Improved noise power spectral density tracking by a MAP-based postprocessor

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

University of Paderborn, Germany

March 28th, 2012
Table of Contents

1 Introduction and problem formulation

2 MAP-based noise PSD estimation

3 Experimental framework

4 Performance evaluation

5 Summary and outlook
Motivation

- Noise PSD estimation is a key component
  - to speech enhancement
  - to robust automatic speech recognition

Basic assumptions

- Many sophisticated algorithms rely on two assumptions:
  - Noise 'more stationary' than speech
  - Noise-only time-frequency bins at regular intervals

Here

- MAP-based (MAP-B) postprocessor
  - Estimate of noise power even if speech is dominant in time-frequency bin
  - Initial estimate of the current speech power required
MAP-B as a postprocessor

\[ \hat{x}_t^{(I)} \]

first speech enhancement stage

Noise estimation

\[ \hat{\sigma}_N^{2, k, l} \]

a priori SNR

\[ \hat{\sigma}_X^{2, k, l} \]

Gain

\[ \hat{X}_{k, l}^{(I)} \]

second speech enhancement stage

MAP-B

\[ \hat{\sigma}_N^{2, k, l} \]

a priori SNR

\[ \hat{X}_{k, l}^{(II)} \]

\[ \hat{x}_t^{(II)} \]

\[ x_t \] - clean speech signal

\[ n_t \] - noise signal

\[ y_t \] - noise signal

STANDARD DISCRETE FOURIER TRANSFORM (STANDARD DFT)

INVERSE STANDARD DISCRETE FOURIER TRANSFORM (INVERSE STANDARD DFT)

Improved noise power spectral density tracking by a MAP-based postprocessor

A. Chinaev, A. Krueger, D.H. Tran Vu, R. Haeb-Umbach
Problem formulation

Observed

\[ Y_l = N_l + X_l \]

- We consider \( N_l \) as target process 'corrupted' by clean speech \( X_l \) of known variance \( \sigma^2_{X,l} \)

Our goal

- Estimation of noise variance \( \sigma^2_{N,l+1} \) at frame \( l + 1 \), from the noisy observation \( y_{l+1} \) given
  - the current speech power \( \sigma^2_{X,l+1} \)
  - a priori PDF \( p_{\sigma^2_{N}}(\sigma^2) \) of variance \( \sigma^2_{N,l} \)

Approach

- Maximum A Posteriori (MAP) estimation
Bayesian variance estimation: $Y_l = N_l$

If $\sigma^2_{X,l+1} = 0$ (uncorrupted observ.) and $\sigma^2_{N,l+1} = \sigma^2_N$ (stationary) then the textbook problem:

- Scaled inverse chi-square distribution

$$p_{\sigma^2_N}(\sigma^2) \propto (\sigma^2)^{-\frac{\nu_l+2}{2}} \cdot e^{-\frac{\nu_l \lambda^2_l}{2\sigma^2}}$$

with the degrees of freedom $\nu_l$ and the scale factor $\lambda^2_l$ is conjugate prior to normal observation PDF

$$p_{Y_{l+1} | \sigma^2_N}(y_{l+1} | \sigma^2) = \frac{1}{\pi \sigma^2} \cdot e^{-\frac{|y_{l+1}|^2}{\sigma^2}}$$

- Parameter update $\nu_{l+1} = \nu_l + 2$ and $\lambda^2_{l+1} = \frac{2}{\nu_l + 2} |y_{l+1}|^2 + \frac{\nu_l}{\nu_l + 2} \lambda^2_l$

- MAP-estimate of variance

$$\hat{\sigma}^2_{N,l+1} = \arg\max_{\sigma^2} \left[ p_{\sigma^2_N}(\sigma^2 | y_{l+1}) \right] = \frac{\nu_{l+1}}{\nu_{l+1} + 2} \cdot \lambda^2_{l+1}$$
Extension to non-stationary noise

Still 'uncorrupted' noise: $Y_l = N_l$

If $\sigma^2_{N,l+1}$ is time-variant then:

- The parameter $\nu_l$ is kept at a constant value $\nu_{l+1} = \nu_l = \nu_0$
- This results in recursive smoothing of variance estimate

$$\hat{\sigma}^2_{N,l+1} = (1 - \alpha) \cdot \hat{\sigma}^2_{N,l} + \alpha \cdot |y_{l+1}|^2,$$
where $\alpha = \frac{2}{\nu_0 + 4}$

Choice of $\nu_0$

- Trade-off between tracking ability and estimation error in stationary noise
General case

Observation 'corrupted' by speech: \( Y_l = N_l + X_l \)

- If \( \sigma_{X,l+1}^2 \neq 0 \) and is known then the posterior PDF
  \[
  p_{\sigma_{N,l+1}^2 | Y_{l+1}}(\sigma^2 | Y_{l+1}) \propto (\sigma_{X,l+1}^2 + \sigma^2)^{-1} \cdot \left( \sigma^2 \right)^{-\frac{\nu_l+2}{2}} \cdot e^{-\left( \frac{|Y_{l+1}|^2}{\sigma_{X,l+1}^2 + \sigma^2} + \frac{\nu_l \lambda^2}{2 \sigma^2} \right)},
  \]
  is no longer a conjugate prior for the observation PDF.

In order to maintain an efficient MAP estimation procedure we

- approximate the posterior PDF by a scaled inverse chi-squared distribution,
- and match its maximum \( \hat{\sigma}_{N,l+1}^2 \) with the maximum of the posterior PDF,
- which we calculate efficiently using a bisection and Newton approach.
Experimental framework

MAP-B as postprocessor

- First noise PSD estimator: Improved Minima Controlled Recursive Averaging (IMCRA) algorithm [Cohen, 2003]
- Gain function: Optimally-Modified Log-Spectral Amplitude (OM-LSA) estimator [Cohen, Berdugo, 2001]
Performance evaluation

Setup

• Clean speech: TIMIT database, sentences concatenated to 3 minutes length
• Artificially added noise from Noisex\textsuperscript{92} database:
  ▶ Noise types: 'Stationary WGN', 'Triangular WGN', 'Babble' and 'Factory-1'
  ▶ SNR values: $-5, 0, 5, 10, 15$ dB
• MAP-B estimator: we set $\nu_0 = 40$ corresponding to a time constant of $\simeq 0.164$ s

Reference noise PSD

• Recursive temporal smoothing
  $$\sigma_{N,k,l}^2 = 0.95 \cdot \sigma_{N,k,l-1}^2 + 0.05 \cdot |N_{k,l}|^2,$$  with known noise periodogram $|N_{k,l}|^2$
Sample trajectories of noise variance estimates

- 'Babble' noise at frequency bin $k = 97$ (3 kHz)

- 'Triangular WGN' (averaged over all frequency bins)

- MAP-B: continuous update of noise variance estimate

- MAP-B: faster response to rising noise power

Improved noise power spectral density tracking by a MAP-based postprocessor

A. Chinaev, A. Krueger, D.H. Tran Vu, R. Haeb-Umbach
Quantitative evaluation

- Performance measures adopted from [Taghia et al., 2011]

- Fig. (a): minimum averaged log distance $LE_m$ between the reference and estimated PSD

- MAP-B obtains lower error $LE_m$ for all noise types and SNRs less than or equal to 5 dB or 10 dB

- Fig. (b): variance of the logarithmic difference $LE_v$

- MAP-B yields lower variance $LE_v$ for all noise types and SNRs than the IMCRA
Speech enhancement

- Gain in perceptual speech quality (PESQ) scores

\[
PESQ_{\text{Gain}} = PESQ_{\text{MAP-B}} - PESQ_{\text{IMCRA}}
\]

- MAP-B has a favourable effect on speech quality for non-stationary noise types

Improved noise power spectral density tracking by a MAP-based postprocessor

A. Chinaev, A. Krueger, D.H. Tran Vu, R. Haeb-Umbach
Summary

- Proposed MAP-B estimator is able to track the noise statistics even if the speech is dominant
- Low computational complexity
- Single parameter $\nu_0$
- Experimental evaluation: MAP-B obtains
  - lower estimation error under low SNR conditions
  - lower fluctuation of the estimated values under all tested environments
  - slightly improved speech quality for non-stationary noise types

Outlook

- Investigations about dependence of MAP-B algorithm performance:
  - on the first noise PSD estimator
  - on the degrees of freedom $\nu_0$
Thank you for your attention!

Questions?
Periodogramms of the spoken sentence 'Biblical scholars argue history' for 'factory-1' noise at an SNR of 5 dB
(a) cleen speech signal
(b) noisy speech signal
(c) enhanced speech signal based on IMCRA estimates
(d) enhanced speech signal based on MAP-B estimates

MAP-B has a positive influence on periodogramm of enhanced signal particularly clearly seen in the highlighted window
Minimum Statistics (MS) instead of IMCRA

- Performance measures $LE_m$ and $LE_v$ for female speaker signals

- Fig.(a): MAP-B obtains lower estimation error $LE_m$ than the MS for all noise types and SNRs

- Fig.(b): MAP-B yields lower variance $LE_v$ than the MS for all tested setups
Minimum Statistics (MS) instead of IMCRA

- \( PESQ_{\text{Gain}} = PESQ_{\text{MAP-B}} - PESQ_{\text{NoiseEst}} \)

with 'NoiseEst' \( \in [\text{IMCRA'}, \text{'MS'}] \) for female speaker signals

- A favourable effect of MAP-B estimator on speech quality for non-stationary noise types is for MS smaller than for IMCRA

Improved noise power spectral density tracking by a MAP-based postprocessor

A. Chinaev, A. Krueger, D.H. Tran Vu, R. Haeb-Umbach