GitHub repository:
(presentation, arXiv link)

Speech Enhancement Based on Reducing the Detail Portion of Speech Spectrograms in Modulation Domain via Discrete Wavelet Transform

Authors: *Shih-kuang Lee*
Syu-siang Wang
Yu Tsao
Jeih-weih Hung

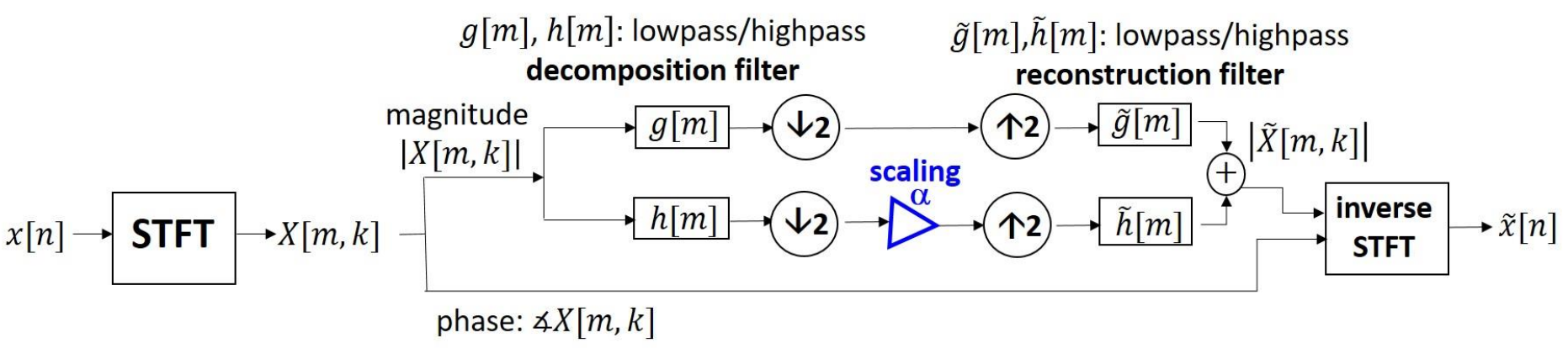
Presenter: *Shih-kuang Lee*

National Chi Nan University, Taiwan
Academia Sinica, Taiwan

A brief summary of this study

- We use **discrete wavelet transform (DWT)** to process the **temporal part of spectrograms**
  - in order to **highlight the human speech part**

- The experimental results show that the presented method **gives better speech quality**



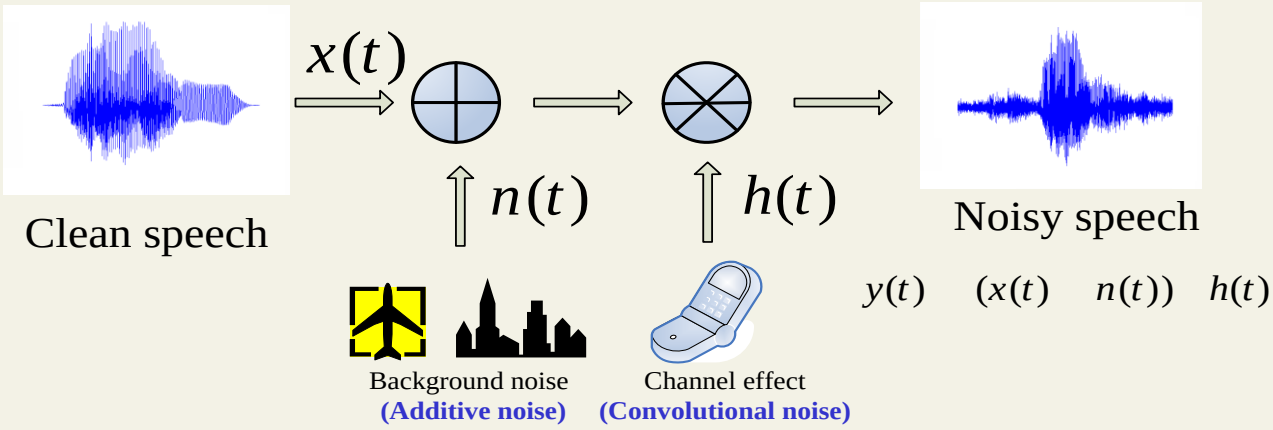
The flowchart of ModWD

Outline

- Introduction
- Proposed Method
- Experimental Setup
- Experimental Results and Discussions
- Conclusions

Introduction (1/2)

- **Speech Enhancement**
  to improve the quality and intelligibility of degraded speech [1]
- **Types of Noise**
  - Additive noise (independent of speech)
    - ✓ Stationary
    - ✓ Non-stationary
  - Convolutional noise (dependent on speech)



[1] Loizou, Philipos C. Speech Enhancement, 2nd Edition. CRC Press, 20130228.

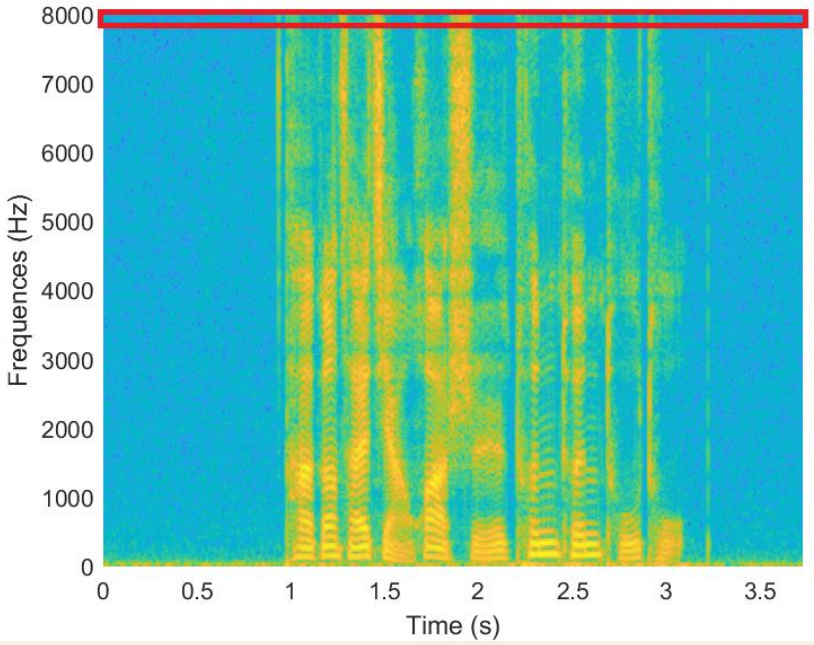
Introduction (2/2)

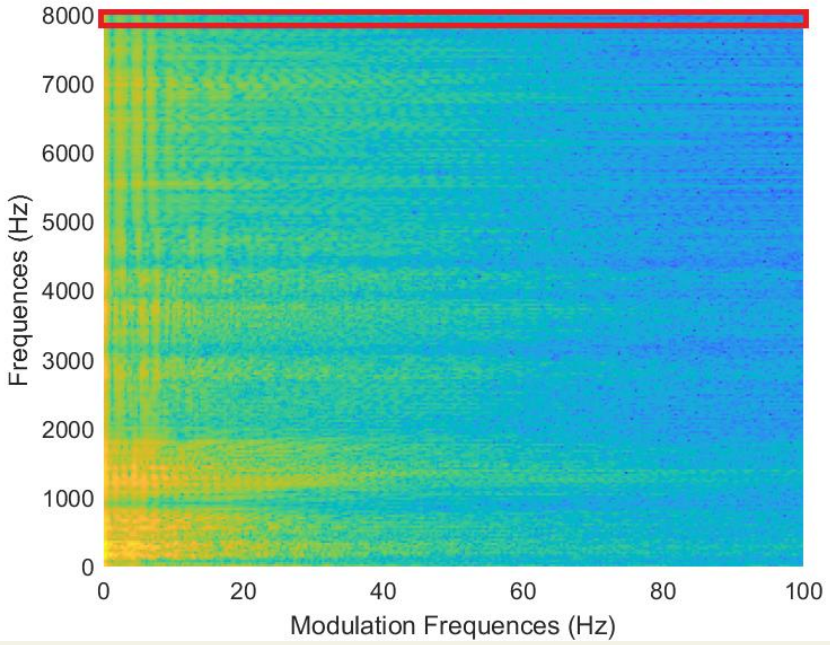
☐ One aspect to enhance the speech part in the noisy signal:

- **taking advantage of human speech characteristics**

  ✓ The linguistic information in human speech is mainly located at the modulation frequencies **within the range 1 – 16 Hz**, dominant **at 4 Hz**.

Spectrograms plot in log scale





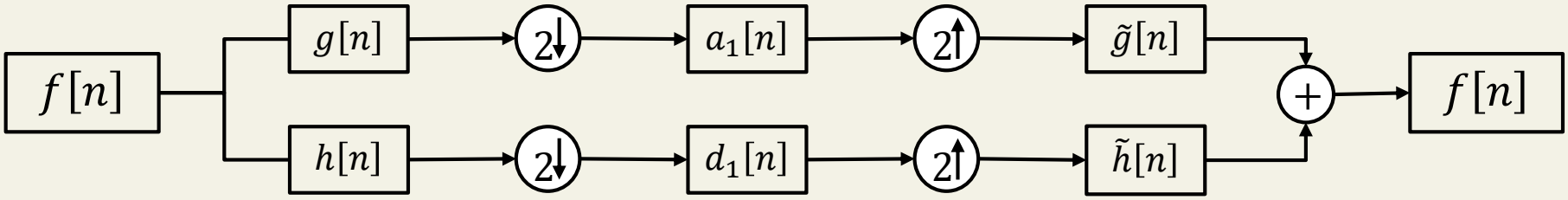
Proposed Method (1/3)

Modulation-domain Wavelet Denoising (ModWD)

- Basic idea:
  using wavelet transform to extract the important modulation spectral component in speech spectrogram
  - the scheme of a one-level **DWT** and **IDWT**

$g[n], h[n]$: *lowpass/highpass* ***decomposition*** *filter*

$\tilde{g}[n], \tilde{h}[n]$: *lowpass/highpass* ***recomposition*** *filter*

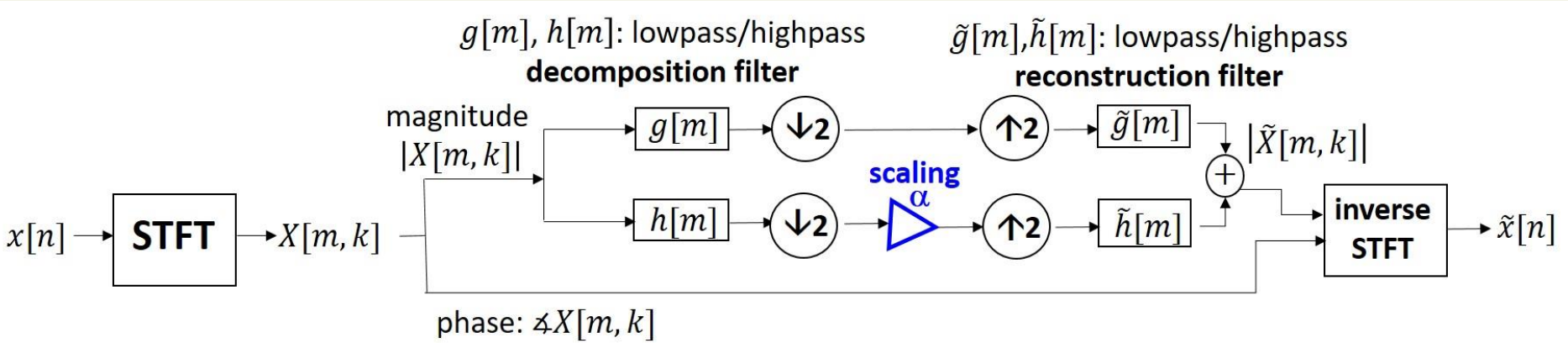


1-level Discrete Wavelet Transform (DWT)

1-level Inverse Discrete Wavelet Transform (IDWT)

Proposed Method (2/3)

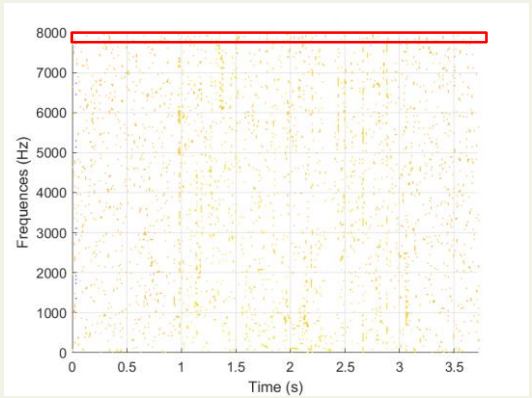
- Modulation-domain Wavelet Denoising (ModWD)
  - The main step:
    deemphasizing the detail part (high-pass filtered part) of the spectrogram in the modulation domain
- Scaling Factor α for the detail part
  - 0 totally discard the detail component
  - 0.25
  - 0.5
  - 0.75

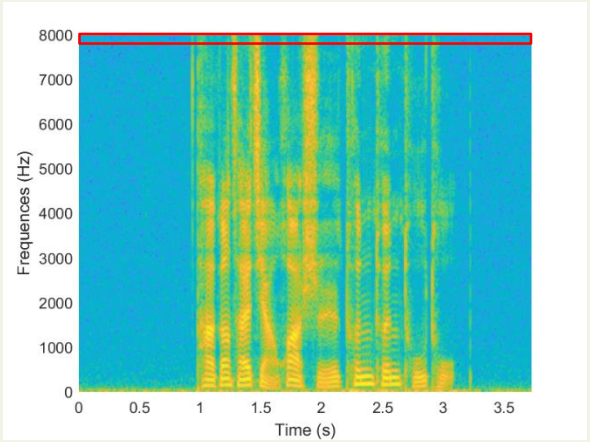


Proposed Method (2/3)

- **Modulation-domain Wavelet Denoising (ModWD)**
  - **demonstration**

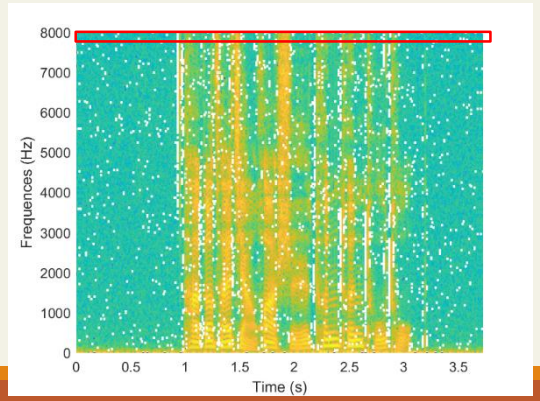
Detail Coefficients



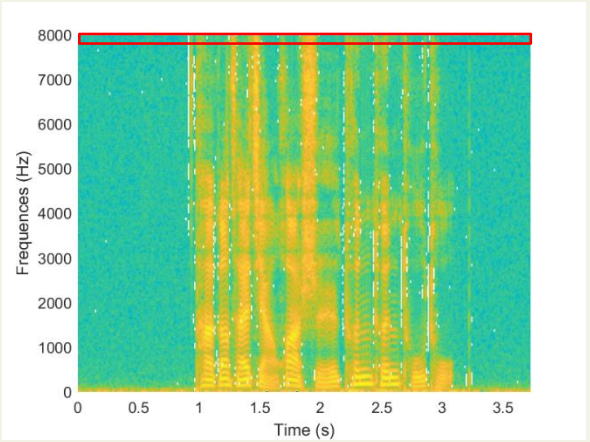


Original Spectrogram

α=0



Approximation Coefficients



Reconstructed Spectrogram

Experimental Setup (1/1)

Speech data

Database: **Aurora-2** (Sampling Rate: **8** kHz)
(**English** digit utterances generated by both female and male speakers)
Utterances: **50** utterances belonging to a single speaker were used.
Noise: airport noise
SNRs: from **0** dB to **20** dB, with a step of **5** dB

Speech enhancement setup

20 ms and **10** ms were employ as frame **size** and frame **rate**
The **biorthogonal 3.7** wavelet basis was used for the **DWT** and **IDWT**
The scaling factor **α** used in ModWD was set to 0, 0.25, 0.5 and 0.75
Bandwidth of Modulation Frequencies: **50** Hz ((1/10ms)/2)

Objective evaluation metric

Perceptual estimation of speech quality (**PESQ**) was used

Experimental Results and Discussions (1/4)

☐ **PESQ** result with **ModWD** only

- Improved PESQ scores at a lower **α in most cases**
  - setting $\alpha = 0.25$ yields the best PESQ results in average.
  - the case of $\alpha = 0$ (discarding detail portion) also improves the PESQ

Table: PESQ results for ModWD with different assignments of the scaling factor α

SNR	0	5	10	15	20	Avg.
Baseline	1.300	1.768	2.060	2.391	2.780	2.060
$\alpha = 0.75$	1.307	1.779	2.070	2.404	2.789	2.070
$\alpha = 0.50$	1.319	1.788	2.078	2.415	2.795	2.079
$\alpha = 0.25$	1.322	1.794	2.083	2.422	2.797	2.084
$\alpha = 0$	1.289	1.798	2.086	2.428	2.797	2.079

The values in red represent the best score. The higher, the better.

Experimental Results and Discussions (2/4)

- *PESQ results for various SE methods including ModWD with* $\alpha = 0$

SNR	0	5	10	15	20	Avg.
ModWD	1.289	1.798	2.086	2.428	2.797	2.079
SS	1.489	2.025	2.341	2.668	3.007	2.306
WF	1.723	2.126	2.412	2.726	2.983	2.394
STSA	1.700	2.141	2.430	2.724	2.974	2.393
logSTSA	1.840	2.234	2.523	2.802	3.058	2.491
ModWD-SS	1.479	2.029	2.366	2.688	3.014	2.315
SS-ModWD	1.500	2.048	2.317	2.694	2.998	2.311
ModWD-WF	1.676	2.079	2.409	2.720	2.967	2.370
WF-ModWD	1.772	2.133	2.441	2.737	2.963	2.409
ModWD-STSA	1.696	2.156	2.449	2.749	2.993	2.409
STSA-ModWD	1.735	2.169	2.445	2.741	2.977	2.413
ModWD-logSTSA	1.813	2.227	2.520	2.819	3.076	2.491
logSTSA-ModWD	1.842	2.240	2.519	2.810	3.054	2.493

The values in red represent the best score. The higher, the better.

Experimental Results and Discussions (3/4)

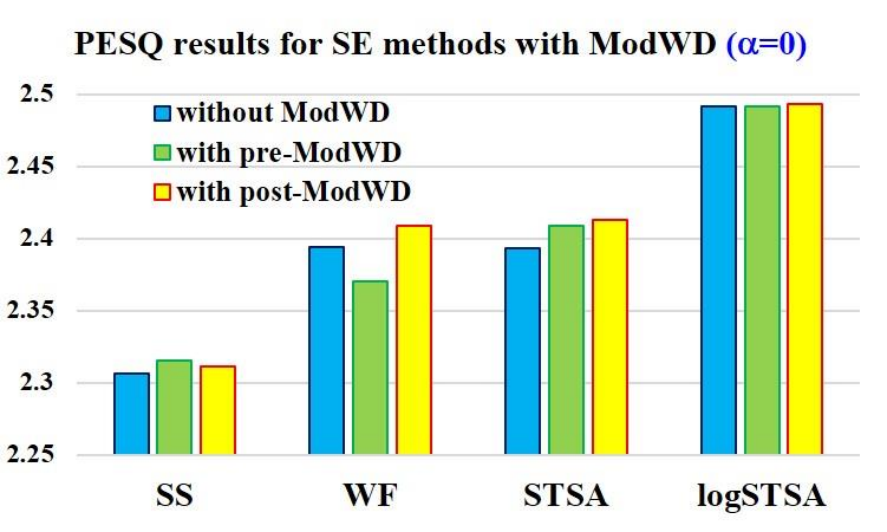
- *PESQ results for various SE methods including ModWD with* $\alpha = 0.25$

SNR	0	5	10	15	20	Avg.
ModWD	1.322	1.798	2.086	2.428	2.797	2.079
SS	1.489	2.025	2.341	2.668	3.007	2.306
WF	1.723	2.126	2.412	2.726	2.983	2.394
STSA	1.700	2.141	2.430	2.724	2.974	2.393
logSTSA	1.840	2.234	2.523	2.802	3.058	2.491
ModWD-SS	1.496	2.025	2.341	2.668	3.007	2.308
SS-ModWD	1.502	2.047	2.364	2.691	3005	2.322
ModWD-WF	1.697	2.087	2.414	2.721	2.988	2.381
WF-ModWD	1.768	2.133	2.445	2.742	2.975	2.413
ModWD-STSA	1.700	2.149	2.429	2.724	2.974	2.395
STSA-ModWD	1.740	2.176	2.445	2.742	2.982	2.417
ModWD-logSTSA	1.840	2.234	2.523	2.802	3.057	2.491
logSTSA-ModWD	1.846	2.246	2.522	2.813	3.062	2.498

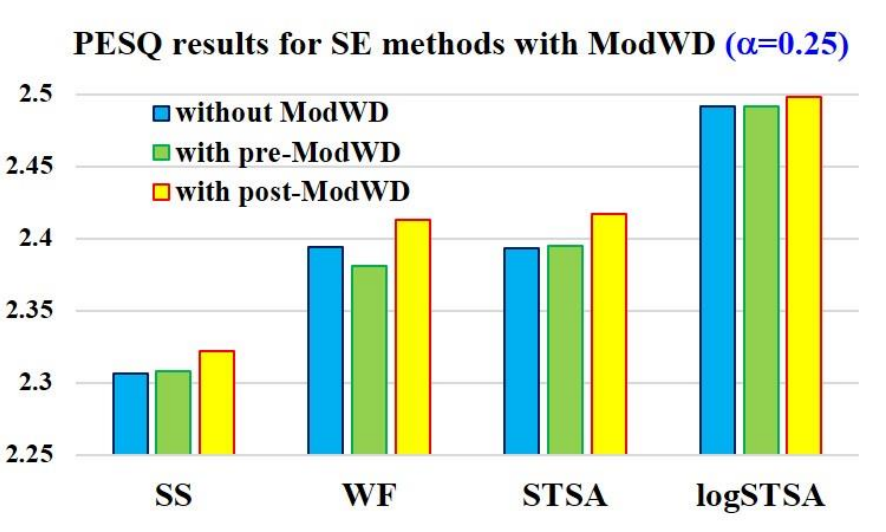
The values in red represent the best score. The higher, the better.

Experimental Results and Discussions (4/4)

- Averaged **PESQ** scores (0 dB to 20 dB)
  - **post-ModWD** integration yields **better** result
  - The case of $\alpha = 0$ (discarding detail portion) still gains improvement



PESQ results averaged over five SNR cases for various SE methods with or without ModWD (α = 0)



PESQ results averaged over five SNR cases for various SE methods with or without ModWD (α = 0.25)

Conclusions and Future Work (1/1)

ModWD, a method for speech enhancement:

- Using DWT to emphasize the approximation portion of spectrograms, which is more likely related to human speech in modulation domain.
- ModWD is additive to speech enhancement algorithms, so as to further improve the speech quality in adverse environments.

Future Work:

- Adopting a multilevel DWT to achieve a high resolution in modulation frequency for the analyzed spectrogram.
- Using a validation set to learn the value of scaling factor α in order to further improve the effectiveness of ModWD.

Thank You