Technion
Israel Institute of Technology

Dept. Electrical Engineering

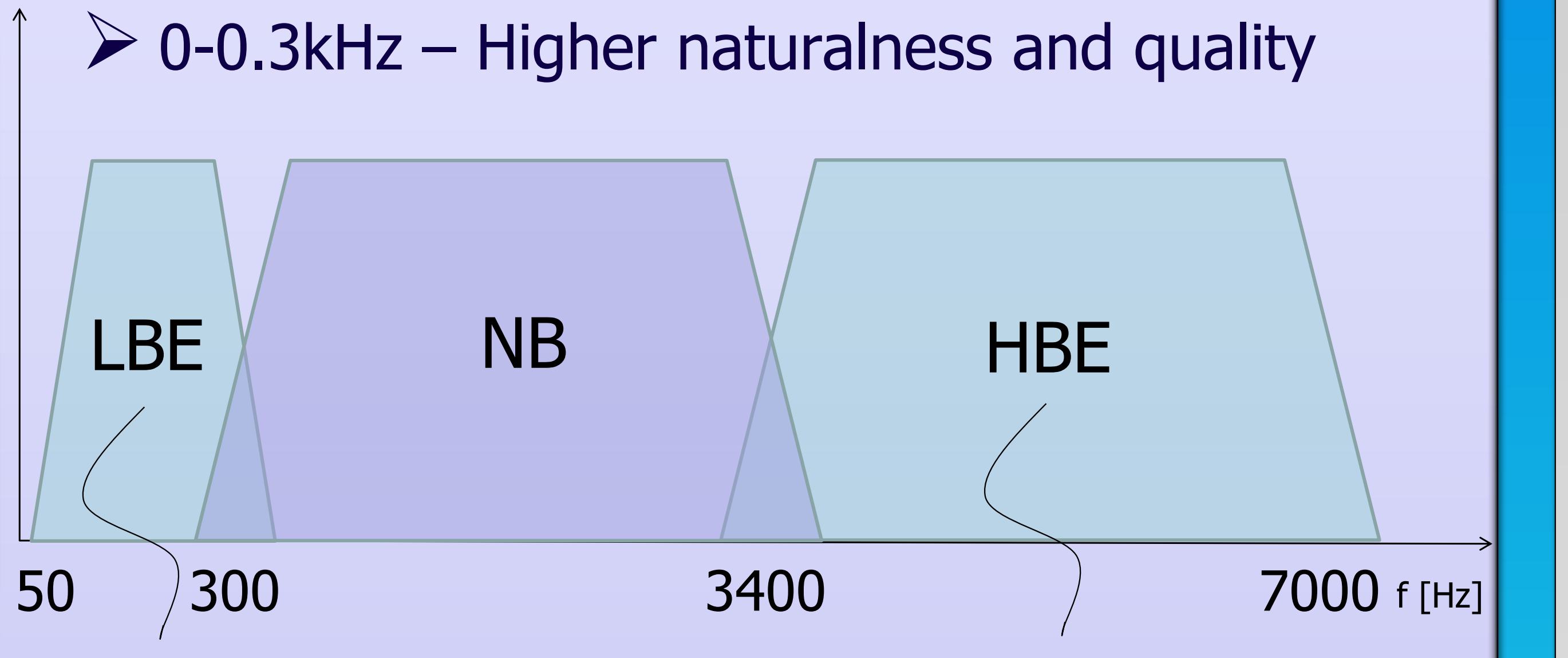
EVALUATION OF A SPEECH BANDWIDTH EXTENSION ALGORITHM BASED ON VOCAL TRACT SHAPE ESTIMATION

Itai Katsir, David Malah and Israel Cohen

SIPL
Signal and Image Processing lab

1. Introduction

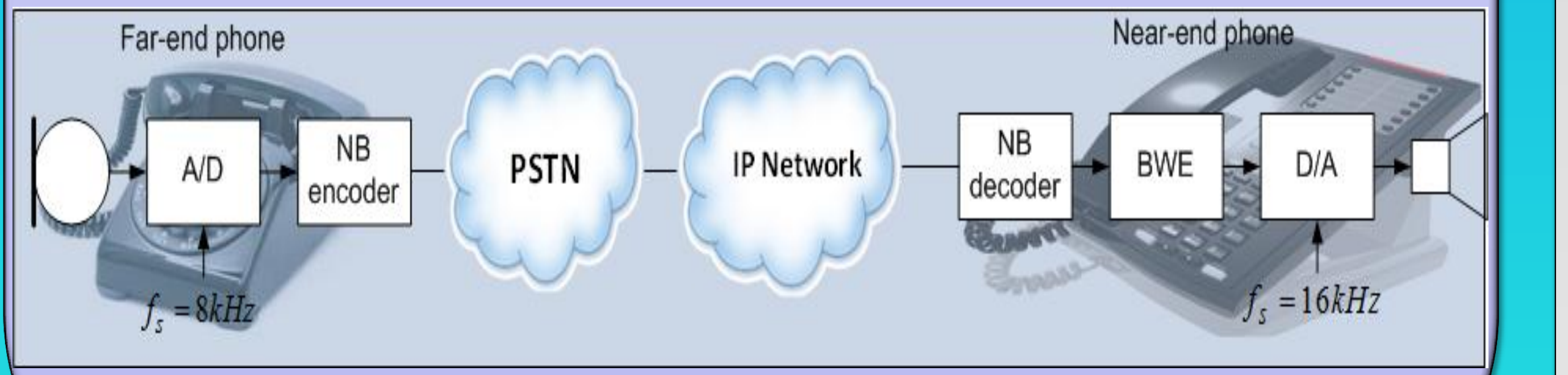
- **Motivation:** Growing consumer demand for high quality, wideband, speech communication of frequency range of 50-7000Hz.
- **Problem:** Reduced quality, narrowband, speech communication of frequency range of 300-3400Hz due to band limitation of analog telephones and traditional speech communication networks.
- **Solution:** Artificially extend speech bandwidth to achieve speech quality enhancement
  - ➢ 3.4-7kHz – Higher intelligibility and quality
  - ➢ 0-0.3kHz – Higher naturalness and quality



+0.6 MOS (theoretical)
Naturalness

+0.6 MOS (theoretical)
Intelligibility

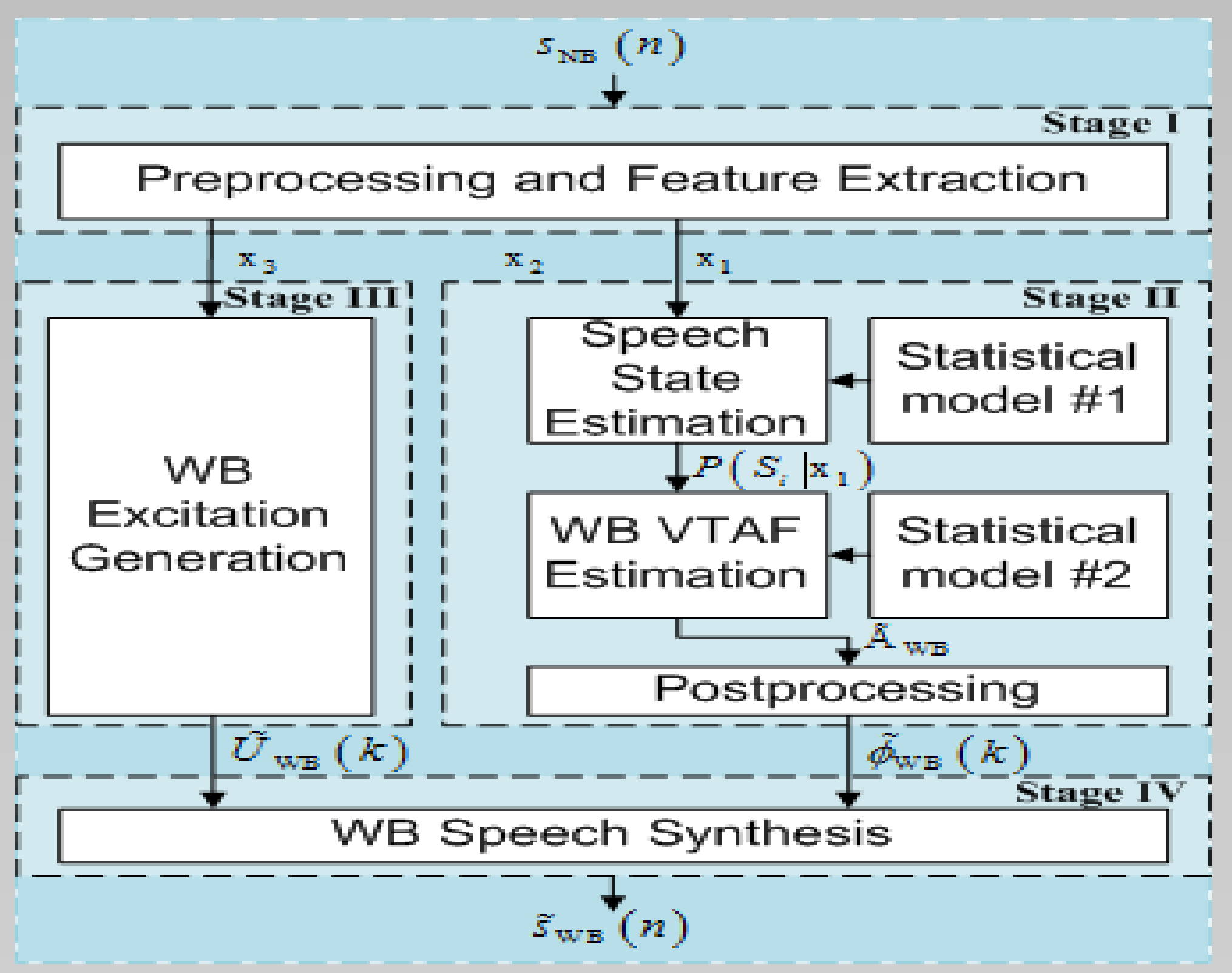
- **Application:** In the transition time to full WB communication networks, BWE can be used in mix NB-WB communication networks.



HD VoIP
Sounds Better

2. BWE Algorithm Overview

- **General Block Diagram:**



- **Preprocessing and Feature Extraction**
  - ➢ Upsampling and equalization by 10dB boost at 300Hz.
  - ➢ x_1 – Frequency-based features for speech state estimation.
  - ➢ x_2 – NB VTAF for WB VTAF estimation.
  - ➢ x_3 – NB excitation for WB excitation generation.
- **WB Spectral Envelope Estimation**
  - ➢ **Speech State Estimation:** making a decision on current frame state (phoneme) by maximizing the a-posteriori PDF:

$$p\left(S_i(m)\middle|\mathrm{X}_1(m)\right)=p\left(\mathrm{x}_1(m)\middle|S_i(m)\right)\cdot\sum_{j=1}^{N_s}p\left(S_i(m)\middle|S_j(m-1)\right)p\left(S_j(m-1)\middle|\mathrm{X}_1(m-1)\right)$$

  - ➢ **WB VTAF Estimation:** Finding closest WB VTAF to extracted NB VTAF using Euclidean distance:

$$\tilde{\mathrm{A}}_{\mathrm{WB}}^{S_i}=\mathrm{A}_{\mathrm{WB}}^{S_i}\left(j^{opt}\right)$$

$$j^{opt}=\arg\min_{j=1}^{N_{\mathrm{CB}}}\left\|\log\left(\mathrm{A}_{\mathrm{NB}}\right)-\log\left(\mathrm{A}_{\mathrm{WB}}^{S_i}(j)\right)\right\|_2^2$$

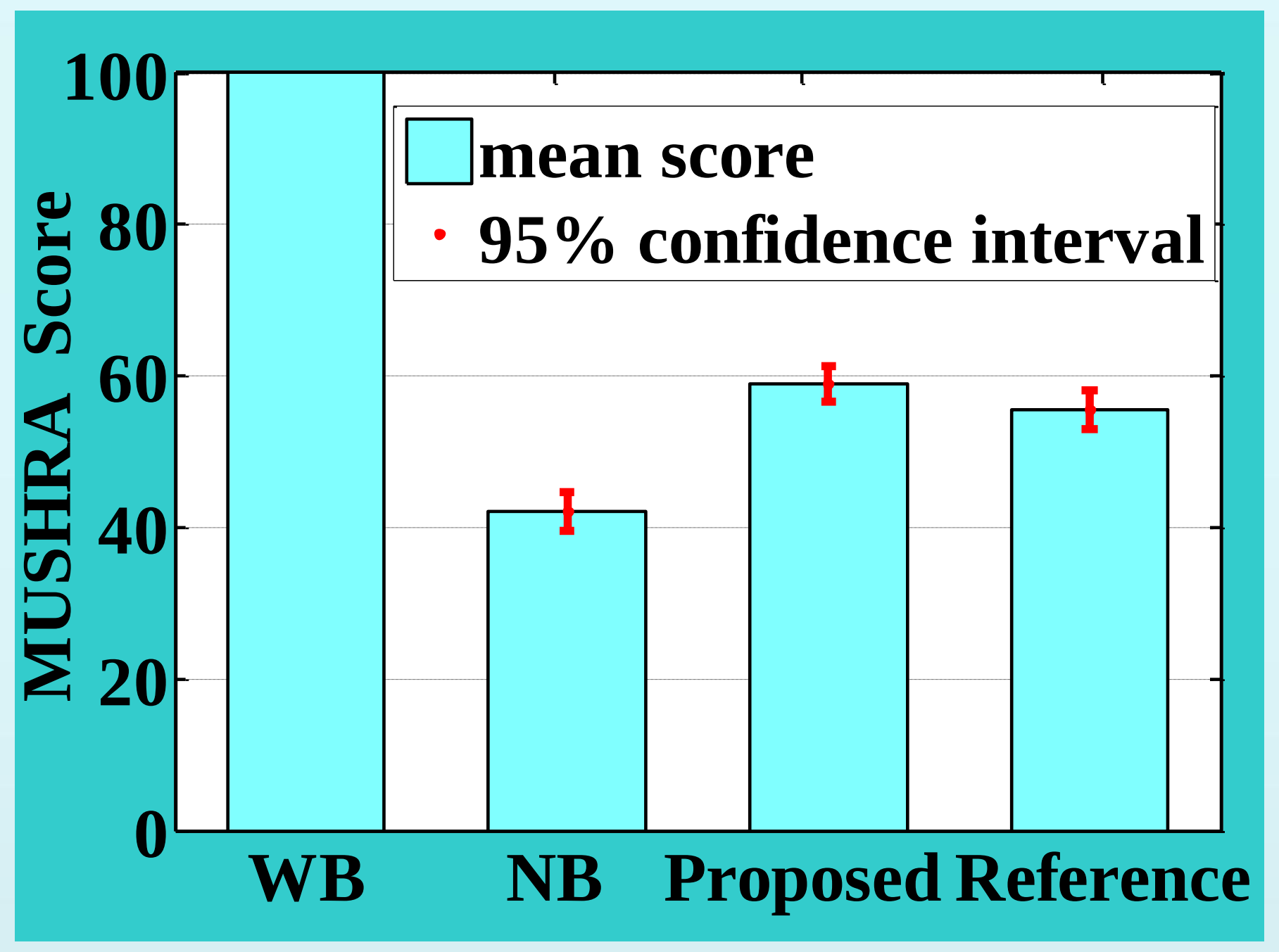
  - ➢ **Postprocessing:** Estimated WB envelope fit to NB envelope by formant frequencies tuning of estimated WB VTAF to allow better gain adjustment to NB envelope. Iterative tuning by VTAF perturbation based on the sensitivity function:

$$\frac{\Delta f_{n_f}}{f_{n_f}}=\sum_{n_A}^{N_A}S_{n_f,n_A}\frac{\Delta A_{n_A}}{A_{n_A}}$$

- **WB Excitation Generation:** HB excitation generation using spectral copy of the NB excitation.
- **WB Speech Synthesis:** Frequency domain synthesis using the estimated spectral envelope and excitation.

3. Performance Evaluation

- **Subjective Evaluation:** MUltiStimulus test with Hidden Reference and Anchor (MUSHRA) - Ranks the processed speech test sentences in comparison to a reference sentence for score between 0-100.
  - ➢ 11 listeners, 6 different experiments, each with different English sentence.
  - ➢ Every experiment included multiple conditions of the speech sentence:
    - ▪ WB reference signal
    - ▪ NB anchor signal
    - ▪ Proposed BWE signal
    - ▪ Reference BWE signal
  - ➢ result illustrates the advantage of using the proposed BWE algorithm when using traditional NB telephone networks.



- **Objective Evaluation:** Average Spectral Distortion Measure (SDM) and Log Spectral Distance (LSD) of estimated spectral envelope with and without the iterative postprocessing step.

$$\mathrm{SDM}_m=\frac{1}{k_{\mathrm{high}}-k_{\mathrm{low}}+1}\sum_{k=k_{\mathrm{low}}}^{k_{\mathrm{high}}}\xi_m\ \ k$$

$$\xi_m\ \ k\ =\begin{cases}\Delta_m\ \ k\ \ e^{\alpha\Delta_m\ \ k\ -\beta k},\ \text{if}\ \Delta_m\ \ k\ \ \geq 0\\ \ln\ -\Delta_m\ \ k\ +1\ e^{-\beta k},\ \text{else}\end{cases}$$

$$\Delta_m\ \ k\ =10\log_{10}\frac{\tilde{\phi}_m\ \ k}{\phi_m\ \ k}$$

Measured	SDM [dB]	LSD[dB]
Without iterative process	13.64	9.98
With iterative process	9.89	9.91

  - ➢ result illustrates the effectiveness of the iterative tuning process in reducing estimation artifacts and especially in reducing gain over-estimation
- **Complexity Evaluation:** Average processing time of a 20 msec speech frame of main BWE algorithm processing blocks.

Algorithm Processing Block	Computation Time [msec]
Preprocessing and feature extraction	1.27
State estimation	19.39
WB VTAF estimation	0.59
Postprocessing (iterative process)	7.69
Postprocessing (gain adjustment)	0.36
WB excitation generation	0.04
WB speech synthesis	0.57
Total	29.91

  - ➢ result illustrate the high complexity of the phoneme estimation step and the iterative processing block of the postprocessing step