Employing Median Filtering to Enhance the Complex-valued Acoustic Spectrograms in Modulation Domain for Noise-robust Speech Recognition

Hsin-Ju Hsieh[1,2], *Berlin Chen*[2] *and Jeih-weih Hung*[1]
Presenter: Jeih-weih Hung

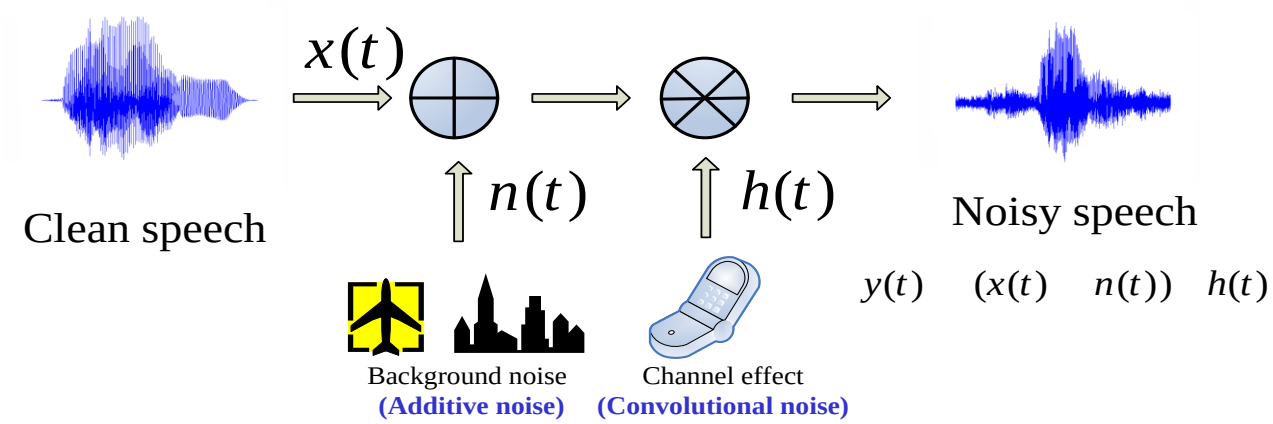
[1]Department of Electrical Engineering
National Chi Nan University, Taiwan
[2]Department of Computer Science & Information Engineering
National Taiwan Normal University, Taiwan

Outline

- Introduction
- Median filtering
- Proposed method
- Experiments and results
- Conclusions

Introduction

- The effect of noise in speech signal



- Extending our previous work in the task of spectrogram enhancement
  - leveraging median filtering (MF) ***to reduce the relatively fast-varying anomaly in the modulation spectra*** for acoustic spectrograms ***caused by noise***
  - the magnitude and phase components of acoustic spectrograms can be enhanced implicitly

Median filtering

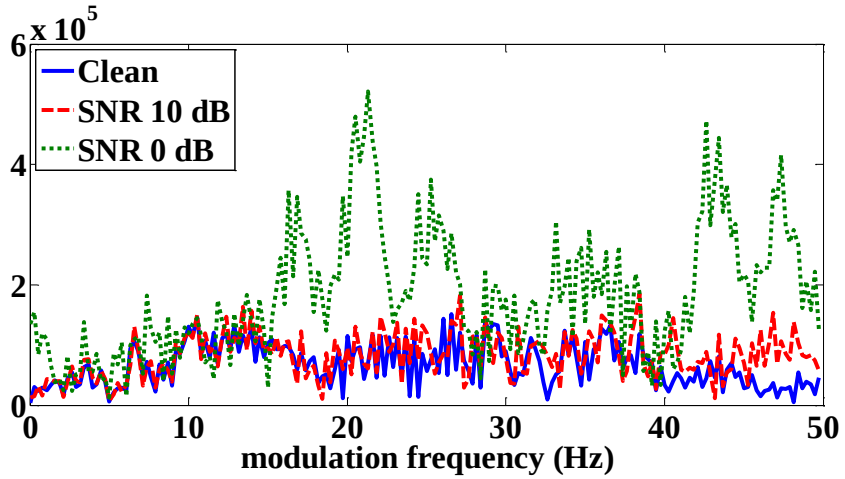
- **The idea of median filtering**:
  - Median filtering (MF) is a rank-order filtering technique, which is a nonlinear operation
    - often used to remove speckle noise in data while preserving the embedded sharp contrasts
  - The procedure of MF:

$$\tilde{x}_t = \mathrm{med}(x_{t-d}, \ldots, x_t, \ldots, x_{t+d})$$

    - $\mathrm{med}(\cdot)$ denotes the median operator, which sorts the elements within the length-$(2d+1)$ sliding window and gives the d^{th} largest element as the output

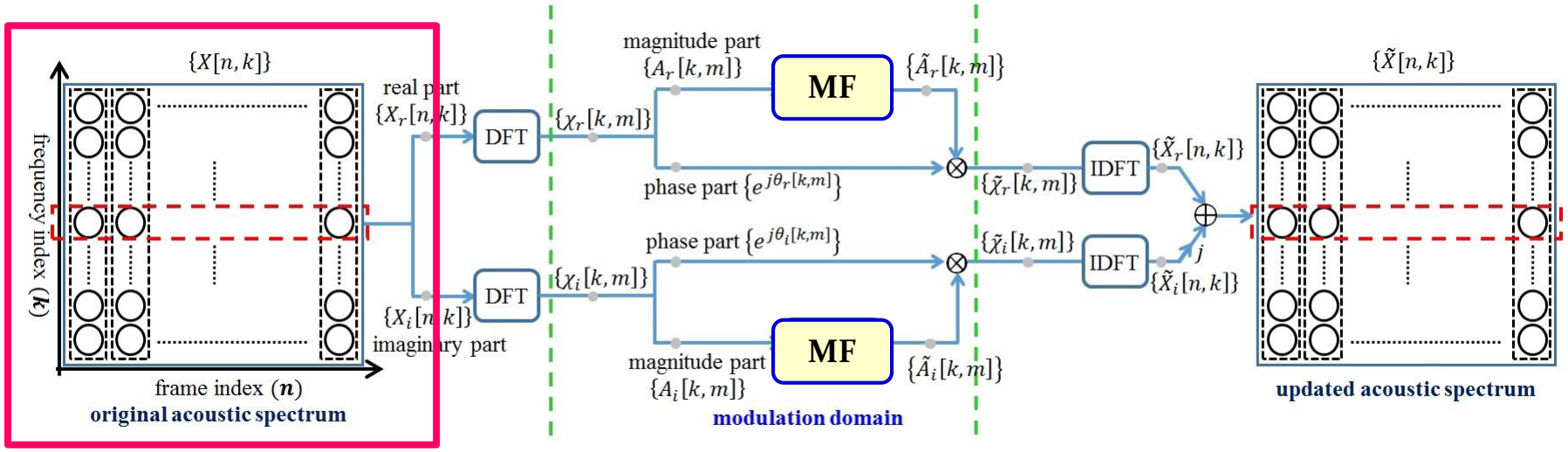
Proposed method (1/4)

- Motivations:
  - The odB-SNR curve contains larger and sharper fluctuations than the clean noise-free one



The magnitude modulation spectral curves of the imaginary acoustic spectrograms at acoustic frequency 375 Hz under three SNR cases (noise type: airport) for one utterance in the Aurora-2 database

Proposed method (2/4)



$$X[n,k] = X_r[n,k] + jX_i[n,k]$$

The diagram of MAS-MF

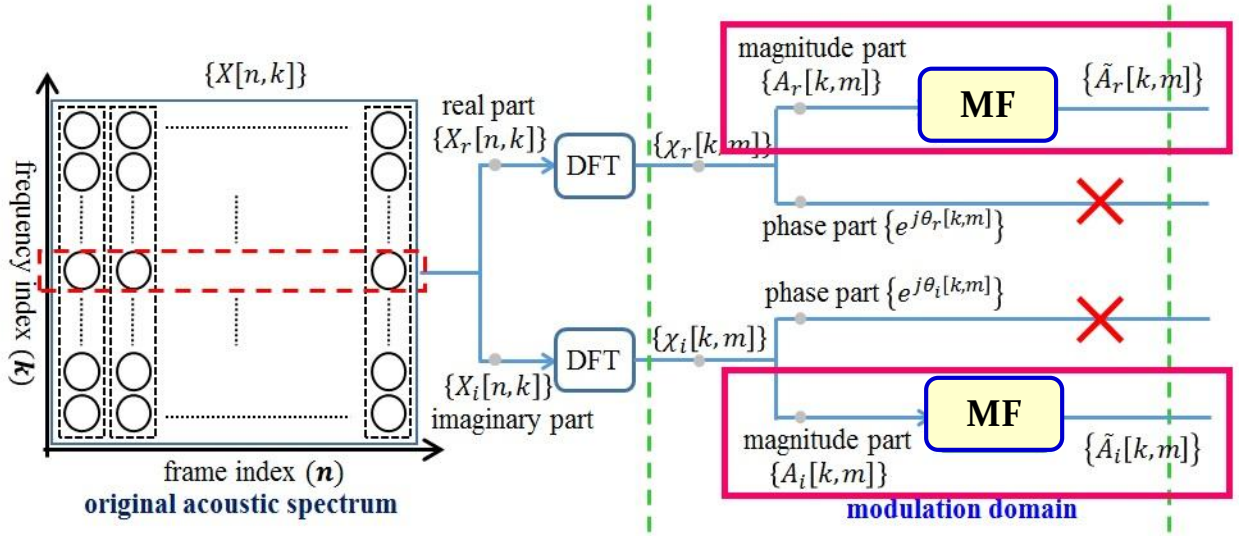
- **Step 1.** Compute the ***modulation spectrum*** for the complex-valued acoustic spectrogram with respect to ***each single acoustic frequency***

$$\text{real part:} \quad \chi_r[k,m] = DFT\{X_r[n,k]\}$$

$$\text{imaginary part:} \quad \chi_i[k,m] = DFT\{X_i[n,k]\}$$

Proposed method (3/4)

- For both the training and testing utterances:

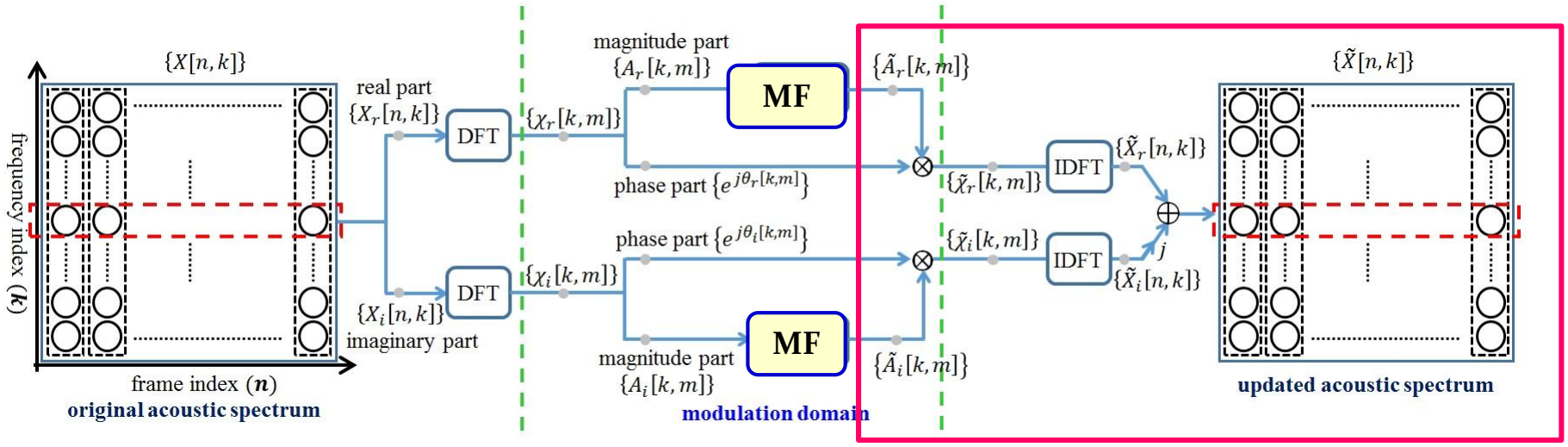


- **Step 2.** Update the **magnitude modulation spectrum** via median filtering (MF)

$$\tilde{A}_r[k,m] = MF\{A_r[k,m]\} = med\{A_r[k,m-d], \ldots, A_r[k,m+d]\}$$

$$\tilde{A}_i[k,m] = MF\{A_i[k,m]\} = med\{A_i[k,m-d], \ldots, A_i[k,m+d]\}$$

Proposed method (4/4)



The diagram of MAS-MF

- **Step 3.** Synthesize the acoustic spectrogram, which is then converted to speech features, MFCC

$$\text{real part:} \quad X_r[n,k] = IDFT\{\tilde{A}_r[k,m]e^{j\theta_r[k,m]}\}$$

$$\text{imaginary part:} \quad X_i[n,k] = IDFT\{\tilde{A}_i[k,m]e^{j\theta_i[k,m]}\}$$

Experiments and results (1/4)

- Speech Corpus

Aurora 2.0	
Speech content	0, 1, 2, …, 8, 9, oh
Training data	8440 clean sentences
Testing data	Set A: 28028 (subway, babble, car, exhibition)
	Set B: 28028 (restaurant, street, airport, train station)
	Set C: 14014 (MIRS subway, MIRS street)
SNR	clean, 20dB,15dB, 10dB, 5dB, 0dB, -5dB

- Hidden Markov Model
  - Each digit model contains 16 states
  - Each state contains 20 Gaussian mixture

Experiments and results (2/4)

Recognition accuracy rates (%) on the **Aurora-2 task**, achieved by baseline MFCC and various robustness methods

	Set A	Set B	Set C	Avg.	RR
MFCC baseline	54.87	48.87	63.95	54.29	-
CMS	66.81	71.79	67.64	68.97	32.12
CMVN	75.93	76.76	76.82	76.44	48.46
HEQ	80.03	82.05	80.10	80.85	58.11
AFE	87.68	87.10	86.27	87.17	71.93
MAS-HEQ	86.49	88.13	84.98	86.84	71.21
MAS-MF (3)	**87.71**	**88.58**	**86.23**	**87.76**	**73.22**
MAS-MF (5)	**87.52**	**88.81**	**86.39**	**87.81**	**73.33**
MAS-MF (6)	**87.65**	**88.90**	**86.47**	**87.91**	**73.55**
MAS-MF (7)	**87.53**	**88.85**	**86.31**	**87.81**	**73.33**
MAS-MF(9)	**87.57**	**88.84**	**86.11**	**87.79**	**73.29**

RR (%) is the relative error rate reduction over the MFCC baseline

Experiments and results (3/4)

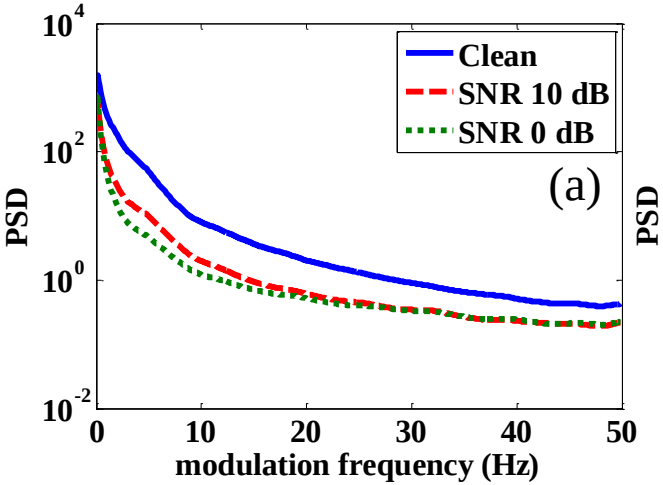
Recognition accuracy rates (%) on the **Aurora-4 task** (viz. Sets 8 to 14), achieved by baseline MFCC and various robustness methods

	MFCC	AFE	MAS-MF (3)
Clean	63.45	79.34	**76.35**
Car	37.27	72.56	68.03
Babble	30.31	61.58	**63.13**
Restaurant	34.30	55.65	**58.27**
Street	26.13	58.31	57.72
Airport	31.85	60.55	**61.36**
Train	26.95	60.88	59.56
Average	35.75	64.12	63.49

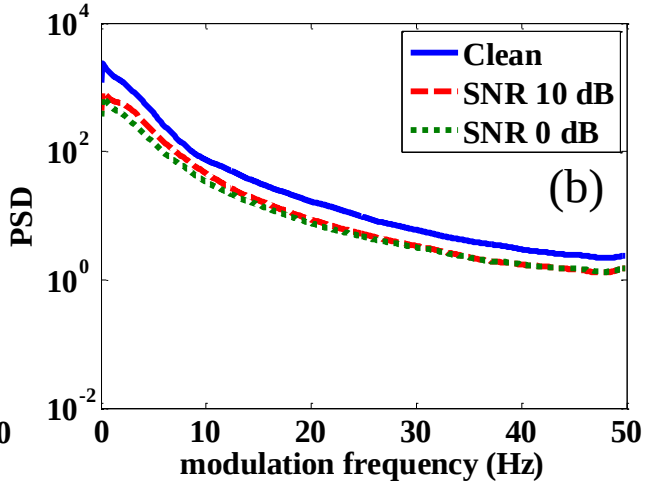
RR (%) is the relative error rate reduction over the MFCC baseline

Experiments and results (4/4)

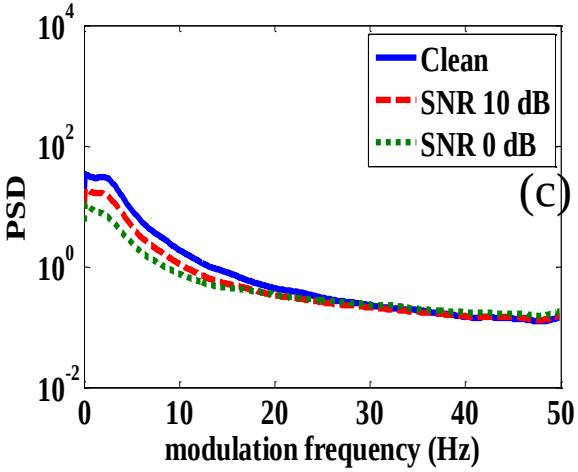
The first cepstral (c_1) PSD (power spectrum density) curves



(a) MFCC baseline



(b) AFE



(c) MAS-MF (6)

- For the unprocessed case, the environmental noise results in a significant PSD mismatch over ***the entire modulation frequency range*** [0 50 Hz]
- The PSD mismatch can be considerably suppressed after performing either of AFE and MAS-MF

Conclusions

- **This work provides a novel feature extraction method, MAS-MF, for robust speech recognition**
  - **Joint normalization** of real and imaginary acoustic spectra
  - The spectral distortion is **reduced via MF**
    - **MF is nonlinear and has little to do with learning**
  - **Recognition accuracy** can be obviously promoted

- **Several directions as to the future work**
  - combining MAS-MF with other robustness methods
  - examining MAS-MF on the state-of-the-art **deep neural network (DNN)** scenario

Thank you for your attention