Denoising-and-Dereverberation Hierarchical Neural Vocoder for Robust Waveform Generation

Yang Ai[1], Haoyu Li[2], Xin Wang[2], Junichi Yamagishi[2], Zhenhua Ling[1]

[1]University of Science and Technology of China, P.R.China
[2]National Institute of Informatics, Japan

Paper ID: 1255

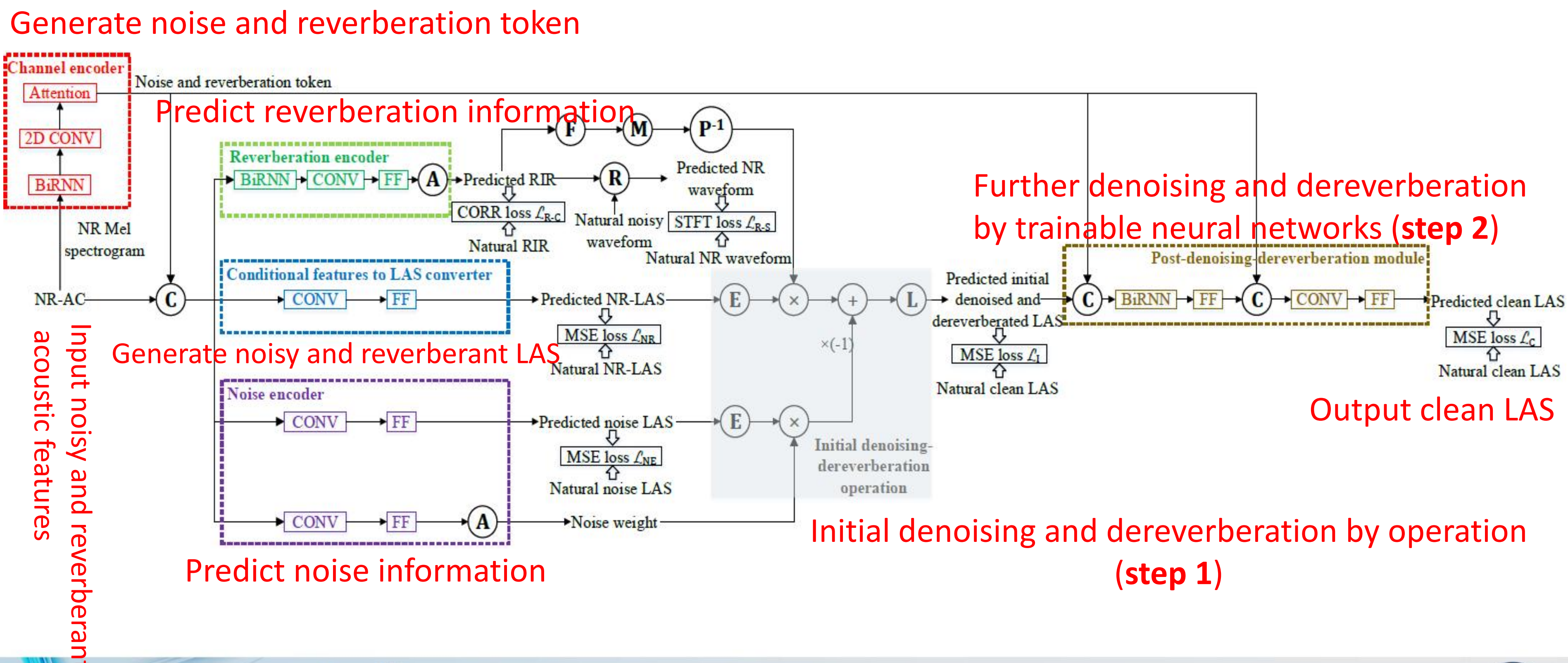
SLT 2021

Proposed method

- Propose a denoising and dereverberation hierarchical neural **vocoder** (DNR-HiNet): convert noisy and reverberant acoustic features into a clean speech waveform;
  - Denoising and dereverberation amplitude spectrum predictor (DNR-ASP)
  - Phase spectrum predictor (PSP)

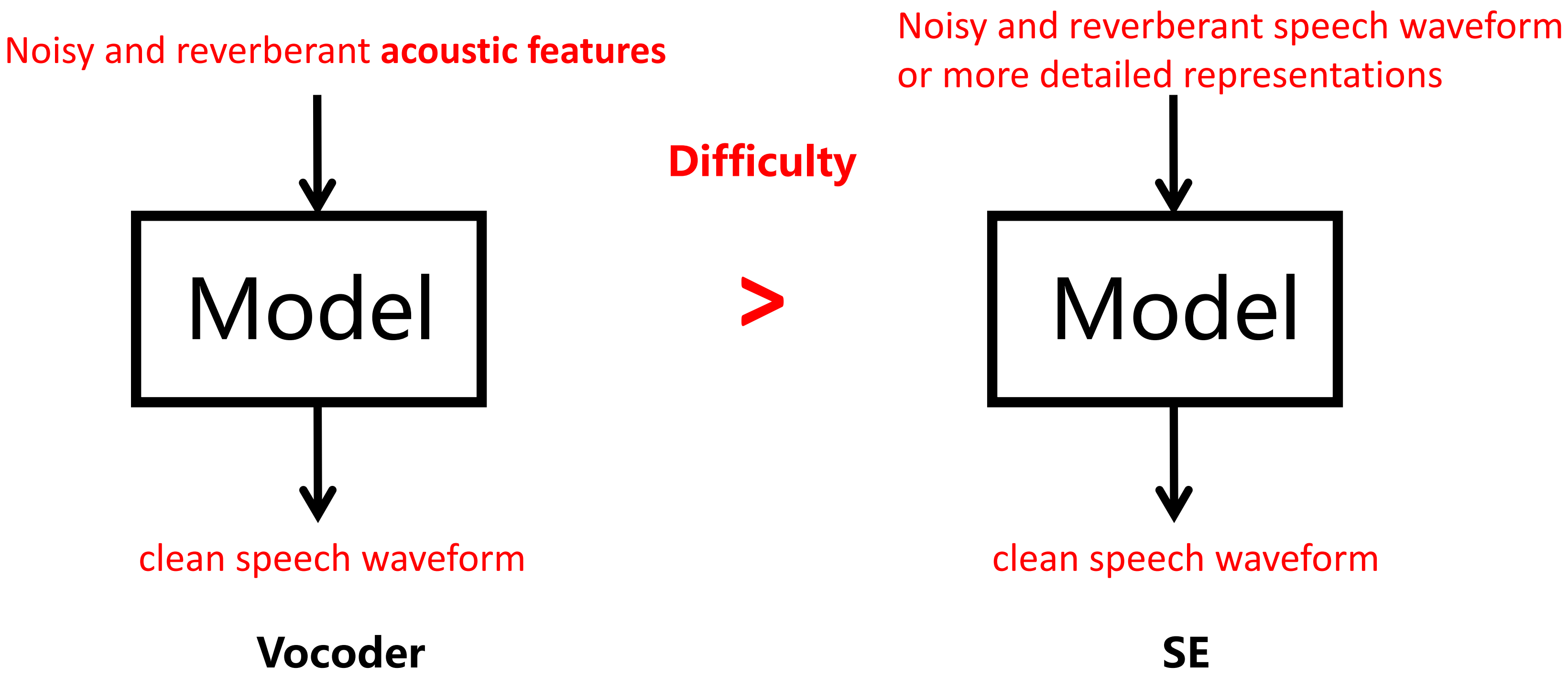
Overview of DNR-ASP

- Overview of DNR-ASP:
  - ➢predict clean log amplitude spectra (LAS) from input noisy and reverberant acoustic features



Comparsion between DNR vocoder and SE method

- The difference between denoising and dereverberation vocoder and SE methods:



Experimental results

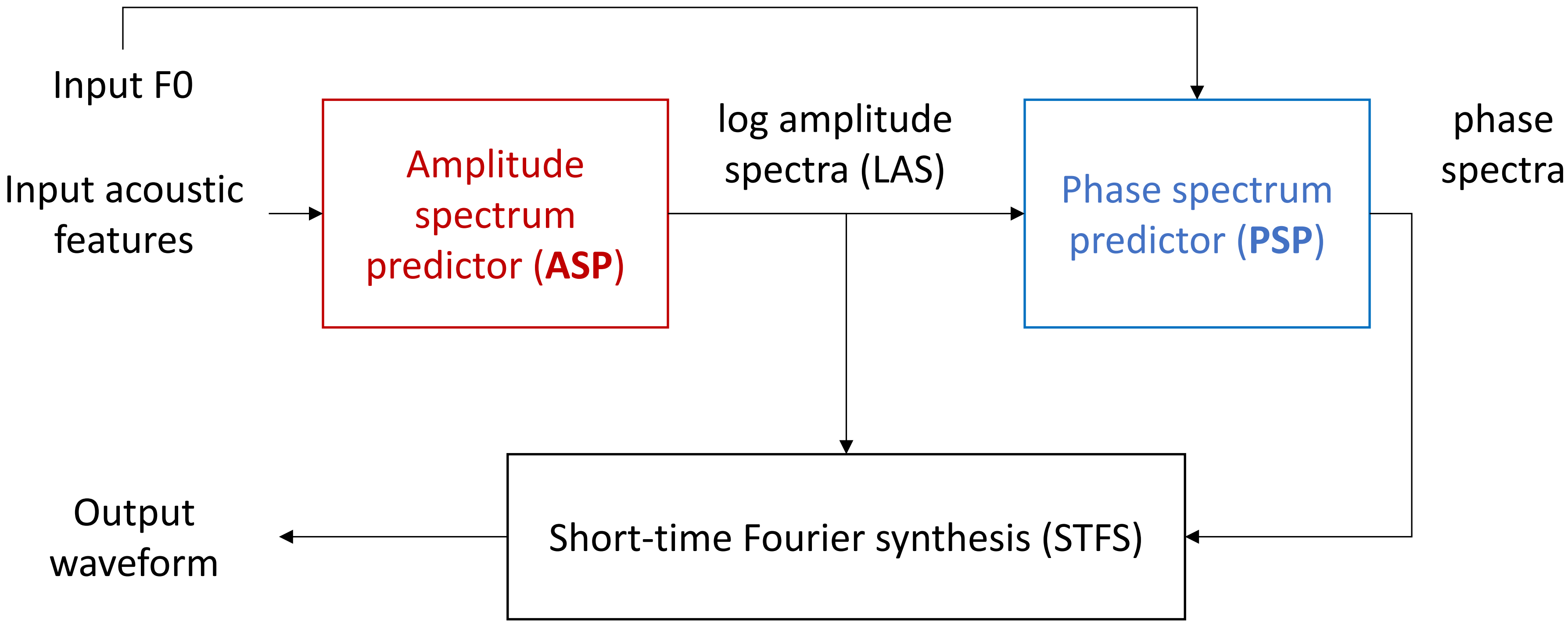
- The DNR-HiNet vocoder achieved better performance than the original HiNet vocoder and a few other **vocoder**s
- The DNR-HiNet vocoder achieved competitive performace with several advanced speech enhancement (**SE**) methods.

Contents

- Review of HiNet vocoder
- Theory
- Experiments
- Problems and future works
- Demos

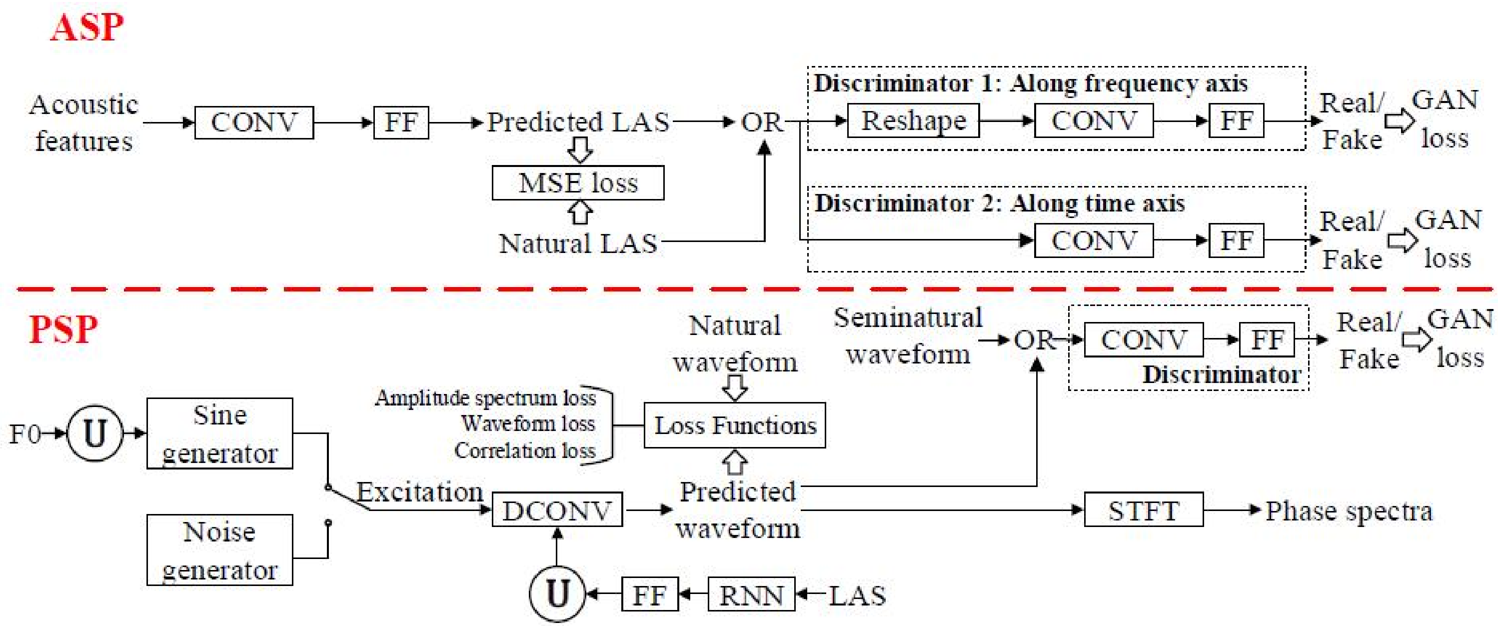
Review of HiNet vocoder

$$x(t) \Longleftrightarrow X(j\omega) = |X(j\omega)| e^{j\angle X(j\omega)}$$



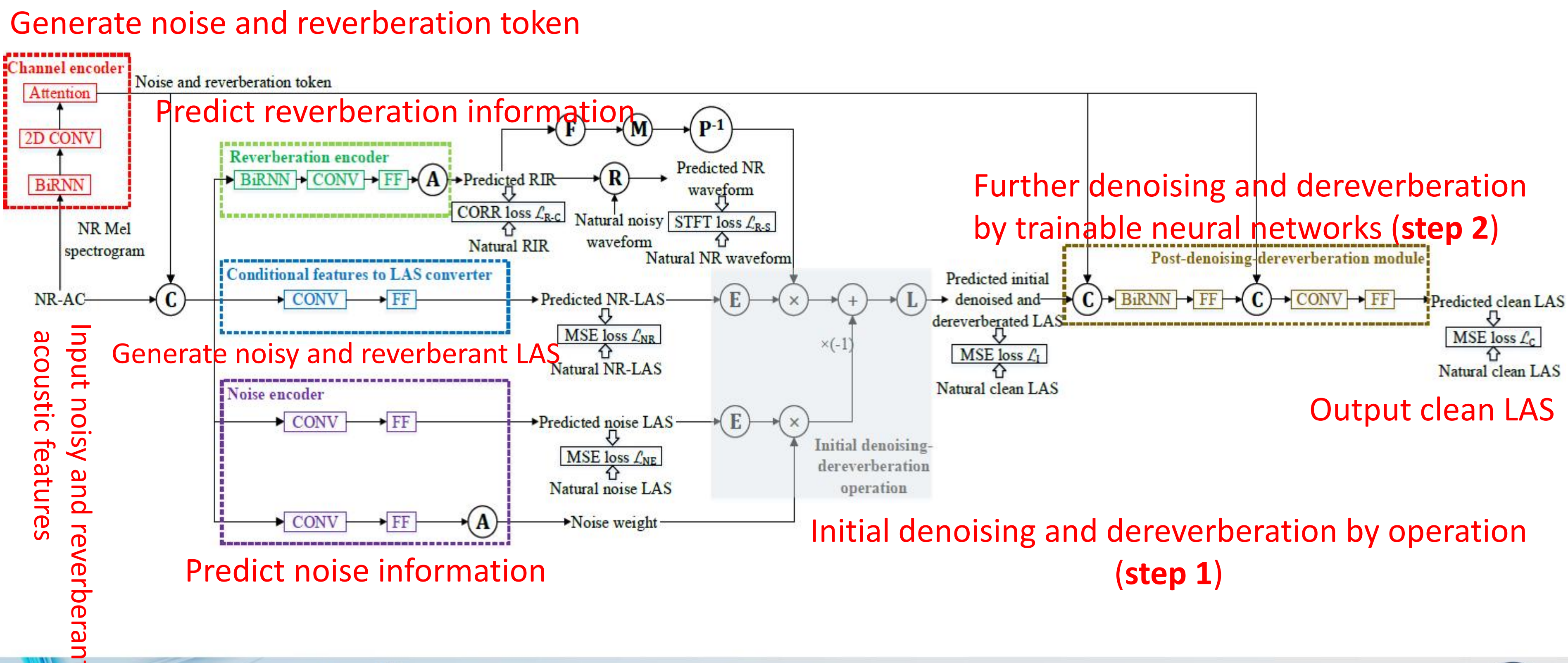
Review of HiNet vocoder

Implement DNR-HiNet mainly by modifying the ASP in the original HiNet vocoder: Design denoising and dereverberation ASP (DNR-ASP)



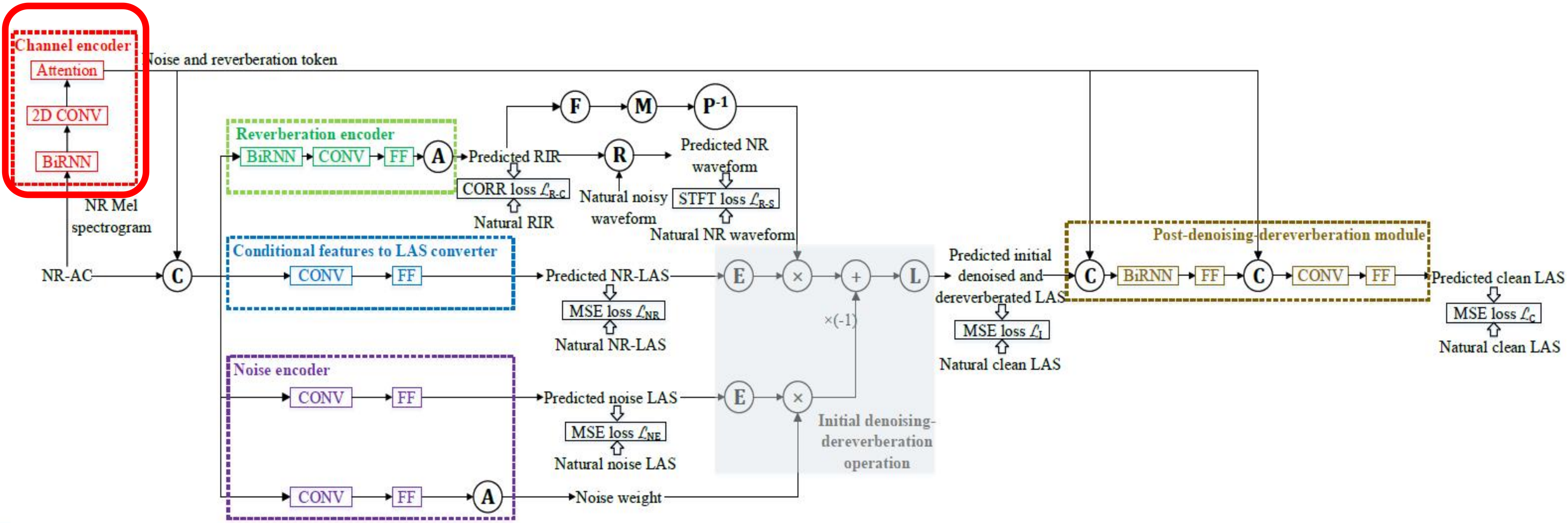
Theory

- Overview of DNR-ASP:
  - predict clean log amplitude spectra (LAS) from input noisy and reverberant acoustic features



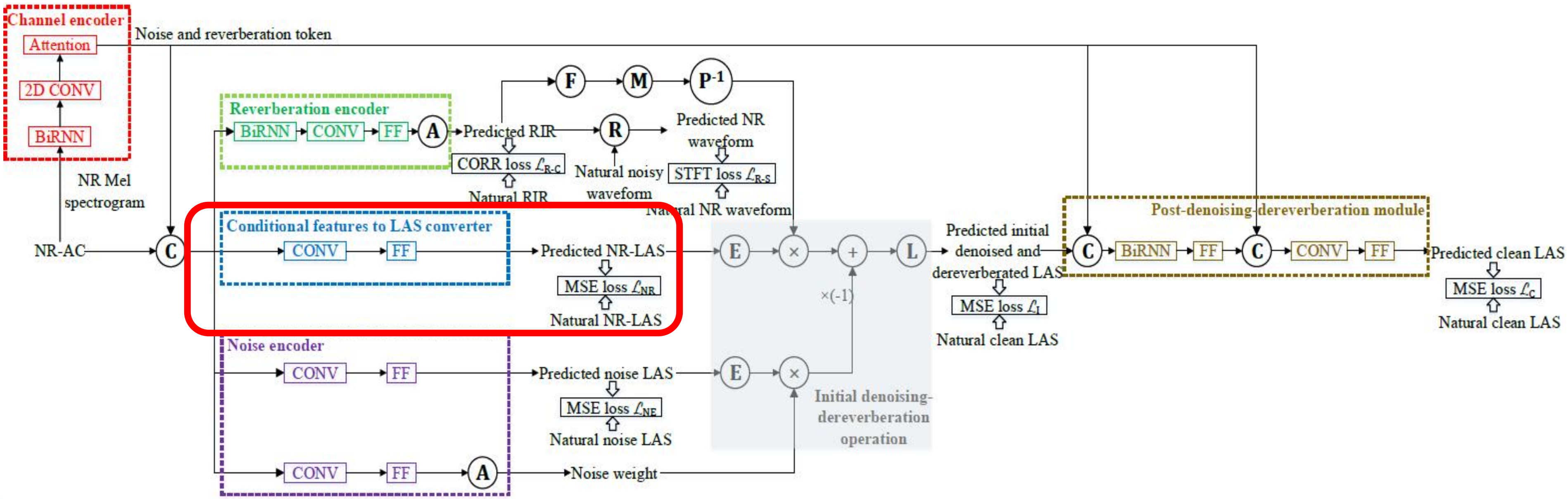
Theory

- DNR-ASP->Channel encoder:
  - Aim: Distinguish different types of noise and reverberation and generalize with unseen types in the test set
  - Input: Noisy and reverberant Mel spectrogram
  - Output: Noise and reverberation token



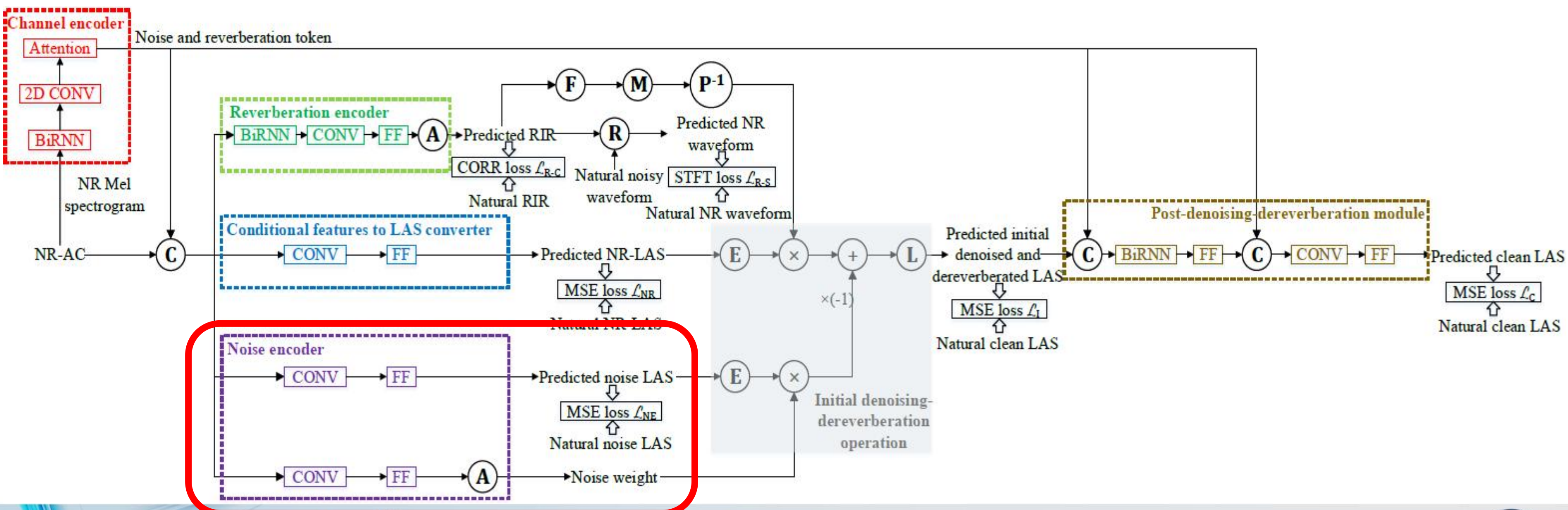
Theory

- DNR-ASP->Conditional features to LAS converter:
  - ➢Aim: Predict noisy and reverberant LAS for initial denosing and dereverberation
  - ➢Input: Noisy and reverberant acoustic features + token
  - ➢Output: noisy and reverberant LAS
  - ➢Loss function: MSE



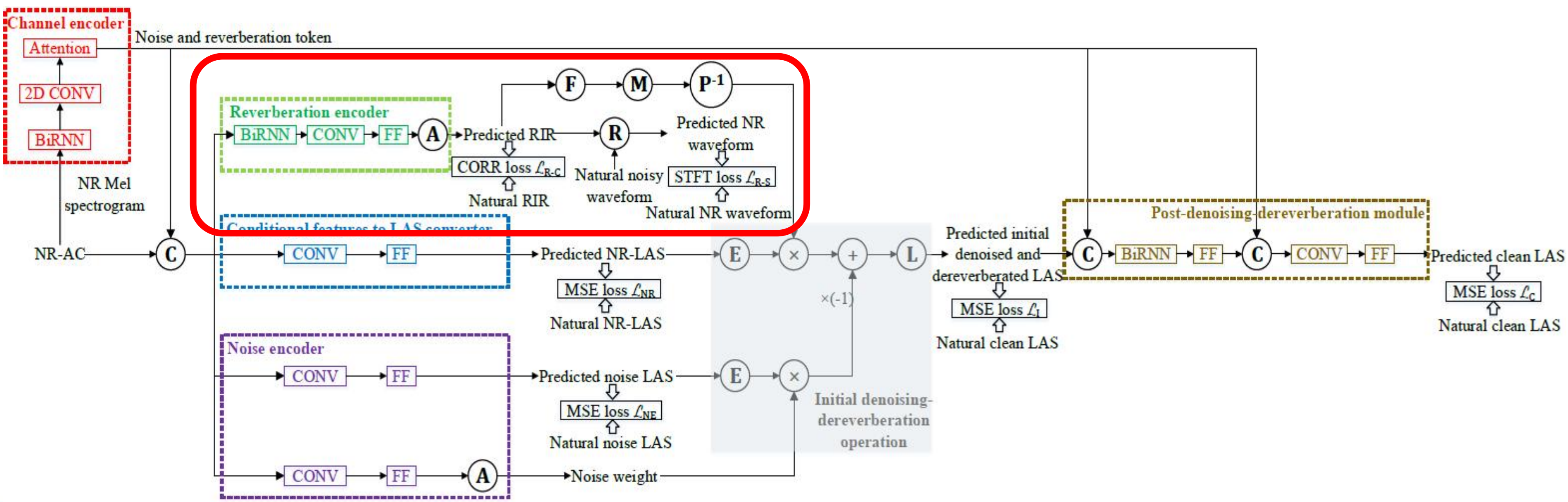
Theory

- DNR-ASP->Noise encoder:
  - Aim: Predict noise-related information for initial denosing and dereverberation
  - Input: Noisy and reverberant acoustic features + token
  - Output: noise LAS and the weight of noise amplitude spectra
  - Loss function: MSE



Theory

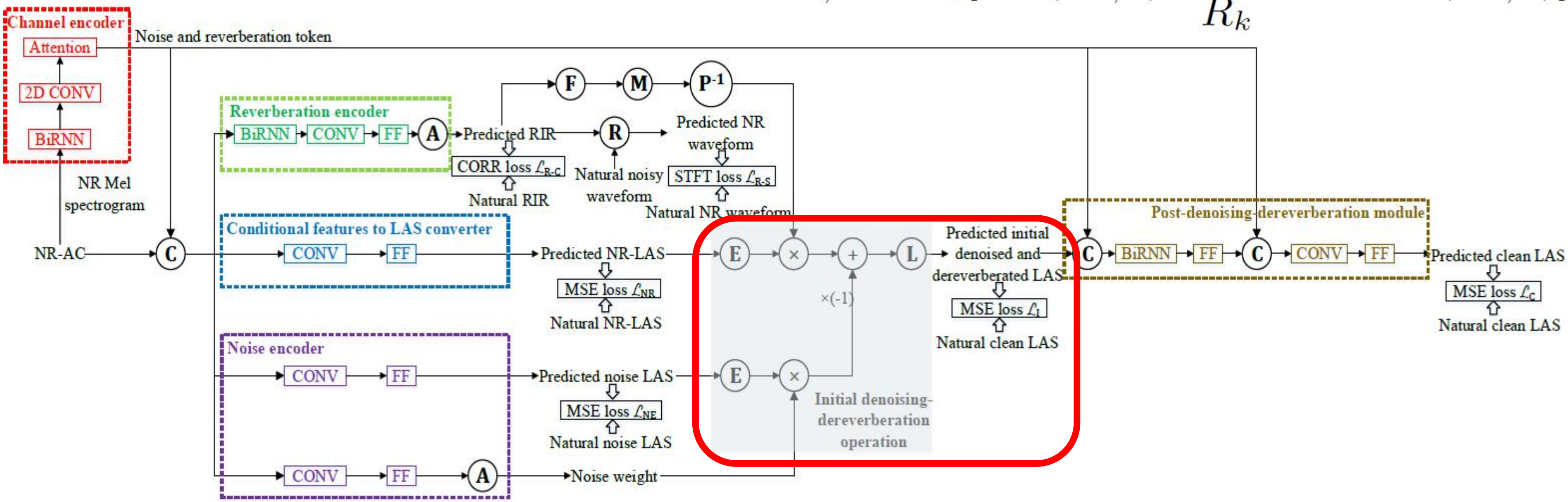
- DNR-ASP->Reverberation encoder:
  - Aim: Predict reverberation-related information for initial denosing and dereverberation
  - Input: Noisy and reverberant acoustic features + token
  - Output: Room impulse response (RIR)
  - Loss function: CORR loss and STFT loss



Theory

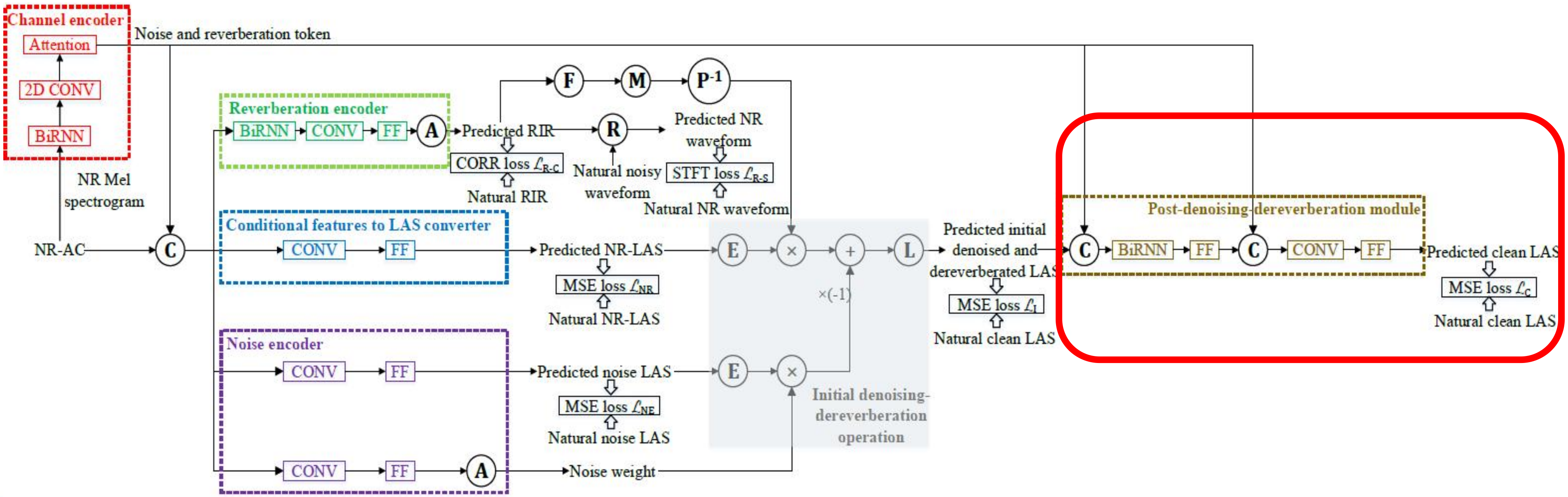
- DNR-ASP->Initial denoising-dereverberation operation:
  - Aim: Initially remove the noise and reverberation from the noisy and reverberant LAS by operation
  - Input: Noisy and reverberant LAS, noise LAS, weigth of noise amplitude spectra and RIR
  - Output: Initial denoised and dereverberated LAS
  - Loss function: MSE loss

$$\tilde{L}^{C}_{n,k} = \log[\exp(\hat{L}^{NR}_{n,k}) \cdot \frac{1}{\hat{R}_k} - \alpha \cdot \exp(\hat{L}^{NE}_{n,k})]$$



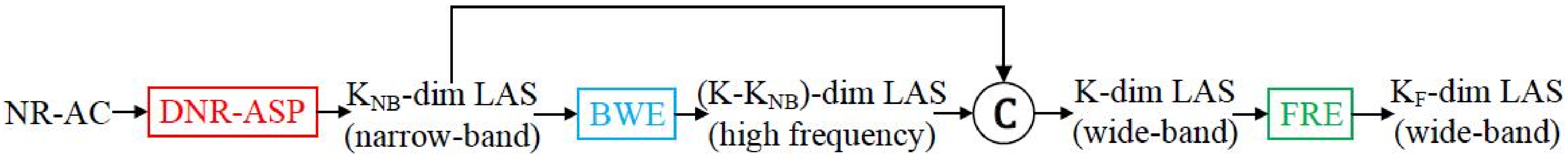
Theory

- DNR-ASP->Post-denoising-dereverberation module:
  - Aim: Further remove the noise and reverberation from the initial denoised and dereverberated LAS by trainable neural networks
  - Input: Initial denoised and dereverberated LAS
  - Output: clean LAS
  - Loss function: MSE loss



Theory

- DNR-ASP->Add two additional models:
  - Bandwidth extension (BWE) model
  - Frequecy resolution extension (FRE) model



Theory

- PSP
  - Training: Using natural clean F0 and LAS as input, using natural waveform as output
  - Generation: Using natural noisy and reverberant F0 and clean LAS predicted by DNR-ASP as input

Experiments

- Data and feature configuration:
  - Training/Validation set: 28 speakers, 11012/560 utterances, 10 noise types and 4 SNRs, 5 reverberation RIR types
  - Test set: (unseen) 2 speakers, 824 utterances, 5 noise types and 4 SNRs, 3 reverberation RIR types
  - Acoustic features: 80-dim Mel spectrogram, 1-dim F0, 1-dim voiced/unvoiced flag

Experiments

- Experimental models--**Vocoders**
  - Baseline-NSF
  - Baseline-NSF': low-bound model
  - Baseline-HiNet
  - Baseline-HiNet': low-bound models
  - DNR-HiNet
  - DNR-HiNet w/ BF: add the BWE and FRE models
- Experimental models--**SE methods**
  - cIRM
  - SEGAN
  - WaveNet
  - T-GSA
  - DNR-HiNet* w/ BF: using natural noisy and reverberant phase spectra

Experiments

- Objective results
  - ➢Comparsion among neural vocoders

Reflect:

	STOI	CSIG	CBAK	COVL
Reflect:	speech intelligibility	MOS on signal distortion	MOS on noise intrusiveness	MOS on overall effect
Noisy and reverberant audio	0.777	2.21	1.84	2.05
Baseline-NSF′	0.740	1.91	1.59	1.70
Baseline-NSF	0.763	2.99	1.98	2.37
Baseline-HiNet′	0.746	2.18	1.76	1.99
Baseline-HiNet	0.705	2.99	2.06	2.48
DNR-HiNet	0.769	**3.25**	2.24	2.69
DNR-HiNet w/ BF	**0.783**	3.24	**2.29**	**2.75**
cIRM	0.701	2.24	1.81	1.98
SEGAN	0.659	1.76	1.26	1.55
WaveNet	0.800	3.35	2.35	2.78
T-GSA	**0.818**	3.32	2.43	2.87
DNR-HiNet* w/ BF	0.803	**3.38**	**2.44**	**2.92**

Experiments

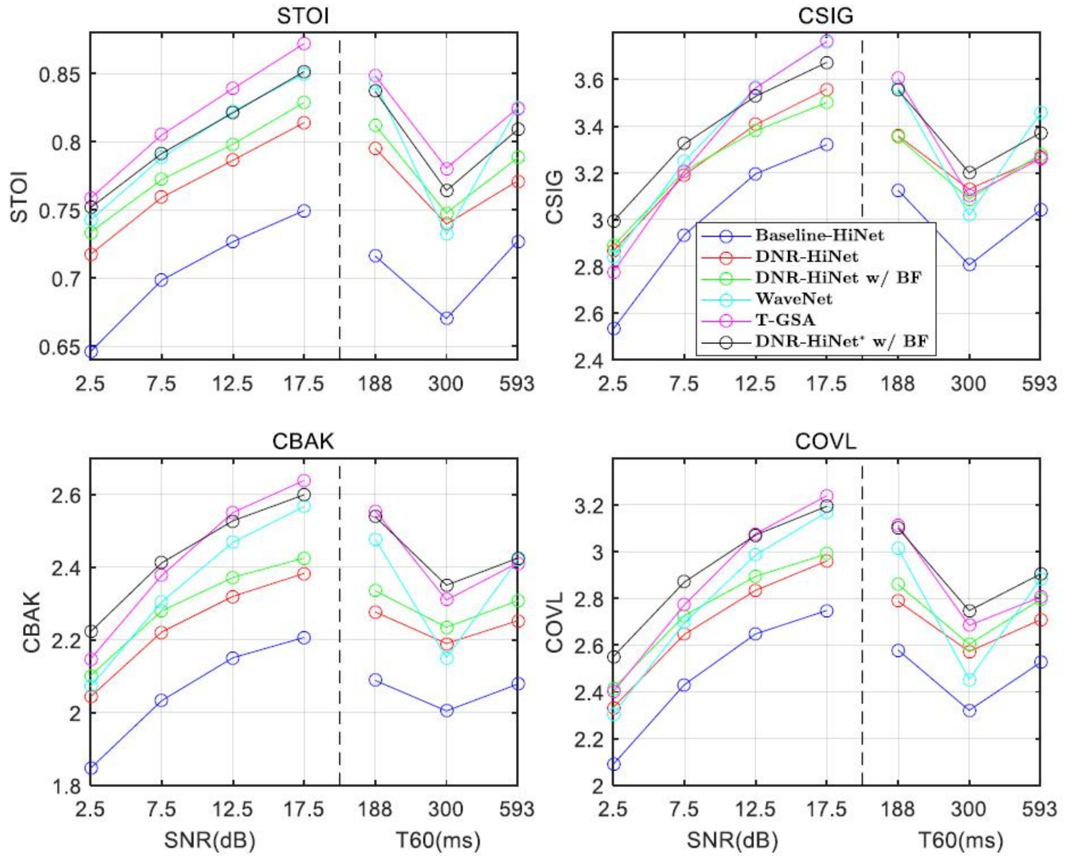
- Objective results
  - Comparsion with SE methods

Reflect:

	speech intelligibility	MOS on signal distortion	MOS on noise intrusiveness	MOS on overall effect
	STOI	CSIG	CBAK	COVL
Noisy and reverberant audio	0.777	2.21	1.84	2.05
Baseline-NSF′	0.740	1.91	1.59	1.70
Baseline-NSF	0.763	2.99	1.98	2.37
Baseline-HiNet′	0.746	2.18	1.76	1.99
Baseline-HiNet	0.705	2.99	2.06	2.48
DNR-HiNet	0.769	**3.25**	2.24	2.69
DNR-HiNet w/ BF	**0.783**	3.24	**2.29**	**2.75**
cIRM	0.701	2.24	1.81	1.98
SEGAN	0.659	1.76	1.26	1.55
WaveNet	0.800	3.35	2.35	2.78
T-GSA	**0.818**	3.32	2.43	2.87
DNR-HiNet* w/ BF	0.803	**3.38**	**2.44**	**2.92**

Experiments

- Objective results
  - ➢Results of different systems under different SNR and RIR conditions of test set



Experiments

- Subjective results
  - Suppression score: Higher score represents better noise and reverberation suppression
  - MUSHRA score: Higher score represents better speech quality

Comparsion among neural vocoders

Comparsion with SE methods

	Systems	Suppression score	MUSHRA score
Group 1	**Baseline-NSF**	5.635±0.131	57.30±1.74
	Baseline-HiNet	5.477±0.133	57.82±1.60
	DNR-HiNet	5.774±0.128	60.51±1.60
	DNR-HiNet w/ BF	**5.939±0.128**	**61.73±1.55**
Group 2	**DNR-HiNet w/ BF**	**5.700±0.129**	**65.38±1.48**
	cIRM	4.975±0.138	55.27±1.88
	SEGAN	4.873±0.155	49.06±2.07
	WaveNet	5.396±0.130	62.18±1.59
	T-GSA	5.624±0.121	62.28±1.58
	DNR-HiNet* w/ BF	**5.703±0.129**	**65.56±1.52**

Problems and future works

- The DNR-ASP model is huge --> Model simplification
- The role of each module needs to be studied --> Ablation test

Demos

- http://home.ustc.edu.cn/~ay8067/DNR/demo.html

Thank you