A Comparison between Power Spectral Subtraction Method and Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise

By

Sravan Kumar Buggaveeti Javier Perez Ramirez

Srinivasu Rao Pudi Yousef Jaradat

Professor

Dr. Charles Cruesere

Guideline

- Introduction
- Power Spectral Subtraction (PSS)
- Multi-Band Spectral Subtraction (MBSS)
- Implementation
- Tests and Results
- Conclusions
- Demo

Introduction

- Mobile communications => Speech Signal => Corrupted by noise
- Remove the noise is difficult because the random nature of it
- There is always a tradeoff between the amount of noise removed and the distortion of the speech signal
- Spectral Subtraction techniques were developed: In this project, PSS and MBSS were studied

Power Spectral Subtraction (PSS)

- Based on the subtraction of noise (power noise spectrum) from the speech signal
- PSS estimates the noise over the silence periods or the initial silence of the speech signal
- Over-subtraction factor α is used to improve the result
  - α depends on the segmental SNR
- After the subtraction, negative values could be obtained => The speech signal in this case is floored.
- This method assumes that the noise present in the speech signal is uniformly distributed in the spectrum => This assumption fails when real noise is considered

Multi Band Spectral Subtraction (MBSS)

- Similar to PSS
- Again, an estimation of the noise is subtracted from the speech signal
- Over-Subtraction α and flooring process of the enhanced speech signal are included
- New coefficient added: δ
    - δ depends on the frequency

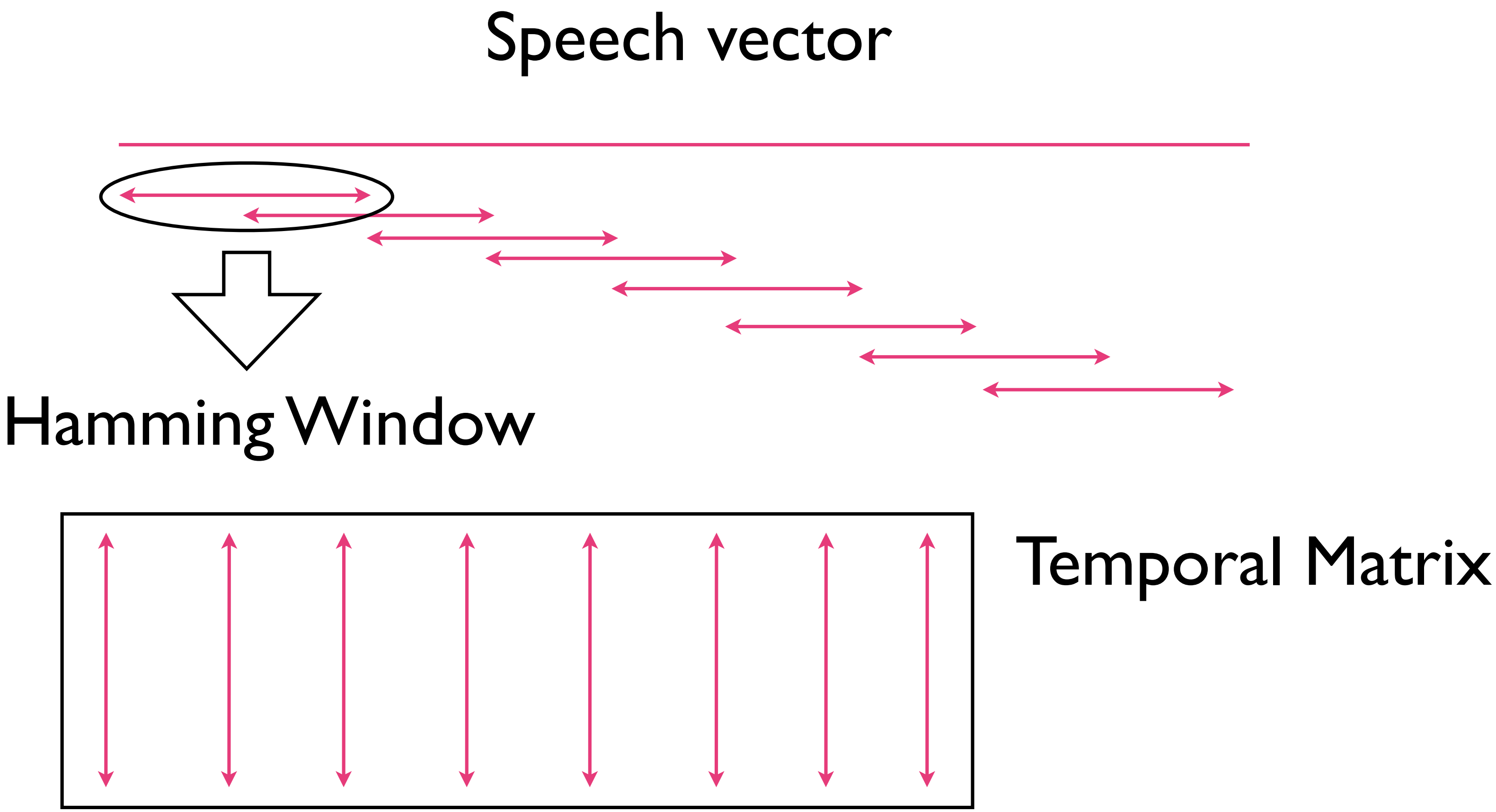
Implementation

- Pre-Processing
- PSS implementation
- MBSS implementation
- Signal Reconstruction

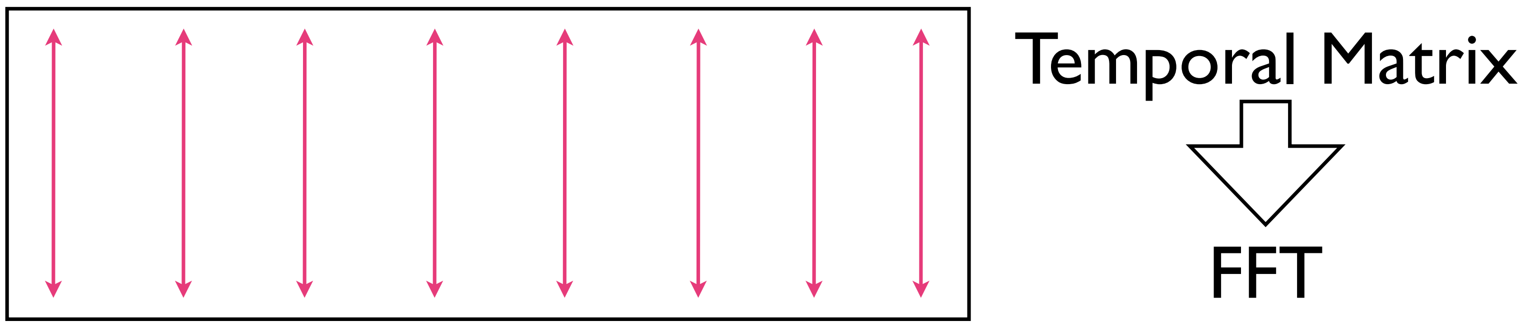
Pre-Processing

- Division of the audio vector into overlapped frames
- Hamming window
- FFT of every frame
- Weighted Spectral Average

Pre-Processing

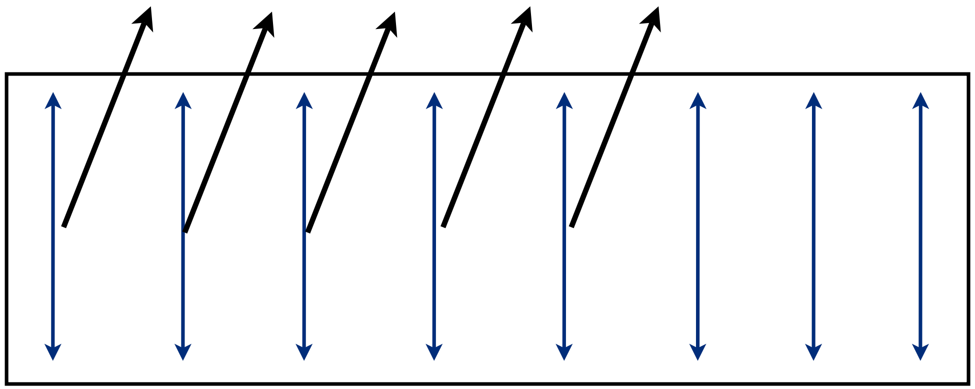


Pre-Processing



Pre-Processing

Weighted Spectral Average



FFT Matrix

PSS implementation

- First frame is used as Estimated Noise
    - With this frame, a threshold for detecting the presence of speech signal is calculated. If the segmental SNR is lower than the threshold, the estimated noise is updated.
- Noise is subtracted from the speech signal by using the following expression:

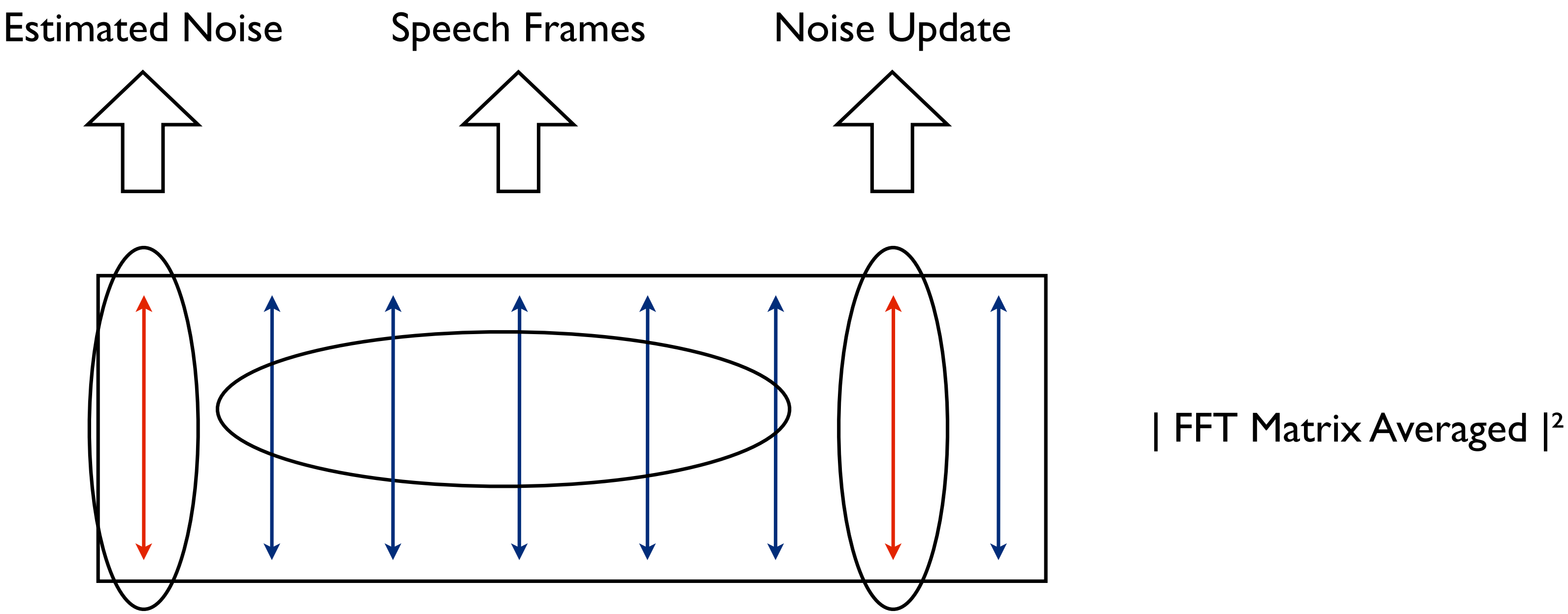
$$|\ddot{S}(k)|^2 = |Y(k)|^2 - \alpha|\ddot{D}(k)|^2$$

PSS implementation

$$SNR_i(dB) = 10log_{10}\left(\frac{\sum_{k=b_i}^{e_i} |Y_i(k)|^2}{\sum_{k=b_i}^{ei} |\hat{D}_i(k)|^2}\right)$$

$$\alpha_i = \begin{cases} 5, & SNR_i < -5 \\ 4 - \frac{3}{20}(\ SNR_i), & -5 \leq SNR_i \leq 20 \\ 1, & SNR_i > 20 \end{cases}$$

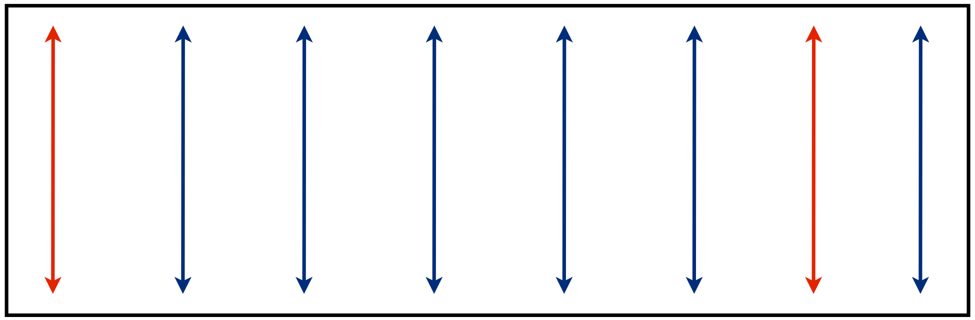
PSS implementation



PSS implementation

- x α = Enhanced Frame

$$|\ddot{S}(k)|^2 = |Y(k)|^2 - \alpha|\ddot{D}(k)|^2$$



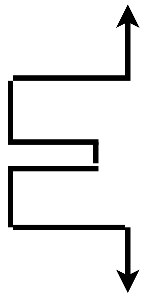
| FFT Matrix Averaged $|^2$

MBSS implementation

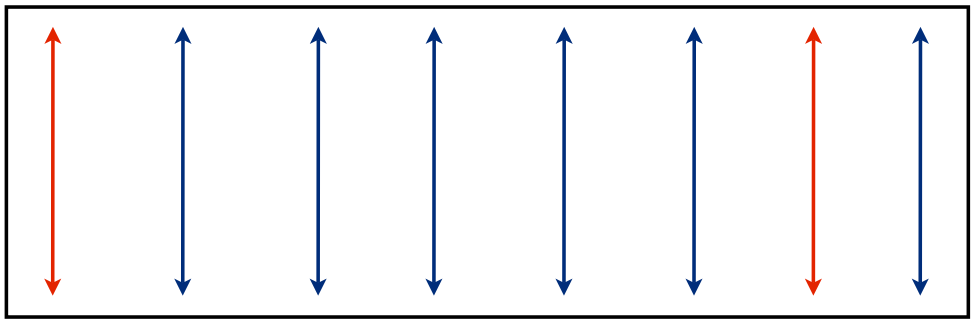
- Similar to PSS
- Now, the spectrum is divided into bands (linear spacing approach)
- A new coefficient is considered: δ
  - δ is a vector defined by:

$$\delta_i = \begin{cases} 1, & f_i < 1\ kHz \\ 2.5, & 1\ kHz < f_i \leq \frac{F_s}{2} - 2\ kHz \\ 1.5, & f_i > \frac{F_s}{2} - 2\ kHz \end{cases}$$

MBSS implementation

\- x  x α = Enhanced Frame

$$|\widehat{Si}(k)|^2 = |Y_i(k)|^2 - \alpha_i\,\delta_i|\widehat{D}_i(k)|^2$$



$|$ FFT Matrix Averaged $|^2$

MBSS implementation

$$|\widehat{S}\imath(k)|^2 = |Y_i(k)|^2 - \alpha_i\, \delta_i |\widehat{D}_i(k)|^2$$

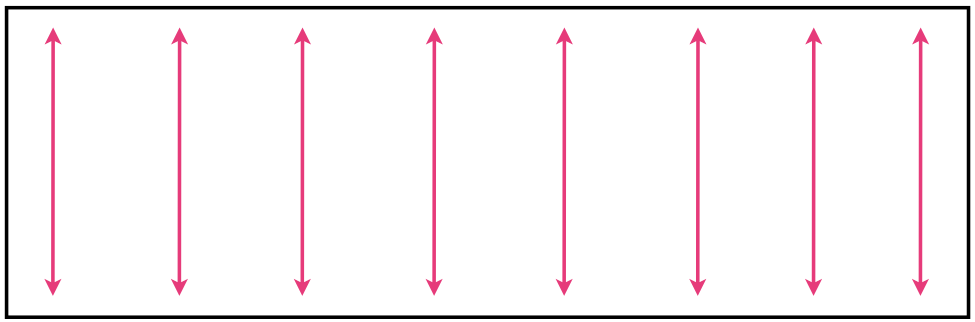
$$|\widehat{S}_i(k)|^2 = \begin{cases} |\widehat{S}_i(k)|^2 & |\widehat{S}_i(k)|^2 > 0 \\ \beta |Y_i(k)|^2 & else \end{cases}$$

Enhanced Speech Signal is floored before reconstruction
β = Flooring Factor

Signal Reconstruction

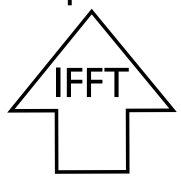
Speech vector

Enhanced Speech vector



Overlap Add Method

New Temporal Matrix



| FFT Matrix Enhanced $|^2$

Tests and Results

- Objective Test
- Subjective Test
- Results

Objective Test

- Perceptual Evaluation of Speech Quality test methodology (PESQ) was used as the method to evaluate the quality of the enhanced speech signal
- PESQ requires the clean and the enhanced speech file
- PESQ scores
  - 0 => When clean and enhanced signals are totally different
  - 4.5 => When clean and enhanced signals are the same

Subjective Test

- Recommendation ITU-P.835 was followed
- 8 subjects were tested
- Analysis of Variance (ANOVA) was used to analyze the data gathered
- The next templates were used for the subjective evaluation of the enhanced speech signal:
    - Speech Signal rating scale
    - Background noise rating scale
    - Overall quality ranking scale

Speech signal rating scale

Attending **ONLY to the SPEECH SIGNAL**, select the category which best describes the sample you just heard.

The **SPEECH SIGNAL** in this sample was

5 – NOT DISTORTED

4 – SLIGHTLY DISTORTED

3 - SOMEWHAT DISTORTED

2 – FAIRLY DISTORTED

1 – VERY DISTORTED

Background noise rating scale

Attending **ONLY to the BACKGROUND**, select the category which best describes the sample you just heard..

The **BACKGROUND** in this sample was

5 – NOT NOTICEABLE

4 – SLIGHTLY NOTICEABLE

3 – NOTICEABLE BUT NOT INTRUSIVE

2 – SOMEWHAT INTRUSIVE

1 – VERY INTRUSIVE

Overall quality rating scale

Select the category which best describes the sample you just heard for purposes of everyday speech communication.

The **OVERALL SPEECH SAMPLE** was

5 – EXCELLENT

4 – GOOD

3 – FAIR

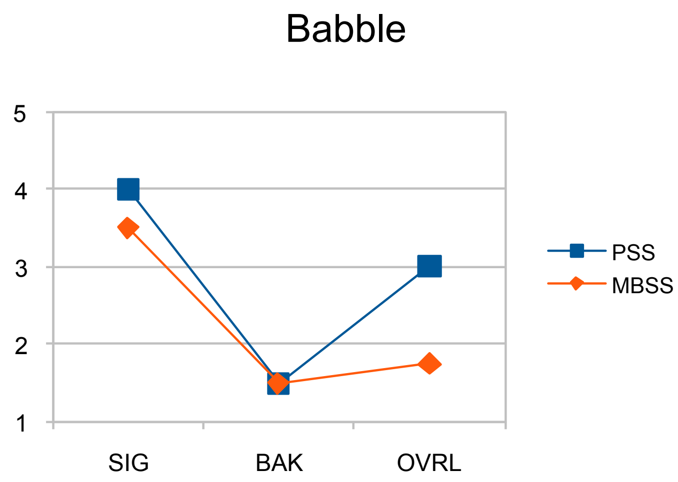
2 – POOR

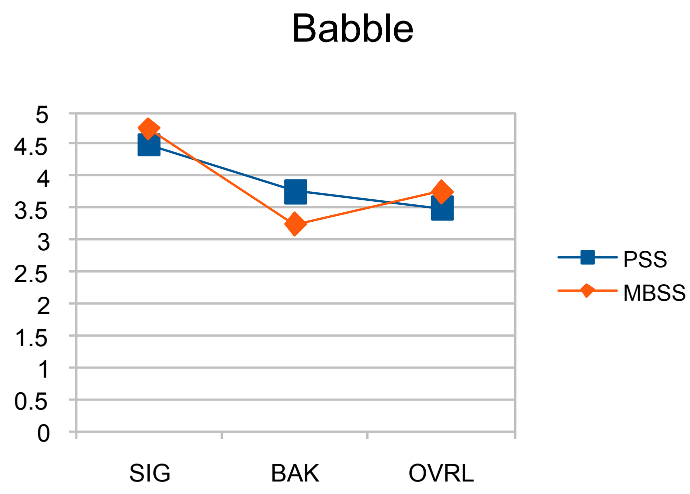
1 – BAD

ITU-P.835

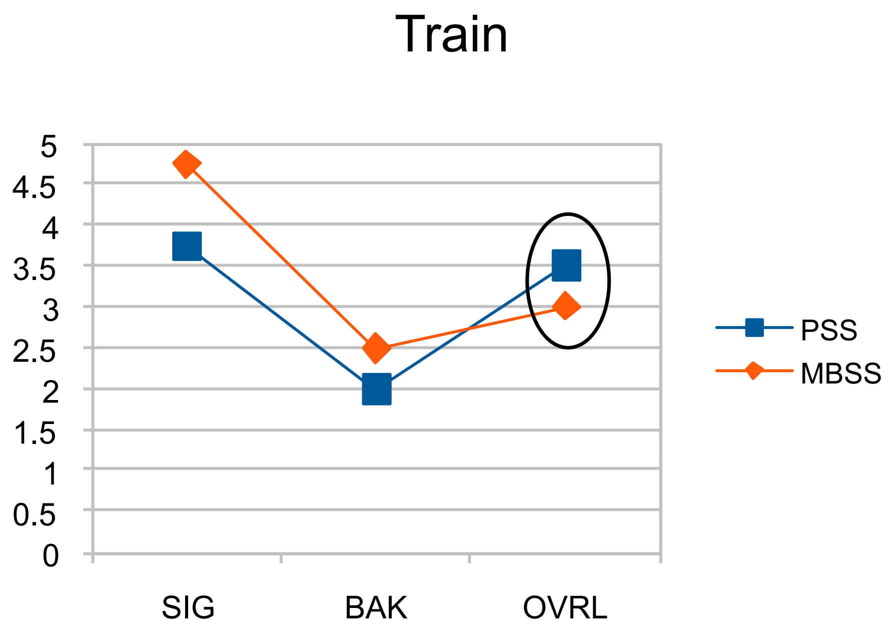
Results

- PESQ Scores for different situations (babble & train)
- Subjective results for different situations (babble & train)
- Parameters:
  - F_s = 8 KHz
  - δ => 8 Bands
  - β = 0.01
  - Hamming Window Size = 20 ms.
  - Speech file length = 2 s.

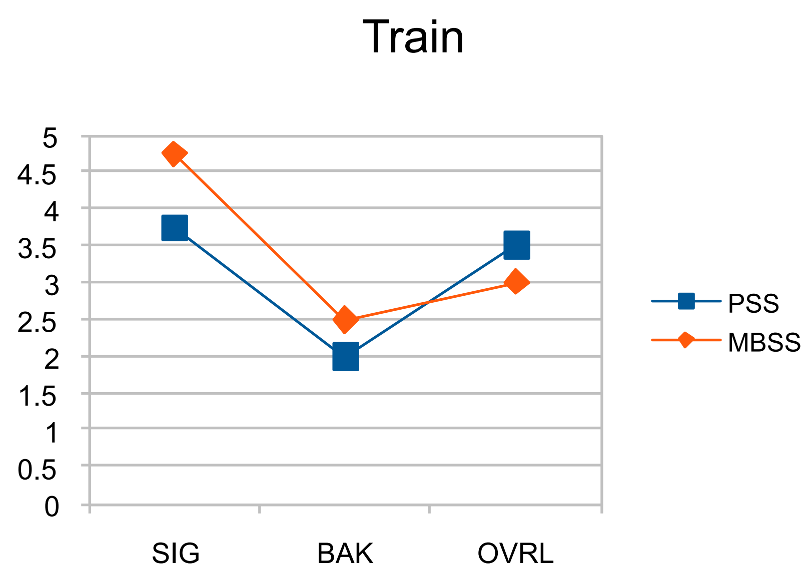


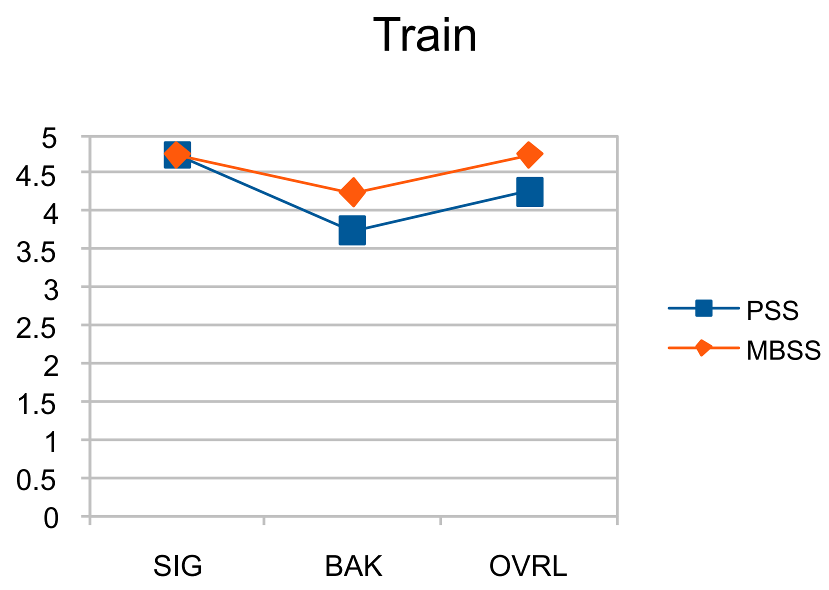


ITU-P.835. Babble (5dB & 10dB)

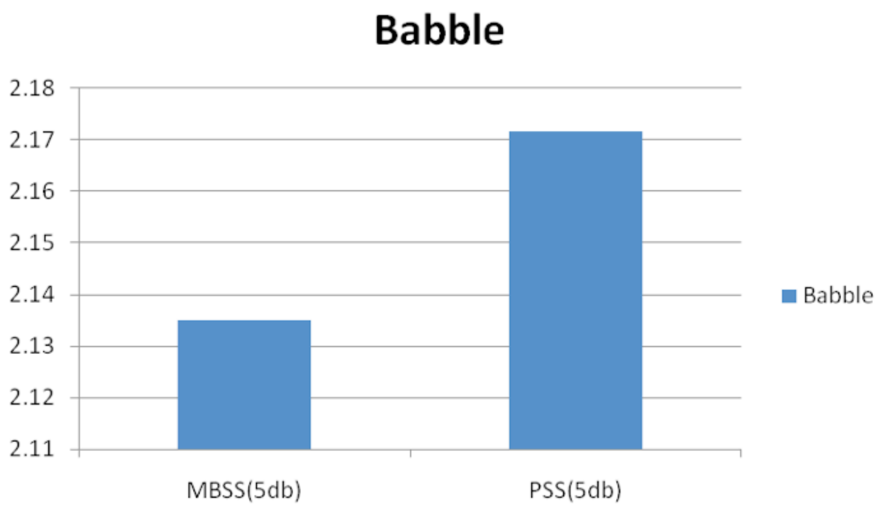


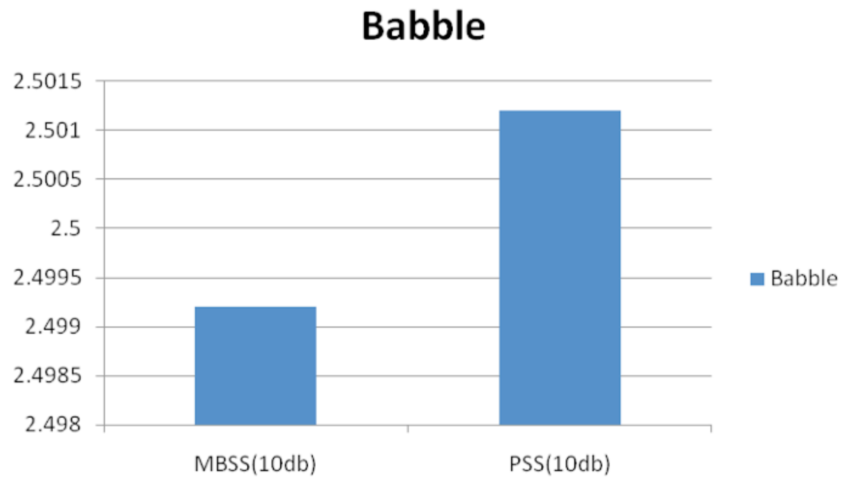
ITU-P.835. Train (5dB)



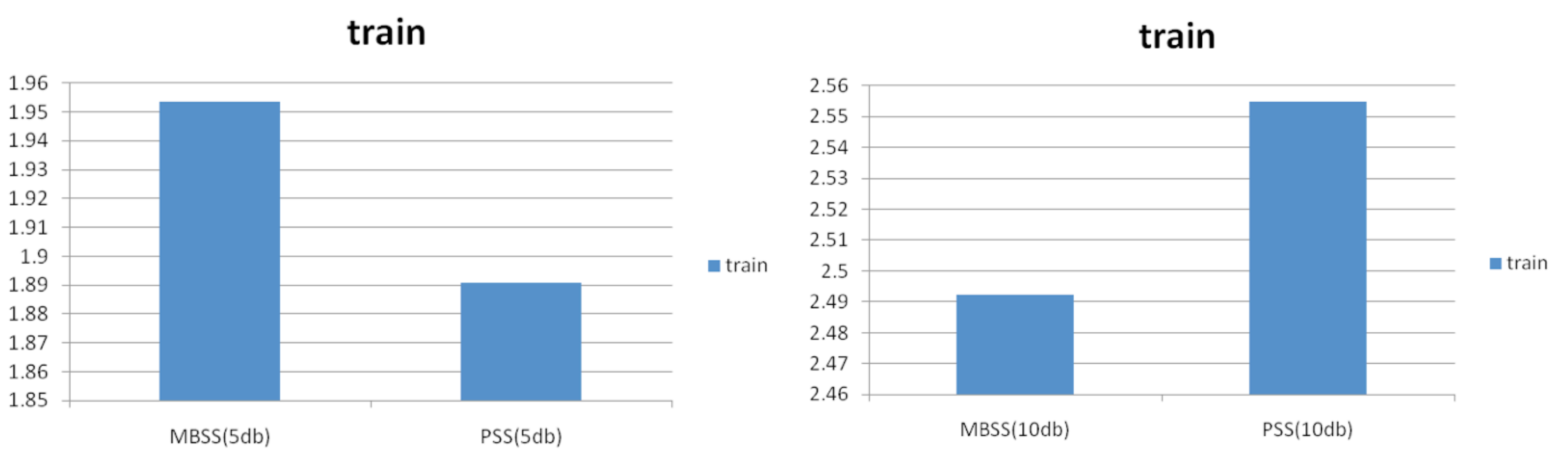


ITU-P.835. Train (5dB &10dB)

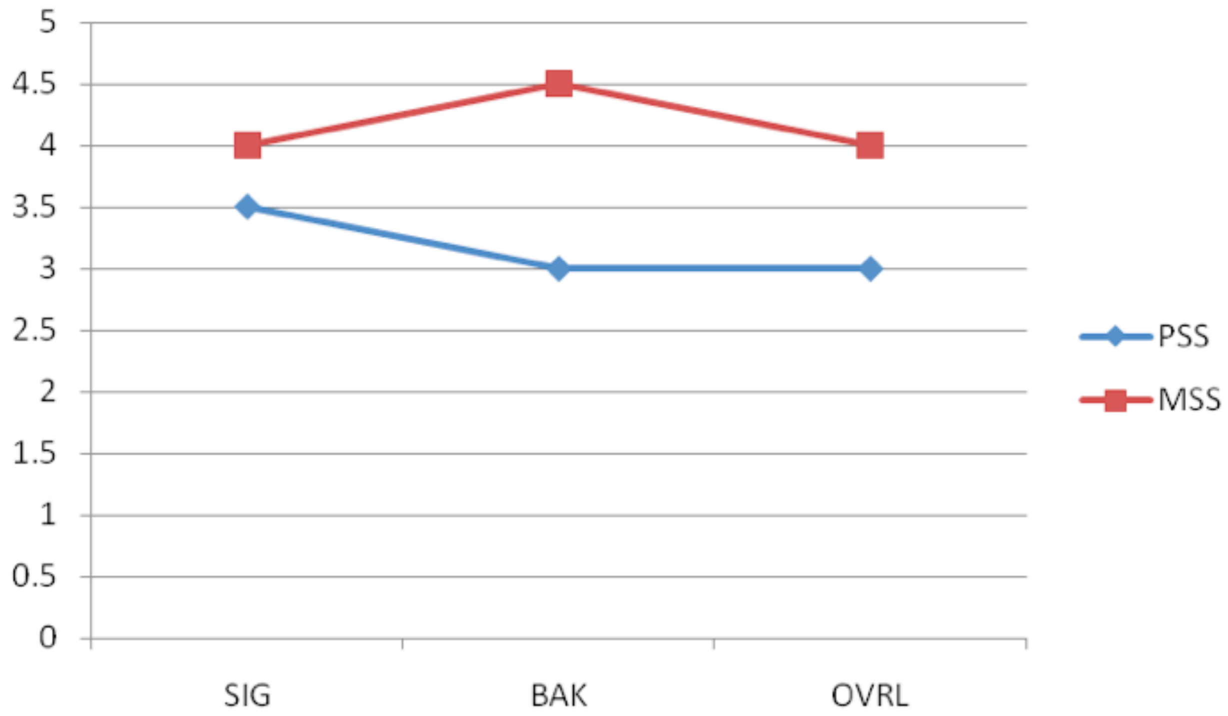




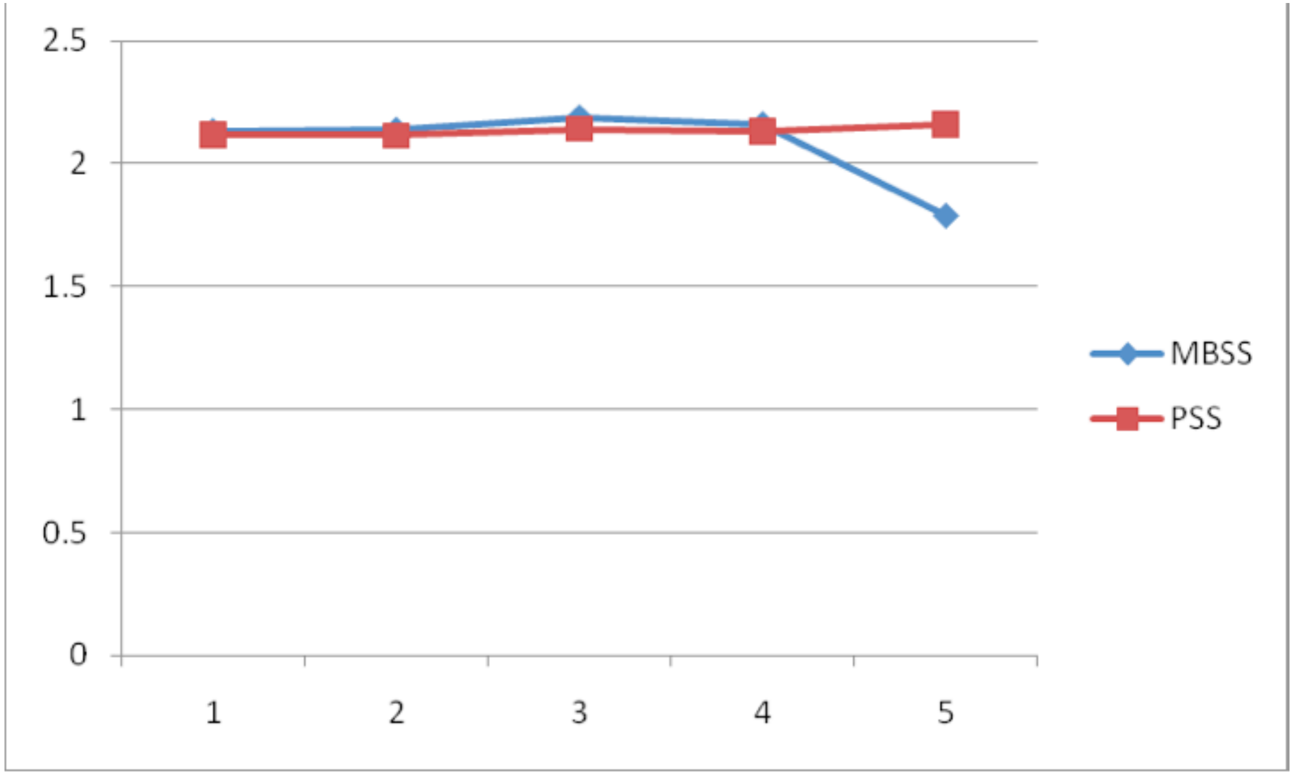
PESQ. Babble (5dB & 10 dB)



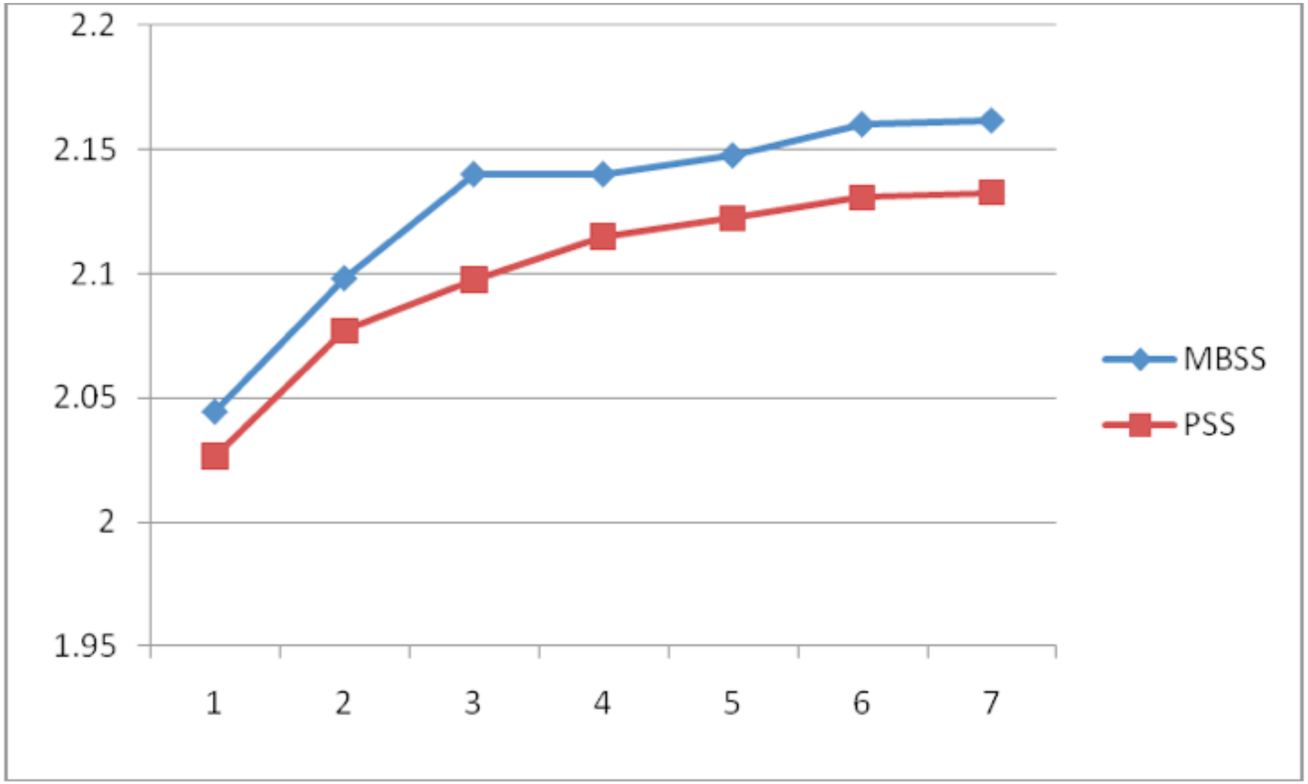
PESQ. Train (5dB & 10dB)



ITU-P.835. Colored Noise. Long Speech (5dB)



PESQ. Hamming Size [0.01-0.5] s
Babble. Longer speech.



PESQ. Beta [0.25-0.001]
Babble .Longer speech

Conclusions

- Generally speaking, both methods perform the same
- Subjective test: PSS performs better (5 dB). MBSS performs better (10 dB).
- Objective test (short speech files): PSS performs better
- Objective test (large speech files): MBSS performs better for all β and small-medium Hamming Window size

Major concepts we have learnt

- Process speech files
- Speech segmentation (Hamming Window & Frame Overlapping)
- Spectral processing (FFT)
- Weighted Spectral Average
- PSS & MBSS techniques
- Signal Reconstruction (IFFT & Overlap - Add)
- Objective & Subjective Testing
- Data analysis (ANOVA)

Questions?