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

Introduction

Motivation

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

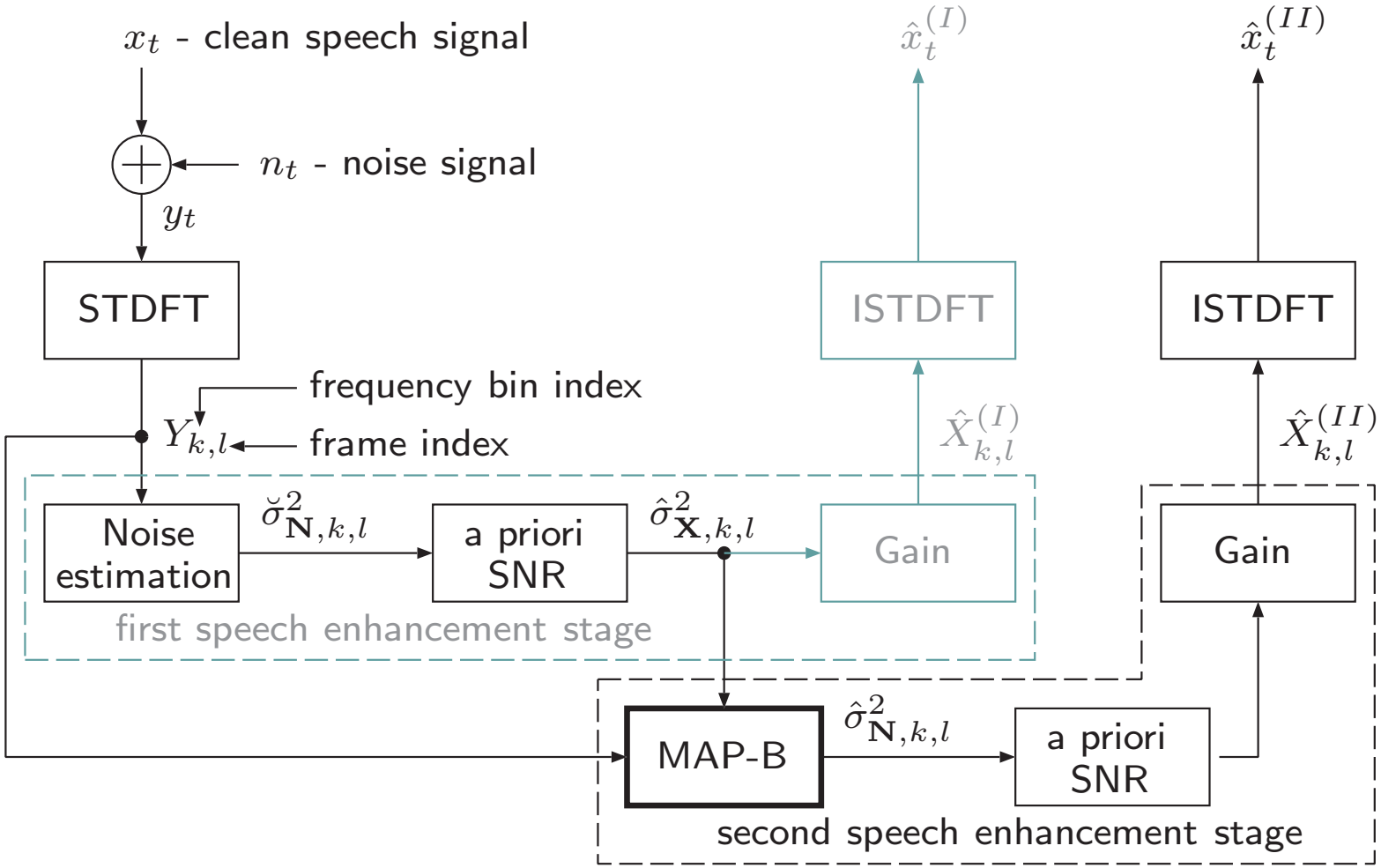
Basic assumptions

- 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



Problem formulation

Observed

$$\mathbf{Y}_l = \mathbf{N}_l + \mathbf{X}_l$$

- We consider $\mathbf{N}_l$ as target process 'corrupted' by clean speech $\mathbf{X}_l$ of known variance $\sigma^2_{\mathbf{X},l}$

Our goal

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

Approach

- Maximum A Posteriori (MAP) estimation

MAP-based noise PSD estimation

Bayesian variance estimation: $\mathbf{Y}_l = \mathbf{N}_l$

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

- Scaled inverse chi-square distribution

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

with the degrees of freedom ν_l and the scale factor λ_l^2 is conjugate prior to normal observation PDF

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

- Parameter update $\quad \nu_{l+1} = \nu_l + 2 \quad$ and $\quad \lambda^2_{l+1} = \frac{2}{\nu_l+2}|\mathbf{y}_{l+1}|^2 + \frac{\nu_l}{\nu_l+2}\lambda_l^2$
- MAP-estimate of variance

$$\boxed{\hat{\sigma}^2_{\mathbf{N},l+1} = \underset{\sigma^2}{\text{argmax}} \left[p_{\sigma^2_{\mathbf{N}}|\mathbf{Y}_{l+1}}(\sigma^2|\mathbf{y}_{l+1})\right] = \frac{\nu_{l+1}}{\nu_{l+1}+2} \cdot \lambda^2_{l+1}}$$

Extension to non-stationary noise

Still 'uncorrupted' noise: $\mathbf{Y}_l = \mathbf{N}_l$

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

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

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

Choice of ν_0

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

General case

Observation 'corrupted' by speech: $\mathbf{Y}_l = \mathbf{N}_l + \mathbf{X}_l$

- If $\sigma^2_{\mathbf{X},l+1} \neq 0$ and is known then the posterior PDF

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

  is no longer a conjugate prior for the observation PDF.

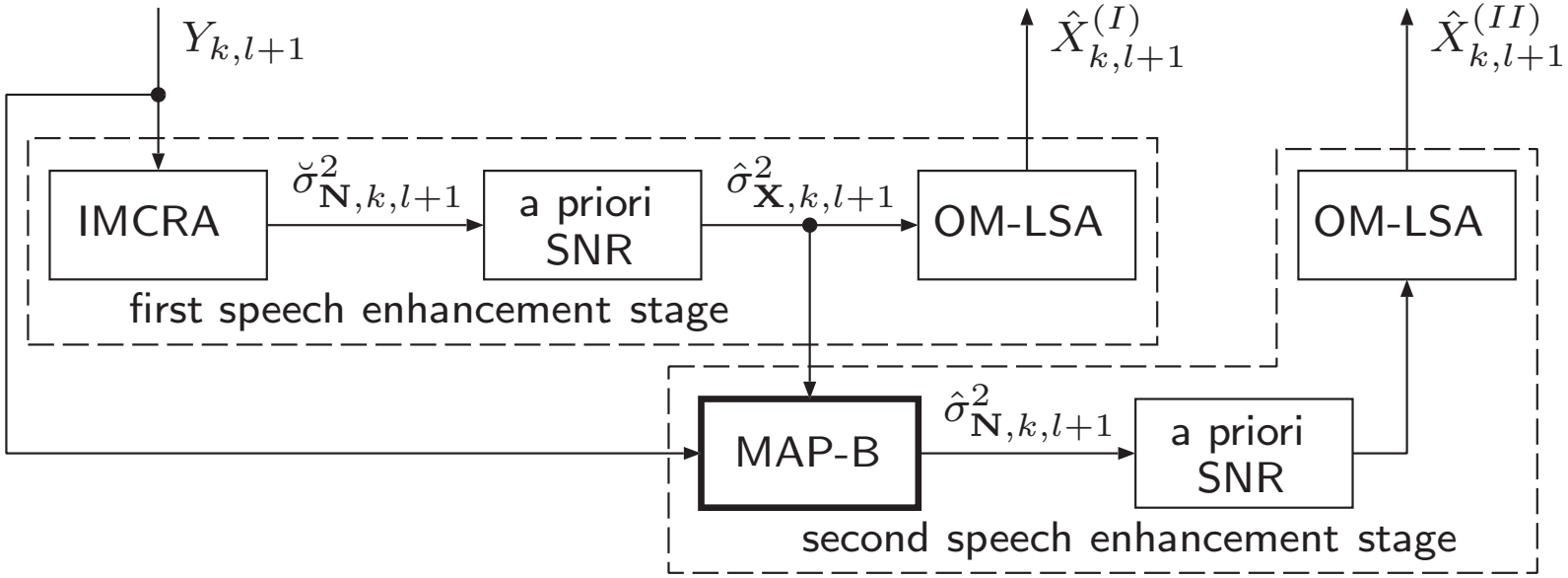
In order to maintain an efficient MAP estimation procedure we

- approximate the posterior PDF by a scaled inverse chi-squared distribution,
- and match its maximum ($\hat{\sigma}^2_{\mathbf{N},l+1}$) 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 Noisex92 database:
  - Noise types: 'Stationary WGN', 'Triangular WGN', 'Babble' and 'Factory-1'
  - SNR values: $-5\,,0\,,5\,,10\,,15\,\mathrm{dB}$
- MAP-B estimator: we set $\nu_0 = 40$ corresponding to a time constant of $\simeq 0.164\,\mathrm{s}$

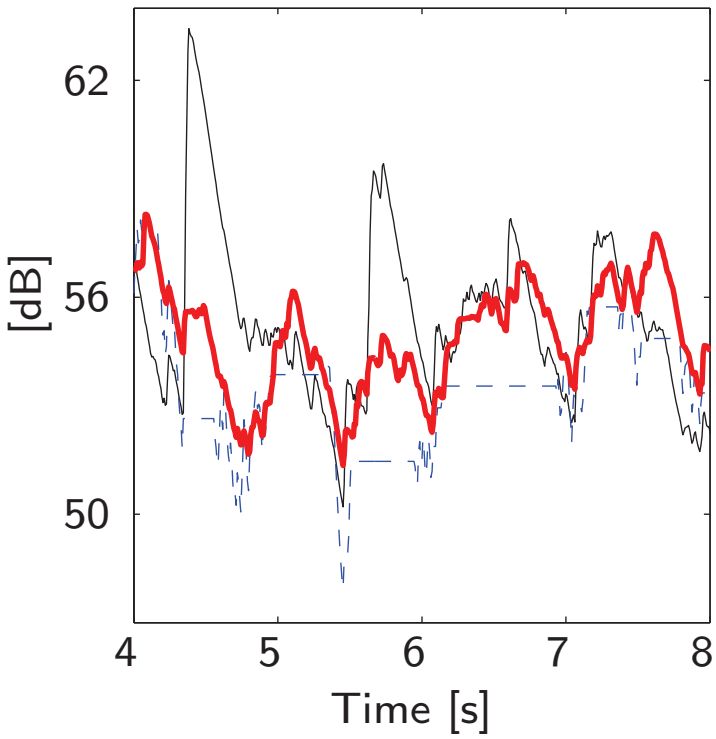
Reference noise PSD

- Recursive temporal smoothing

  $\sigma^2_{\mathbf{N},k,l} = 0.95 \cdot \sigma^2_{\mathbf{N},k,l-1} + 0.05 \cdot |\mathbf{N}_{k,l}|^2$, with known noise periodogram $|\mathbf{N}_{k,l}|^2$

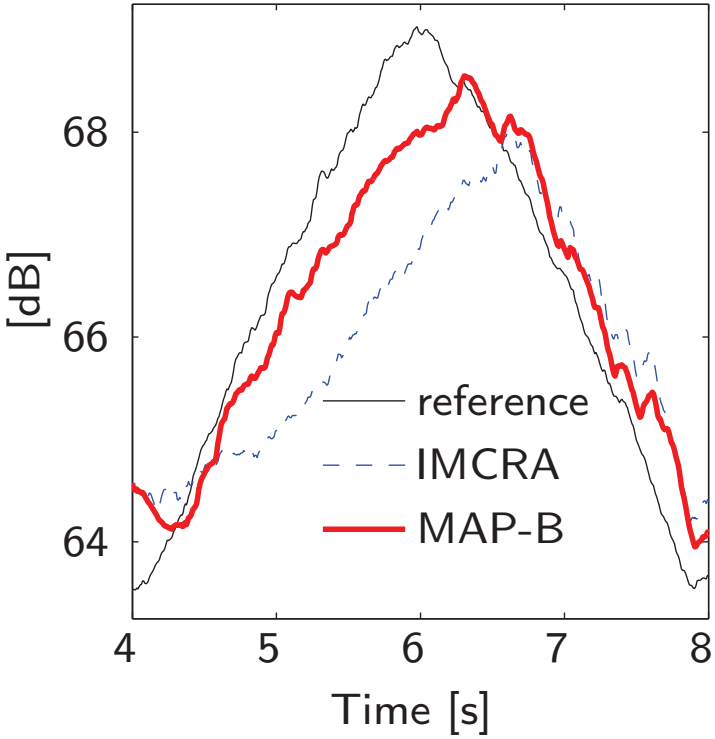
Sample trajectories of noise variance estimates

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



- MAP-B: continious update of noise variance estimate

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

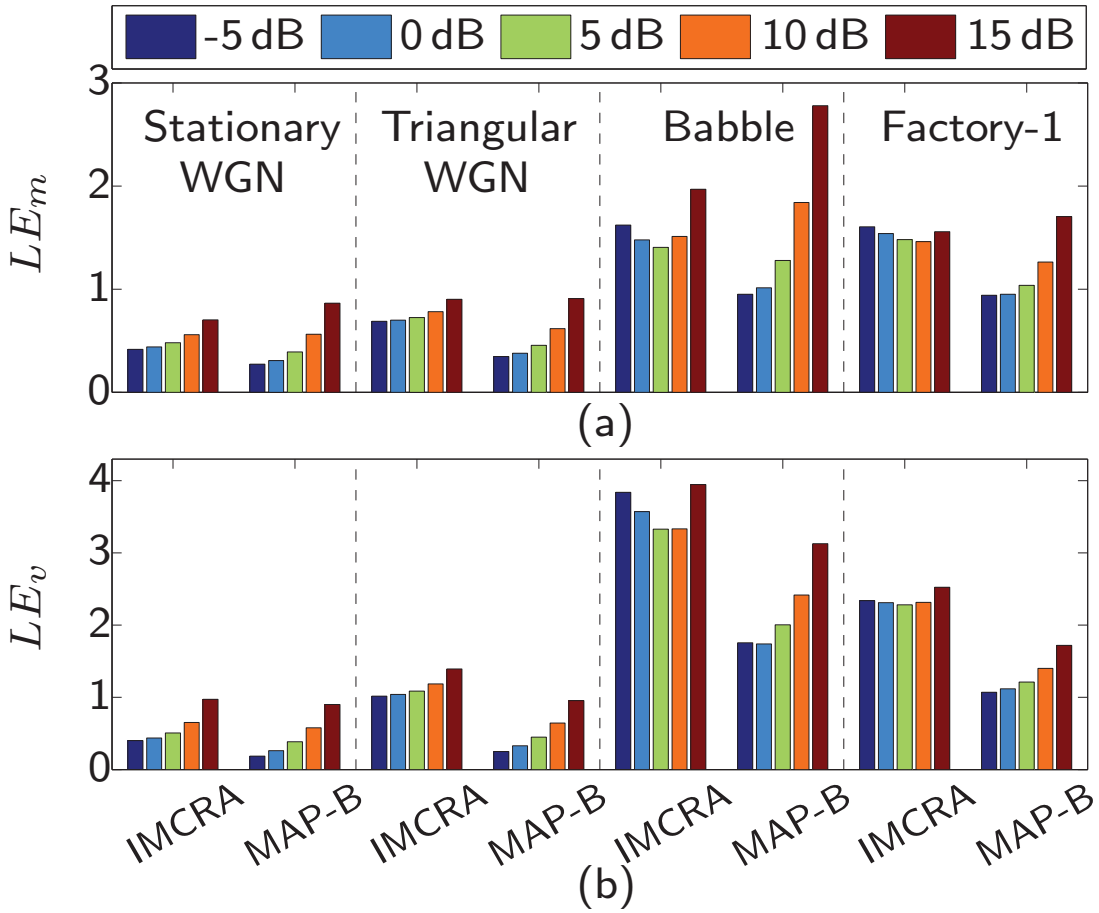


- MAP-B: faster response to rising noise power

Quantitative evaluation

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

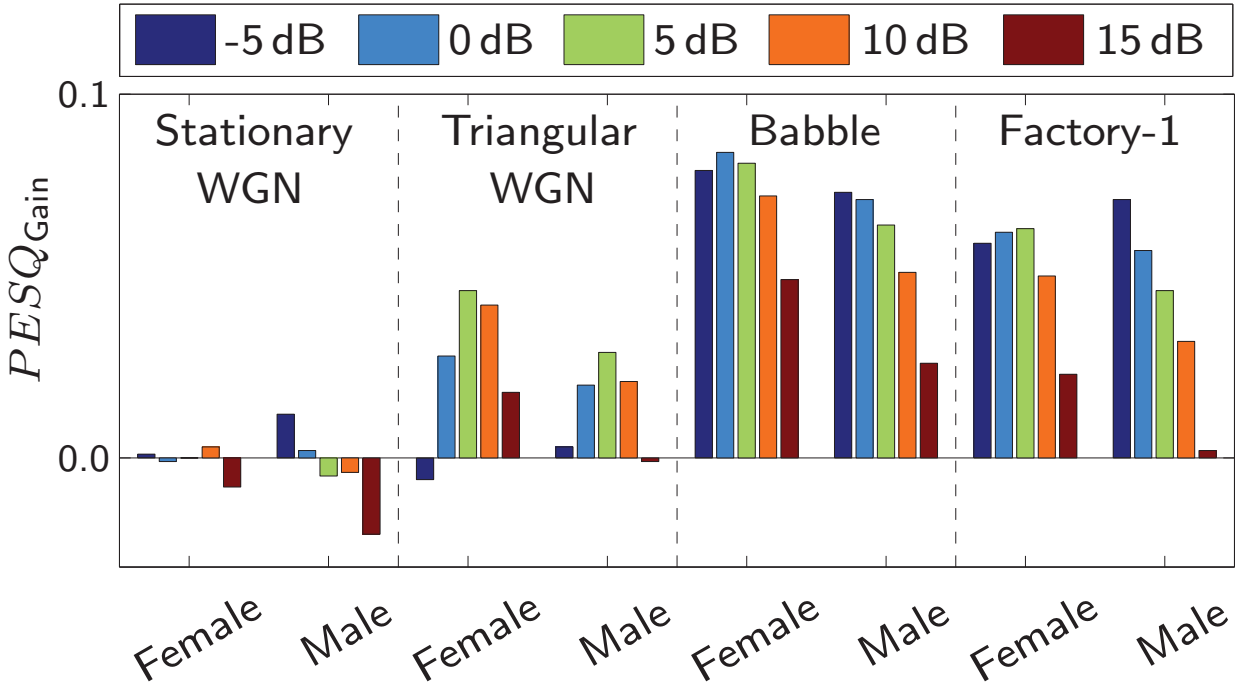
- Fig.(a): minimum averaged log distance LE_m between the reference and estimated PSD
- MAP-B obtains lower error LE_m for all noise types and SNRs less than or equal to 5 dB or 10 dB
- Fig.(b): variance of the logarithmic difference LE_v
- MAP-B yields lower variance LE_v for all noise types and SNRs 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

Summary and outlook

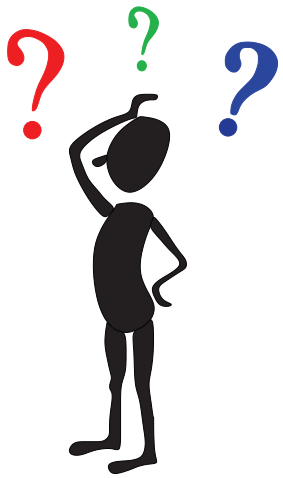
Summary

- Proposed MAP-B estimator is able to track the noise statistics even if the speech is dominant
- Low computational complexity
- Single parameter ν_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 ν_0

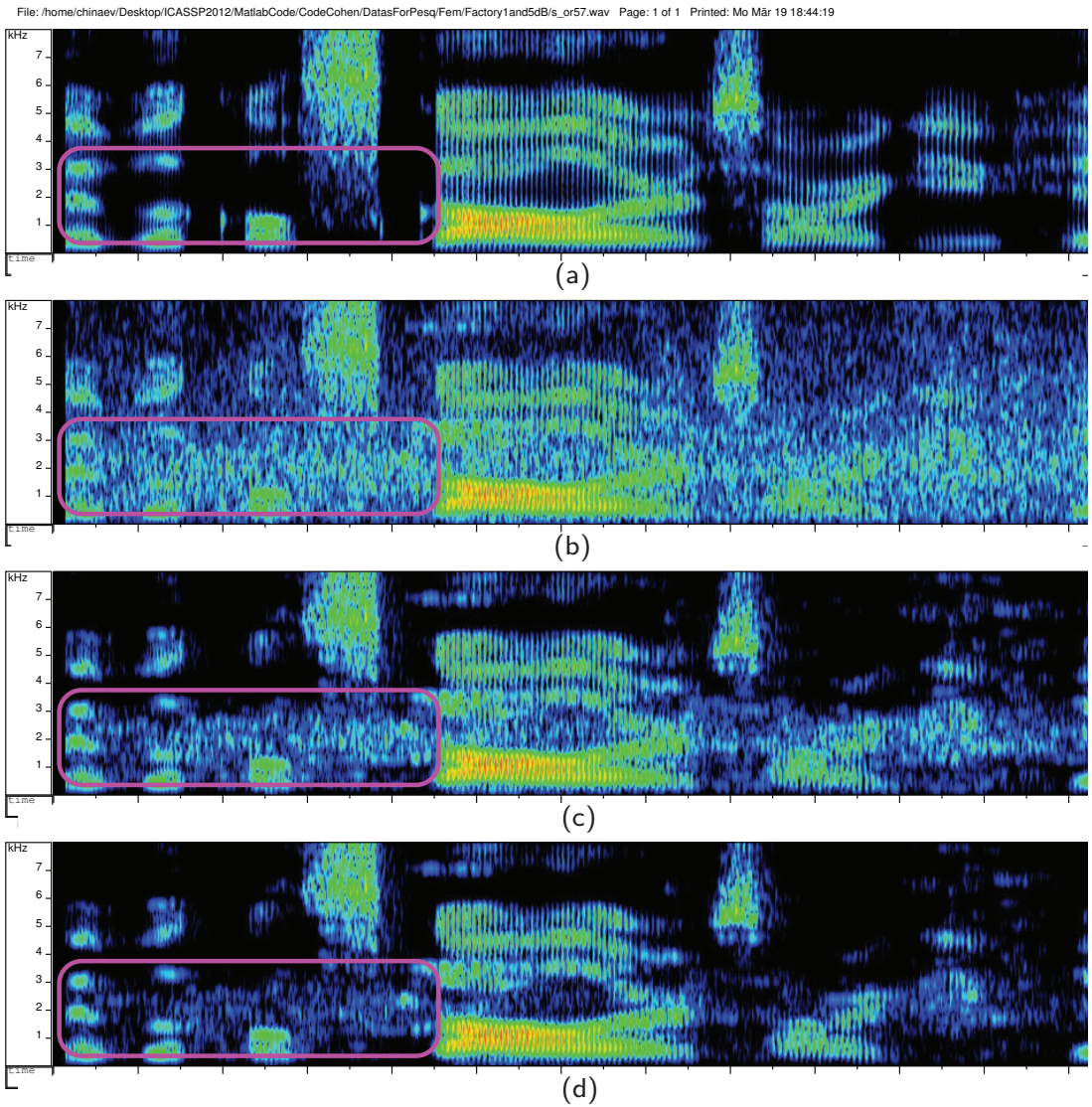
Thank you for your attention!



Questions?

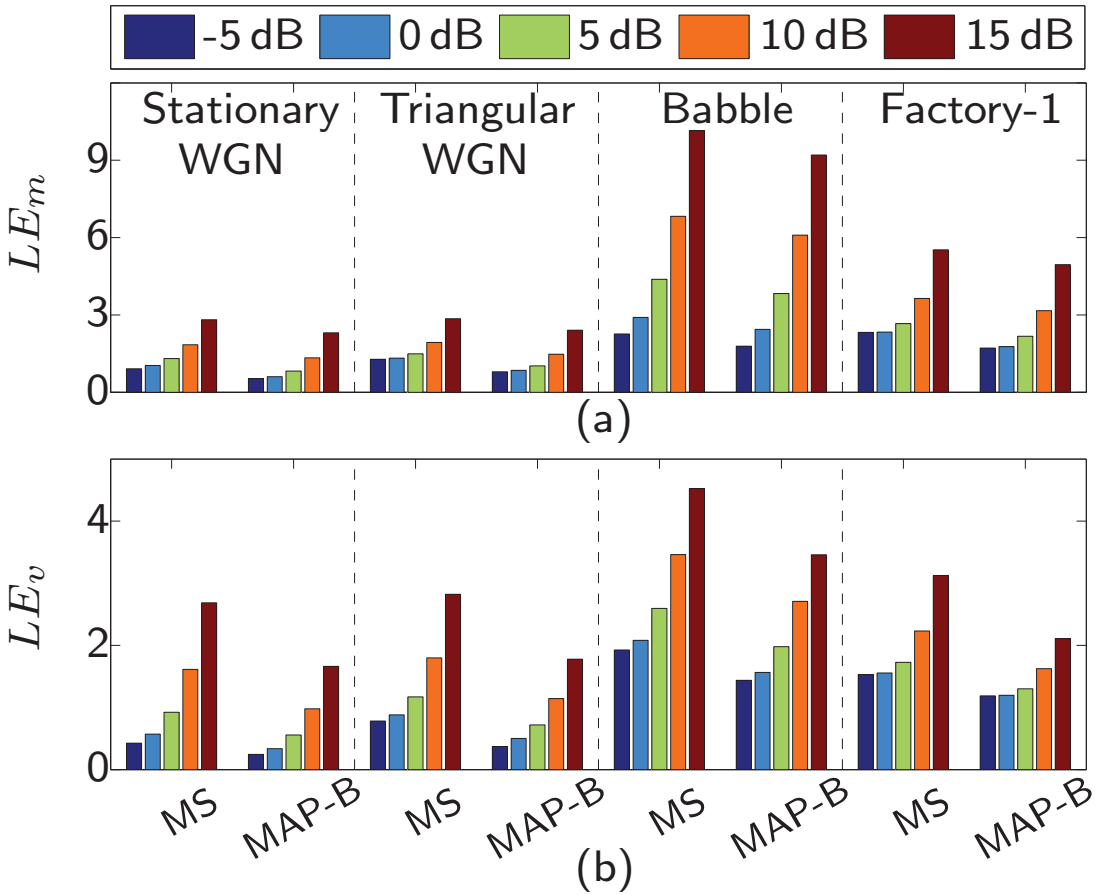
Periodograms of cleen, noisy and enhanced signals

- 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

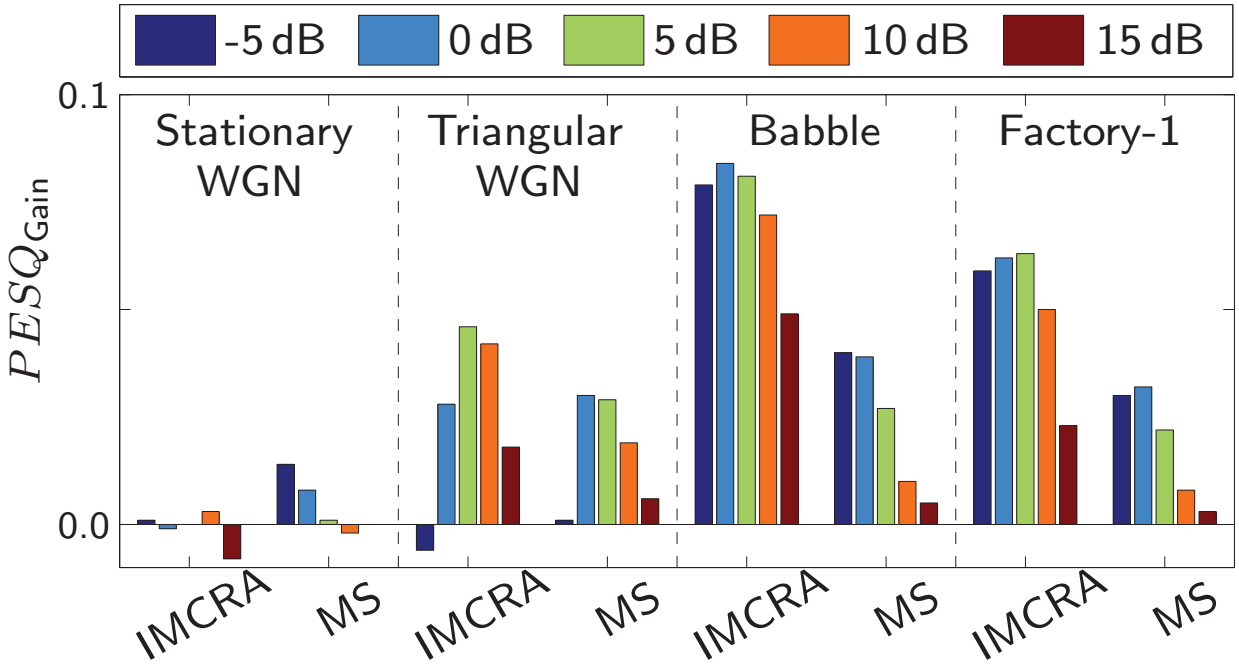


- Fig.(a): MAP-B obtains lower estimation error LE_m than the MS for all noise types and SNRs
- Fig.(b): MAP-B yields lower variance LE_v than the MS for all tested setups

Minimum Statistics (MS) instead of IMCRA

- $PESQ_{\text{Gain}} = PESQ_{\text{MAP-B}} - PESQ_{\text{NoiseEst}}$

with 'NoiseEst' $\in$ ['IMCRA'; '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