improvement is due to the bias reduction and the varying of the bandwidth of the MAP-B approach according to (9) and (10). It should be mentioned that the experimental evaluation of the original MAP-B estimator in [1] has revealed its weakness to track the noise statistics in case of high SNR values. With the proposed modifications, this deficiency has been solved.

However the improved noise tracking hardly affects the quality of the enhanced signal $\hat{X}_{k,l}$, as an additional evaluation, employing the perceptual evaluation of speech quality (PESQ) measure according to ITU-T P.862 [8] has shown. This may be explained by the fact that it is in general difficult to improve the enhanced speech quality solely by improving the noise tracking.

It should be mentioned that similar improvements in noise tracking have been observed for the case of using the IMCRA estimator [2] as the noise tracker in the first ES.

6 Conclusions

A quality analysis of the MAP-B noise PSD estimator has been carried out and the simulation results suggest, that for the case of increasing degrees of freedom parameter ν_l the estimator is asymptotically unbiased only for low SNR. Furthermore, the larger the SNR, the more observations are required to reach a low estimation error. For practical applications with non-stationary noise the parameter, however, should be kept at a constant value $\nu_l = \nu_0$ to avoid an increasingly narrow filter bandwidth. This results in a biased estimate, where the bias is always positive (i.e. the noise power is overestimated) and where the bias grows with increasing SNR and decreasing parameter ν_0.

A striking feature of the MAP-B estimator is that it keeps its low noise estimation variance for a wide range of estimation errors of the input speech power (see Fig. 4b). The variance is furthermore independent of the SNR and, not surprisingly, it grows with decreasing parameter ν_0, i.e. wider filter bandwidth. An analysis of tracking ability has shown that estimator's latency increases with increasing parameter ν_0 as expected. An reasonable value of degrees of freedom seems to be $\nu_0 = 40$.

With the findings of the quality analysis we optimized the MAP-B postprocessor by introducing a SNR-dependent bias reduction and bandwidth adjustment. We then compared the performance of this improved estimator with the original one by experimental evaluation in a single-channel speech enhancement system. The results show that the optimization of the MAP-B postprocessor leads to improved noise tracking, particularly for high SNR values.

References

[1] A. Chinaev, A. Krueger, D. H. Tran Vu, and R. Haeb-Umbach, "Improved noise power spectral density tracking by a map-based postprocessor," *ICASSP 2012*, pp. 4041–4044.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," vol. 11, pp. 466–475, September 2003.

[3] G. S. Fishman, *Monte Carlo: Concepts, Algorithms and Applications*. Springer, 1 ed., 1996.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, pp. 504–512, July 2001.

[5] I. Cohen and S. Gannot, "Spectral enhancement methods," in Handbook of Speech Processing, J. Benesty, M.M. Sondhi, Y. Huang, Chapter 44, pp. 873–901, Springer-Verlag, 2008.

[6] 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.

[7] 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, pp. 247–251, July 1993.

[8] "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.