Improving Neural-Network-Based Speech Enhancement for Noise Reduction and Intelligibility Boosting

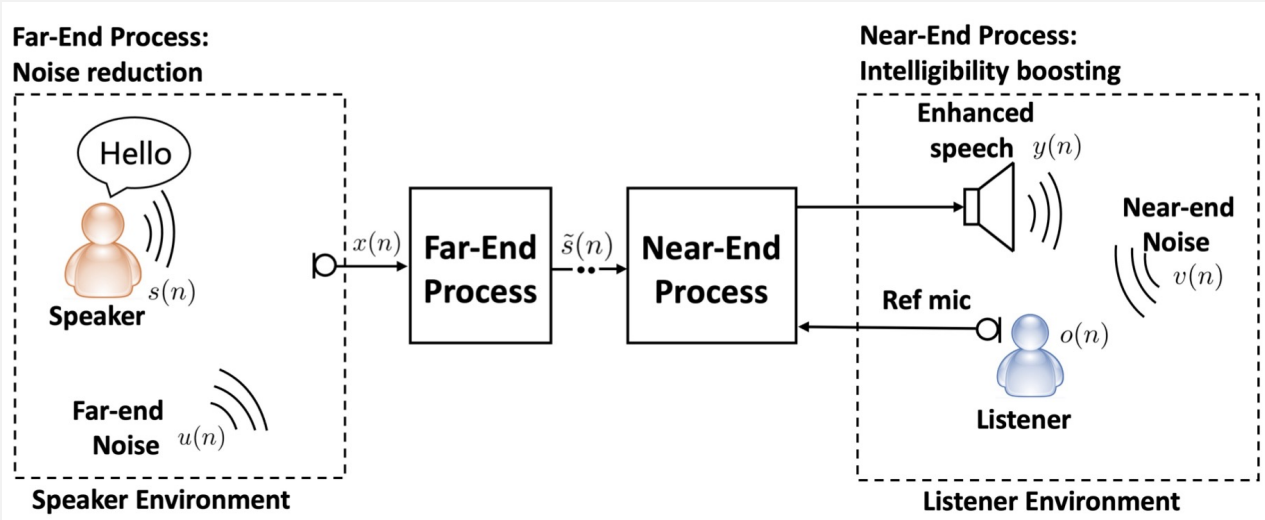
Li Haoyu

2022-07-12

CONTENTS

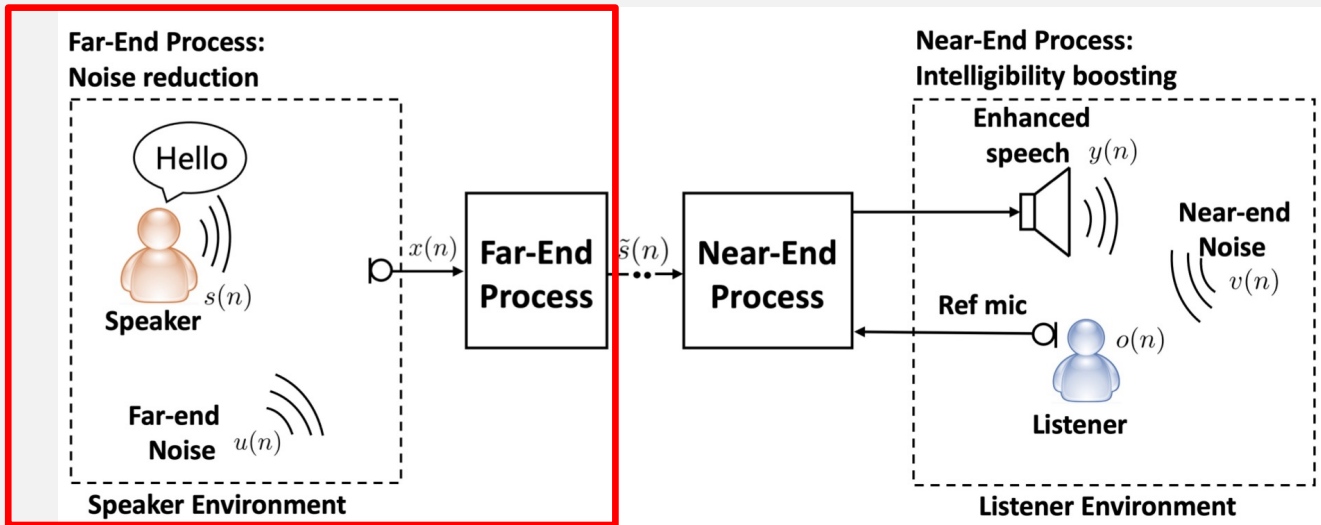
INTRODUCTION

- **Speech enhancement**
  - Speech quality and intelligibility severely degrades in noisy environment
  - Speech enhancement aims to improve the quality and intelligibility of degraded speech
- **Depending on the usage scenario:**
  - Noise reduction
  - Intelligibility boosting



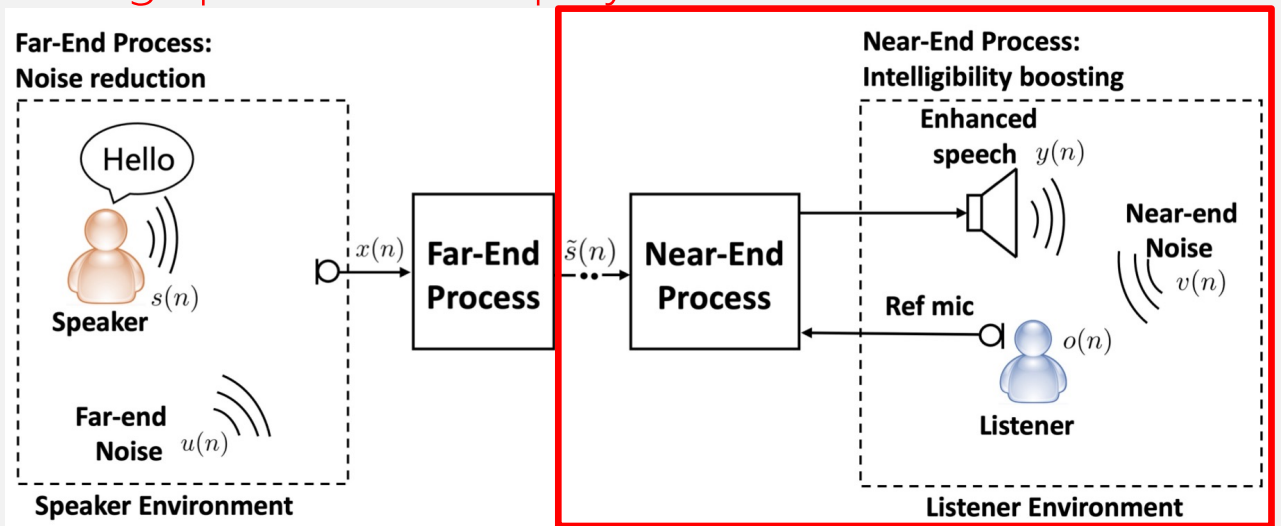
INTRODUCTION

- **Speech enhancement**
  - Speech quality and intelligibility severely degrades in noisy environment
  - Speech enhancement aims to improve the quality and intelligibility of degraded speech
- **Depending on the usage scenario:**
  - Noise reduction -> removing noise before transmission
  - Intelligibility boosting



INTRODUCTION

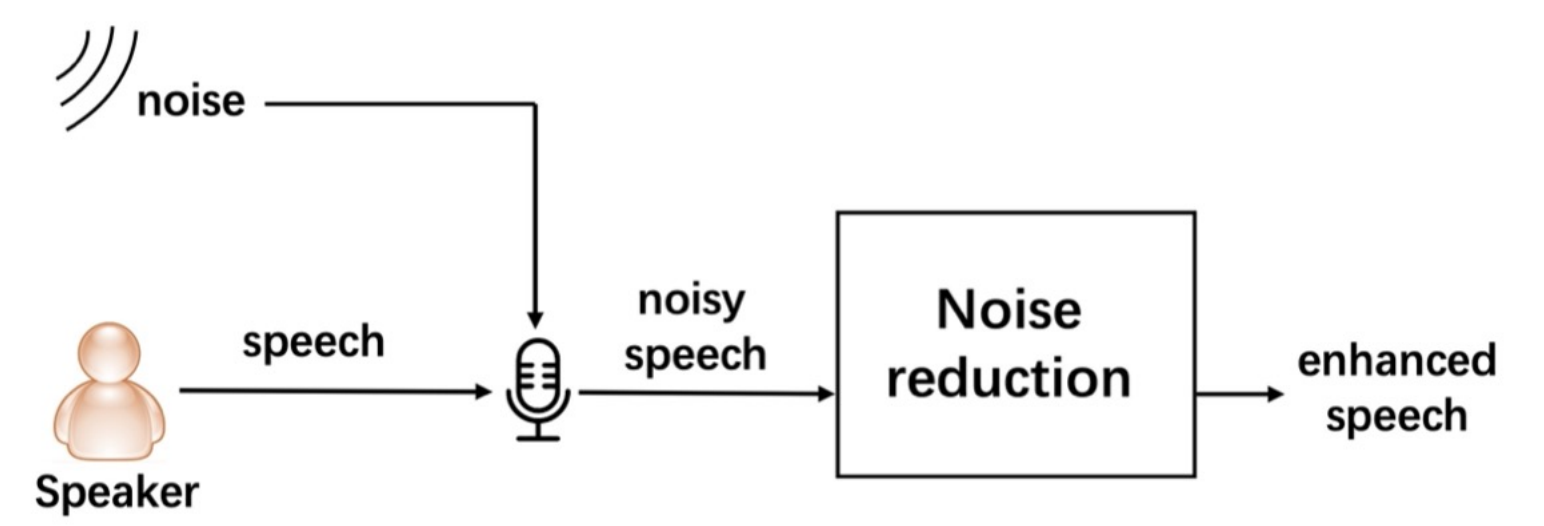
- **Speech enhancement**
  - Speech quality and intelligibility severely degrades in noisy environment
  - Speech enhancement aims to improve the quality and intelligibility of degraded speech
- **Depending on the usage scenario:**
  - Noise reduction
  - Intelligibility boosting -> enhancing speech before playback



INTRODUCTION

1. Noise reduction

- In speaker side, signal captured by the mic is a mixture of speech and noise
- Modify noisy signal to suppress the noise contained in speech
- Applications: Mobile telephony, Hearing aids, robust ASR, etc.

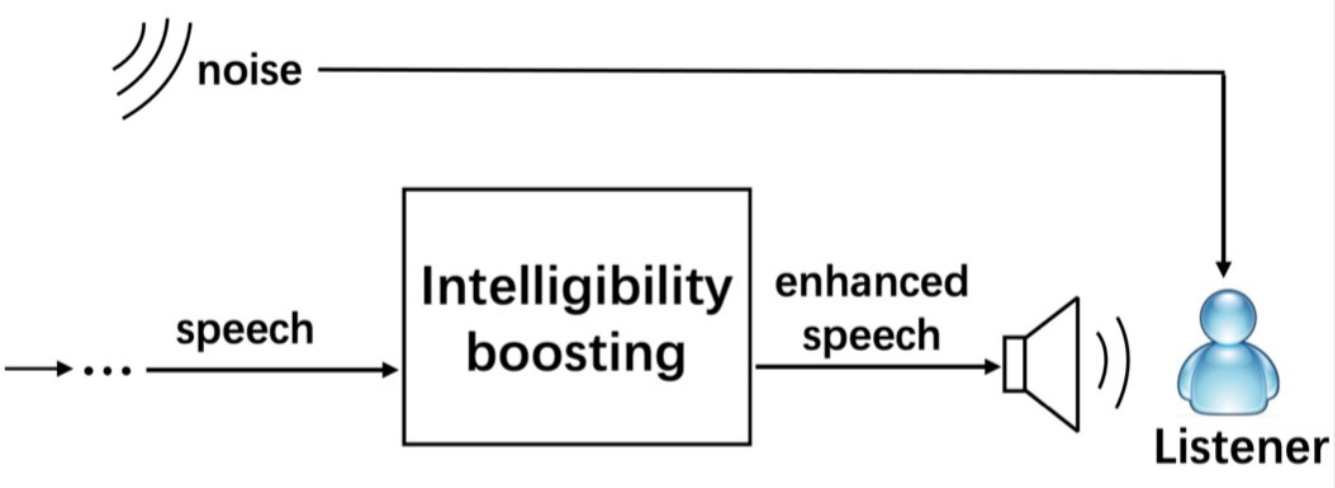


Demo:

- Noisy input
- Processed output

INTRODUCTION

1. **Noise reduction**
2. **Intelligibility boosting**
   - In listener side, noise source is physically present
   - Modify clean signal only to enhance its intelligibility
   - Applications: Mobile telephony, public-address announcement, etc.

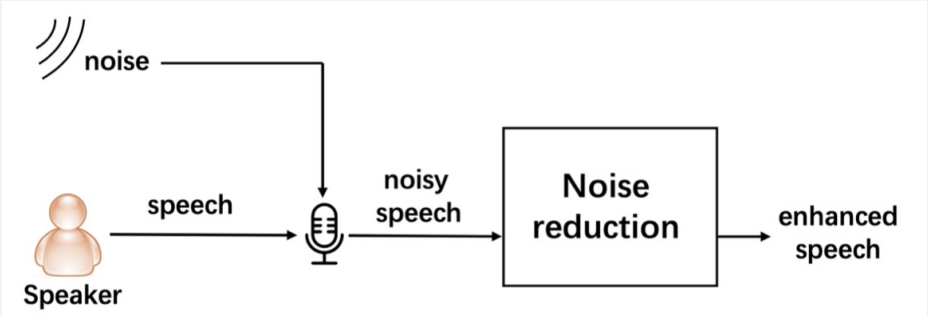


Demo:

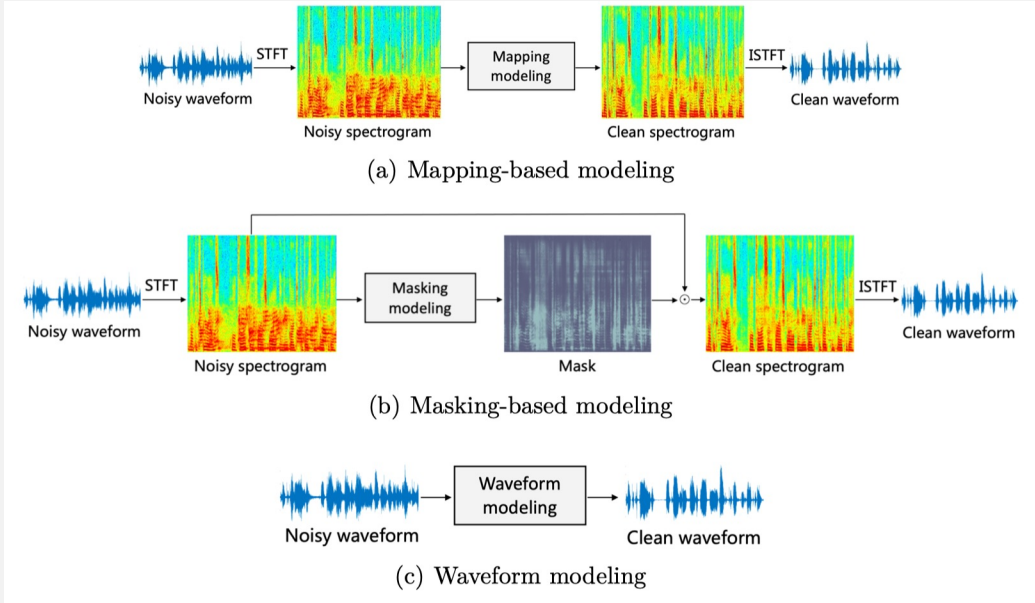
- Clean speech in noise
- Modified speech in noise

THESIS OUTLINE

- Conventional approaches for noise reduction (Chapter 2)

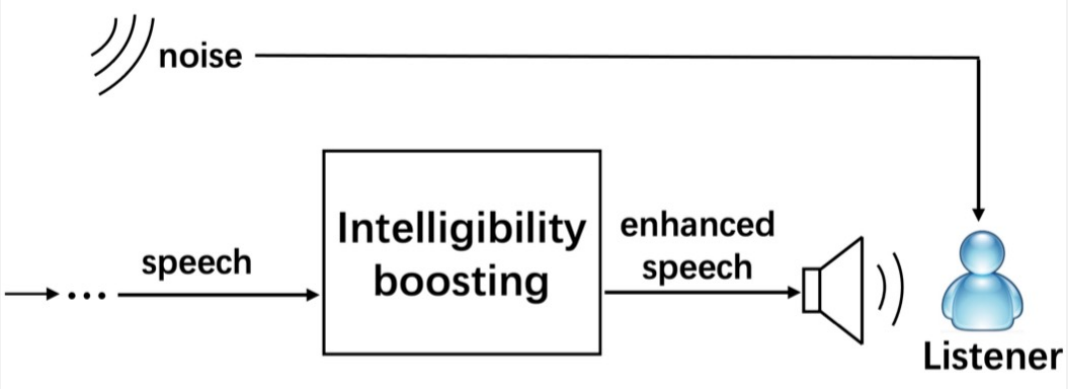


  - Signal processing approaches:
    - Spectral subtraction, Wiener filter, MMSE estimator, etc.
  - Neural approaches:
    - Mapping-based, Masking-based, Waveform modeling



THESIS OUTLINE

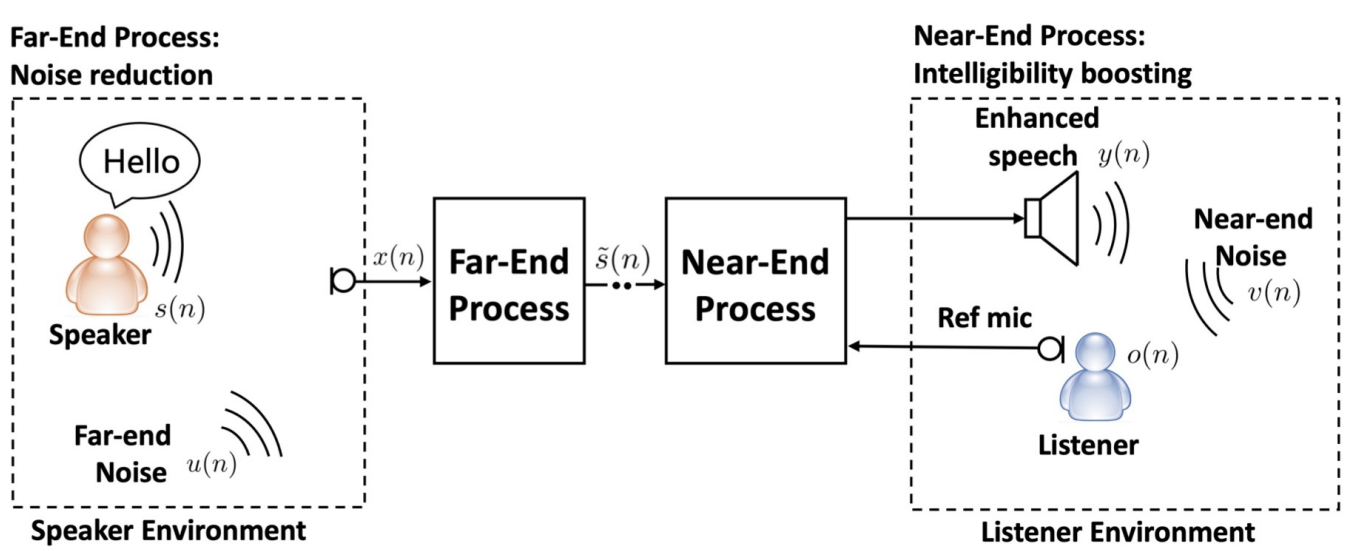
- Conventional approaches for intelligibility boosting (Chapter 2)



  - Most are based on signal processing:
    - Knowledge-based modification, .e.g., formant enhancement, dynamic range compression
    - Lombard-style conversion, i.e., convert to Lombard speech using voice conversion techniques)
    - Metric-oriented optimization -> optimize a certain intelligibility metric

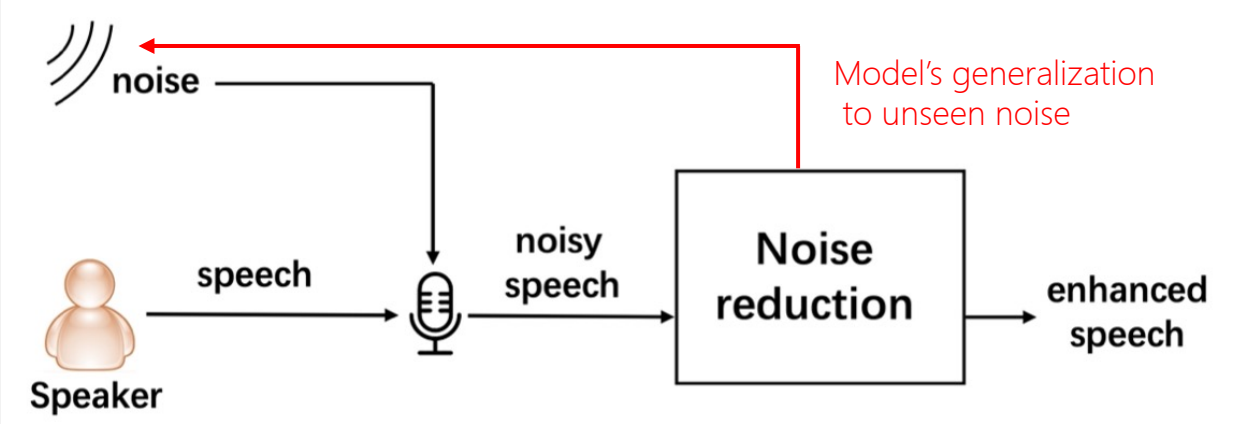
THESIS OUTLINE

- **Thesis issues**
  - Primary goal:
    - Enhance speech quality and intelligibility in speech communication
      - -> Improve noise reduction and intelligibility boosting



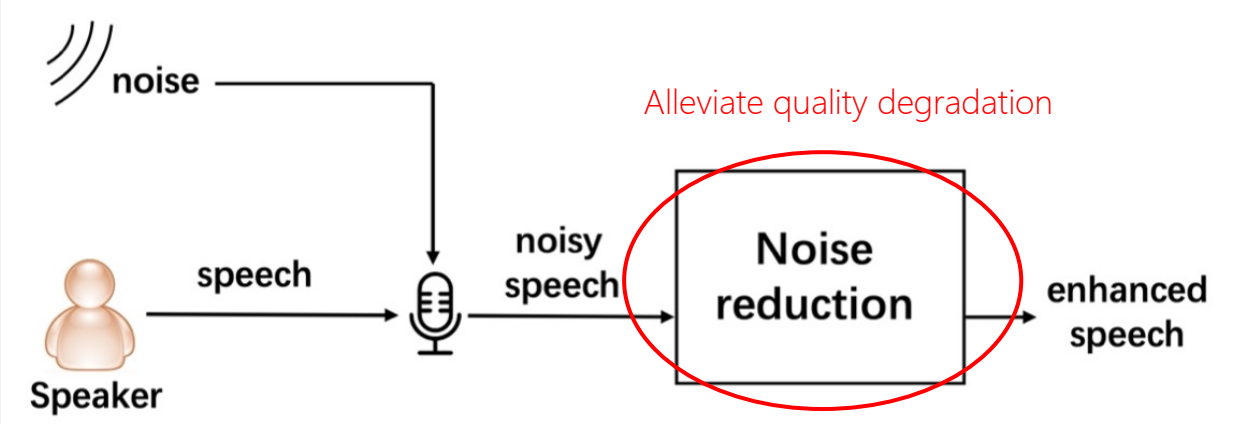
THESIS OUTLINE

- **Thesis issues**
  - For noise reduction:
    - Issue 1: Improving limited noise generalization capability of DNN-based noise reduction model **(Chapter 3)**



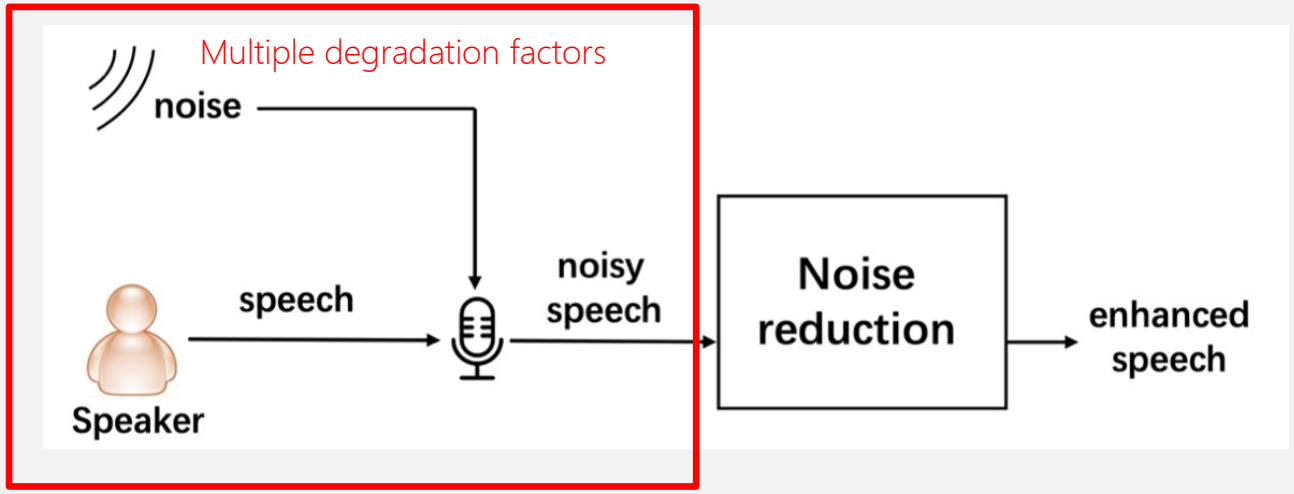
THESIS OUTLINE

- **Thesis issues**
  - For noise reduction:
    - Issue 1: Improving limited noise generalization capability of DNN-based noise reduction model (Chapter 3)
    - Issue 2: Alleviating speech quality degradation caused by inverse STFT with noisy phase (Chapter 3)



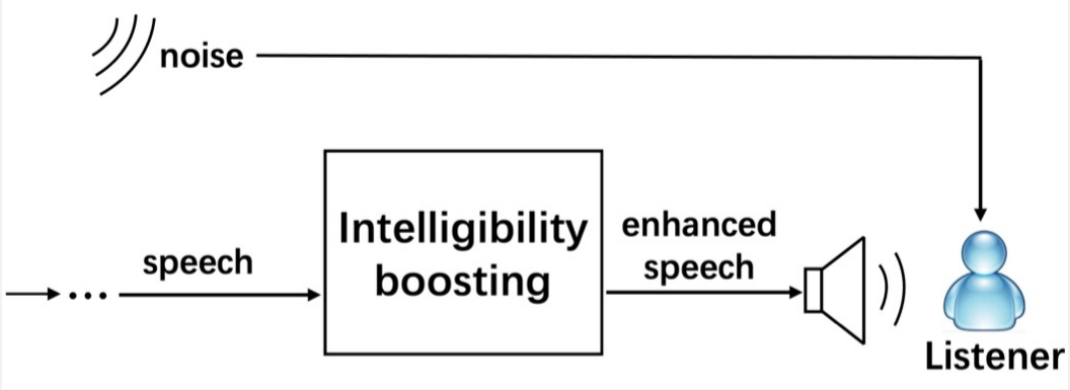
THESIS OUTLINE

- **Thesis issues**
  - For noise reduction:
    - Issue 1: Improving limited noise generalization capability of DNN-based noise reduction model (Chapter 3)
    - Issue 2: Alleviating speech quality degradation caused by inverse STFT with noisy phase (Chapter 3)
    - Issue 3: Improving noise reduction for device-degraded speech (Chapter 4)



THESIS OUTLINE

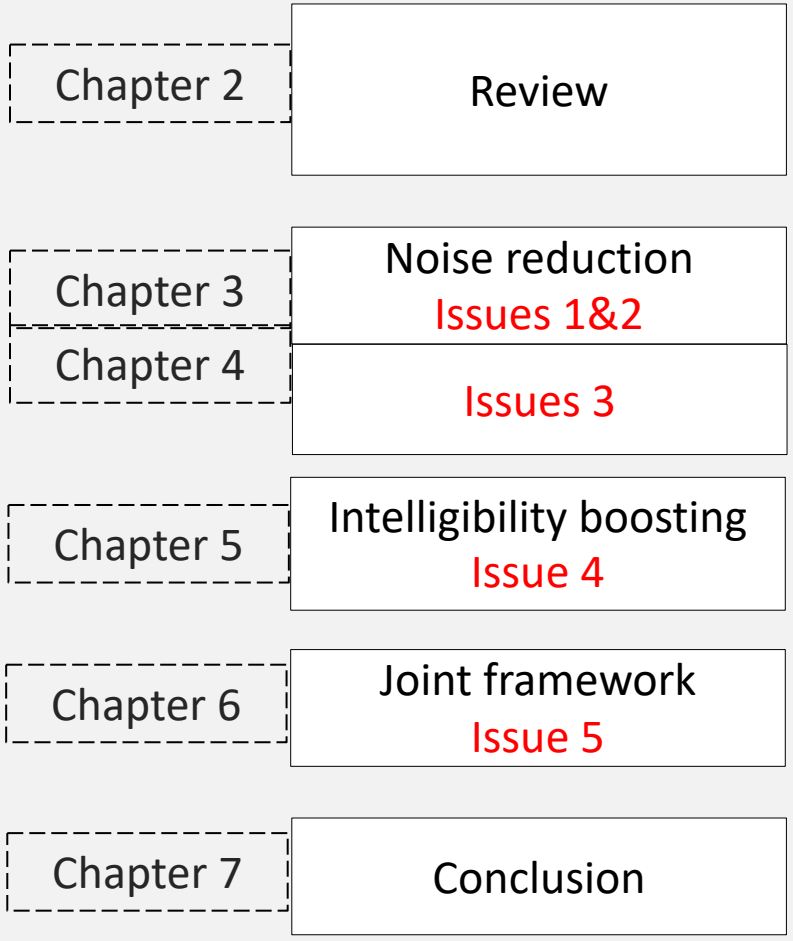
- **Thesis issues**
  - For intelligibility boosting:
    - Issue 4: Improving the performance of intelligibility boosting by leveraging deep learning (Chapter 5)



  - For joint noise reduction and intelligibility boosting:
    - Issue 5: integrating noise reduction with intelligibility boosting for full-end speech enhancement (Chapter 6)

THESIS OUTLINE

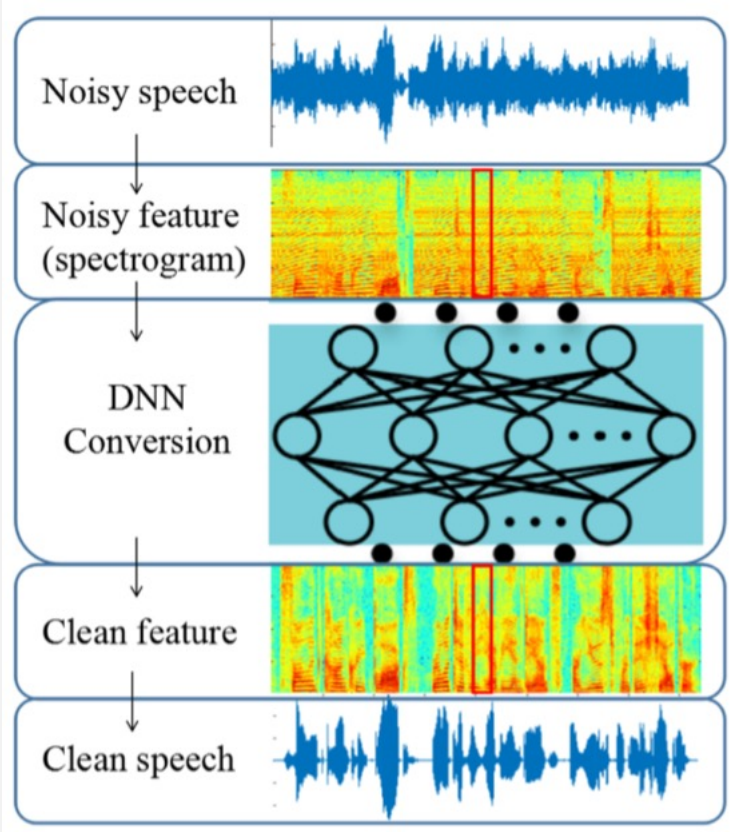
- Roadmap:



CONTENTS

ISSUES 1&2

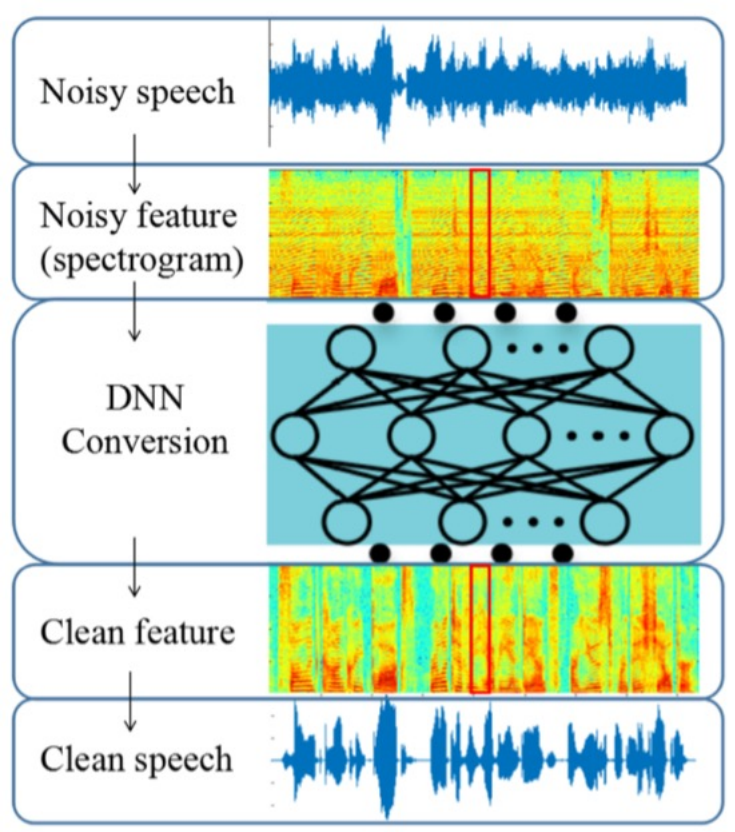
- **Background**
  - Neural noise reduction learns to convert noisy speech features into clean ones
  - Trained with pairs of clean speech and noisy speech
  - Data-driven



ISSUES 1&2

- **Background**
  - Neural noise reduction learns to convert noisy speech features into clean ones
  - Trained with pairs of clean speech and noisy speech
  - Data-driven

- **Evaluation metrics**
  - Objective quality: PESQ, CSIG, CBAK, COVL, etc.
  - Objective intelligibility: SIIB, HASPI, ESTOI, etc.
  - Subjective listening test: MOS score, preference test, etc.



ISSUES 1&2

- **Motivation**
  - Neural noise reduction learns to convert noisy speech features into clean ones
  - As a data-driven model, we are concerned its noise generalizability — Robustness to unseen noises **(issue 1)**
  - Conventional approaches operate only on magnitude spectrogram but disregard phase.

    While phase distortion degrades noise reduction performance **(issue 2)**

ISSUES 1&2

- **Motivation**
  - Neural noise reduction learns to convert noisy speech features into clean ones
  - As a data-driven model, we are concerned its noise generalizability — Robustness to unseen noises **(issue 1)**
  - Conventional approaches operate only on magnitude spectrogram but disregard phase.

    While phase distortion degrades noise reduction performance **(issue 2)**
- **Solutions:**
  - Introduce neural noise temples to improve noise generalizability

ISSUES 1&2

- **Motivation**
  - Neural noise reduction learns to convert noisy speech features into clean ones
  - As a data-driven model, we are concerned its noise generalizability — Robustness to unseen noises **(issue 1)**
  - Conventional approaches operate only on magnitude spectrogram but disregard phase.

    While phase distortion degrades noise reduction performance **(issue 2)**
- **Solutions:**
  - Introduce neural noise temples to improve noise generalizability
  - Investigate applying neural vocoder, instead of ISTFT, for waveform synthesis to alleviate phase distortion

ISSUES 1&2: METHOD

- Key idea:
  - Model noise dynamics and feed it into network training, known as Noise Aware Training (NAT) [1]
  - Conventional NAT relied on separate module to get noise estimation, might be suboptimal

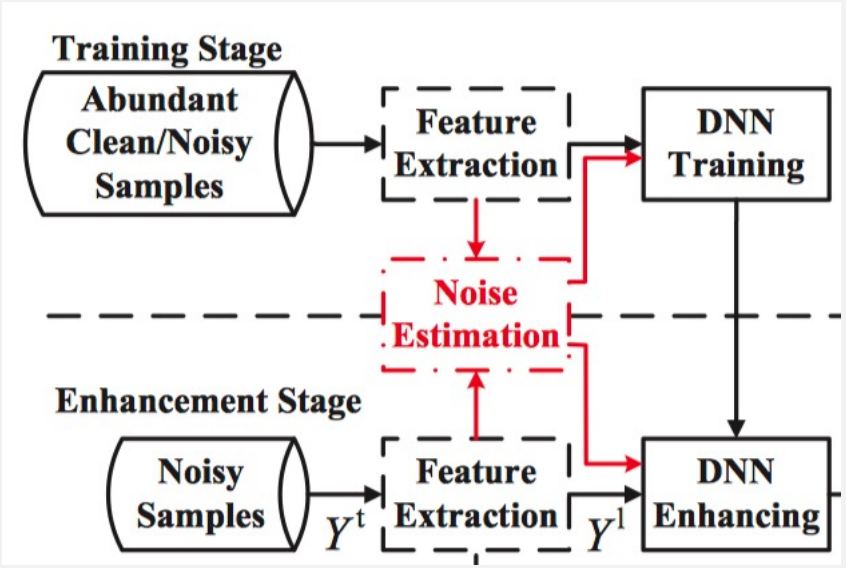


Figure from [1]

[1] Xu, Y., Du, J., Dai, L.R. and Lee, C.H., 2014. Dynamic noise aware training for speech enhancement based on deep neural networks. In Fifteenth Annual Conference of the International Speech Communication Association.

ISSUES 1&2: METHOD

- Key idea:
  - Model noise dynamics and feed it into network training, known as Noise Aware Training (NAT) [1]
  - Conventional NAT relied on separate module to get noise estimation, might be suboptimal
  - Propose trainable neural noise template (*noise tokens*)

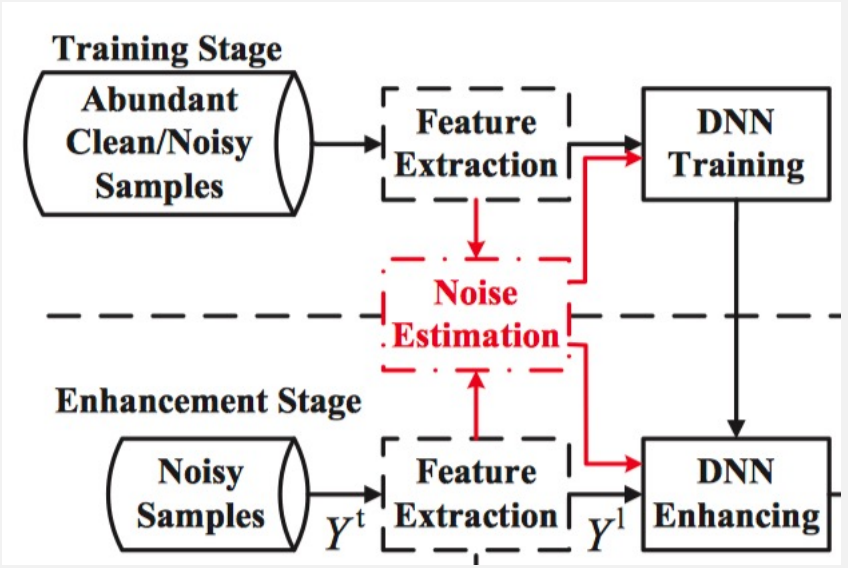
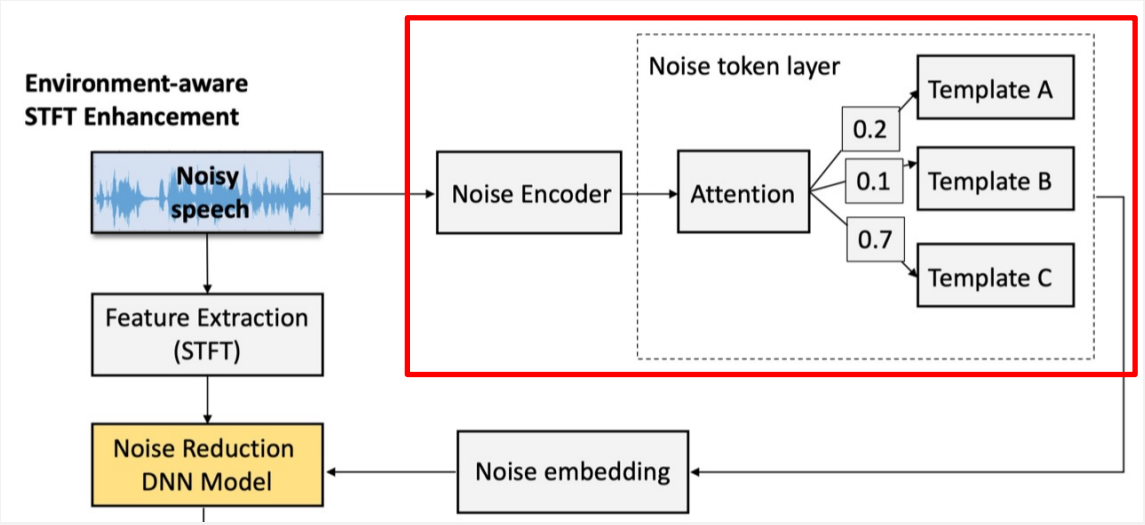


Figure from [1]

[1] Xu, Y., Du, J., Dai, L.R. and Lee, C.H., 2014. Dynamic noise aware training for speech enhancement based on deep neural networks. In Fifteenth Annual Conference of the International Speech Communication Association.

ISSUES 1&2: METHOD

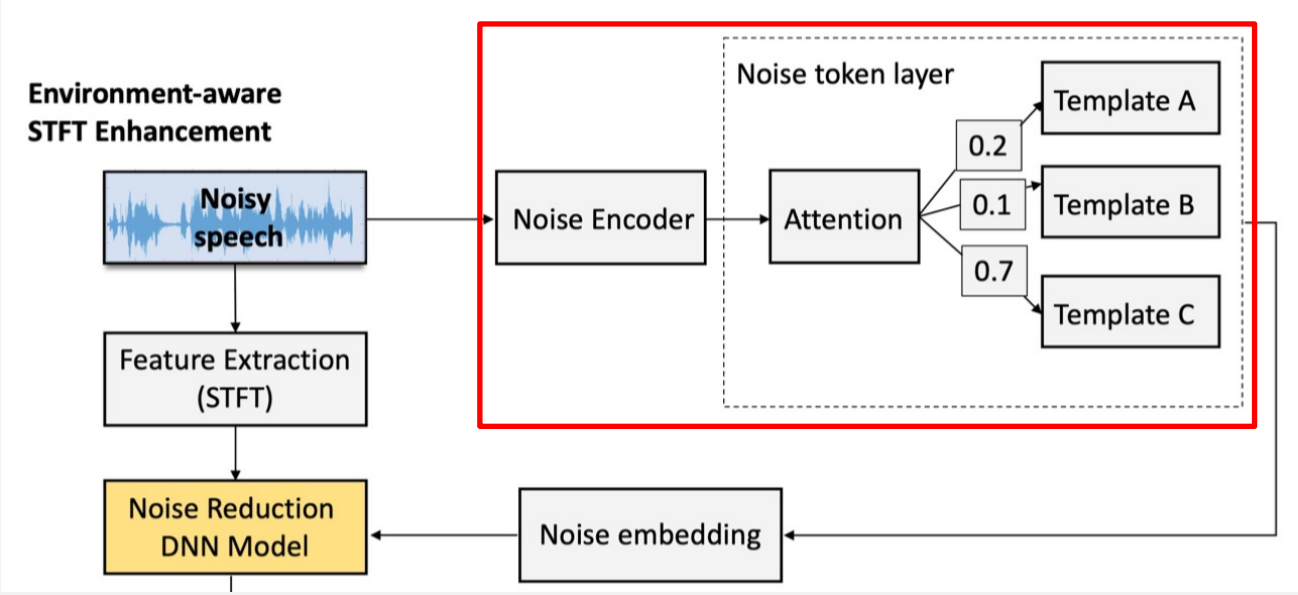
- Neural noise templates —— Noise tokens:
  - Inspired by Style token [2]; Generated noise embedding is in frame-level to capture noise dynamics
  - **Intuition**: Learn noise latent -> Represent new noise into the combination of old templates by assigning weights -> Improve generalization



[2] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y. and Saurous, R.A., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

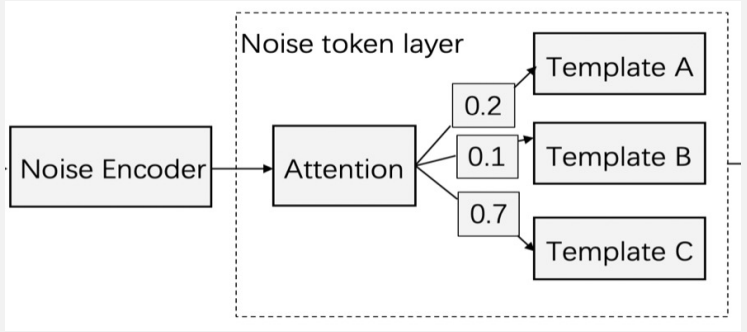
ISSUES 1&2: METHOD

- Neural noise templates —— Noise tokens:
  - Noise tokens are jointly optimized with the whole system, and can be regarded as dictionaries (*analogy to noise dictionaries in NMF-based approach)
  - Unseen noises are factorized and then represented as the weighted sum of trained noise tokens through attention module
  - Unseen noise --> linear combination of trained templates (which are seen to the model)



ISSUES 1&2: EXPERIMENT

- Experiment 1: Performance analysis with noise tokens
  - Noise token module: 16 learnable tokens; Multi-head attention with 8 heads
  - Data:
    - 50-hour noisy speech with 21 training noise types
    - Tested on 4 unseen noises; SNR={-2.5, 2.5, 7.5, 12.5, 17.5} dB
  - Three DNN architectures:
    - Standard 2-layer **BLSTM**
    - CNN+BLSTM used in **VoiceFilter [3]**
    - Gaussian-weighted Transformer used in **T-GSA** [4] (Simplified by reducing number of blocks)



[3] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," arXiv preprint arXiv:1810.04826

[4] J. Kim, M. El-Khamy, and J. Lee, "Transformer with gaussian weighted self-attention for speech enhancement," arXiv preprint arXiv:1910.06762, 2019.

ISSUES 1&2: EXPERIMENT

- **Result 1: Performance analysis with noise tokens**
  - Three tested systems (all used ISTFT for synthesis):
    - w/o noise embedding
    - DNAT (Dynamic noise aware training): w/ noise embedding; Obtained by a separate noise estimation module using a noise tracking algorithm
    - NTs (Noise tokens): w/ noise embedding; Obtained by proposed noise token module

Architectures	w/o embedding		with DNAT		with NTs	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
BLSTM	2.686	0.896	2.692	0.898	**2.858**	**0.914**
VoiceFilter	2.792	0.904	2.771	0.902	**2.907**	**0.916**
T-GSA	2.754	0.906	2.726	0.902	**2.808**	**0.912**

ISSUES 1&2: EXPERIMENT

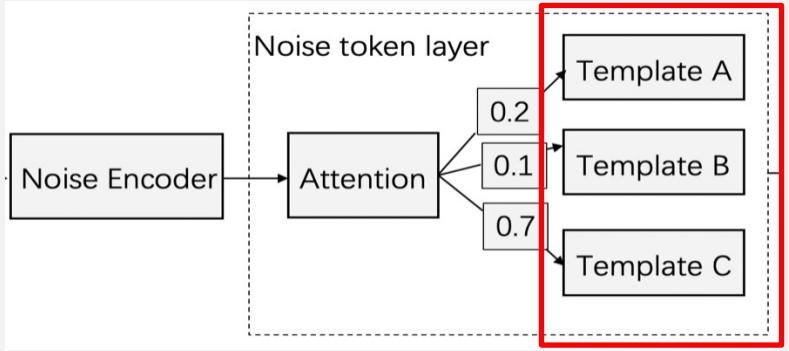
- Result 1: Performance analysis with noise tokens
  - Three tested systems (all used ISTFT for synthesis)

➡ Noise token consistently improves the noise reduction performance across all three tested architectures under unseen noise conditions

Architectures	w/o embedding		with DNAT		with NTs	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
BLSTM	2.686	0.896	2.692	0.898	**2.858**	**0.914**
VoiceFilter	2.792	0.904	2.771	0.902	**2.907**	**0.916**
T-GSA	2.754	0.906	2.726	0.902	**2.808**	**0.912**

ISSUES 1&2: EXPERIMENT

- Experiment 2: Impact of noise diversity
  - Effect of learned noise tokens:
    - We expect that learned tokens can model or capture noise patterns
    - More noise types fed into training, learned tokens have better noise representation ability
  - Data:
    - Four noisy datasets, all have 50 hours duration
    - Only differed in the number of noise types included
    - The more types a dataset includes, the more diverse it is
    - Generated {N7, N12, N16, N21}, each includes {7, 12, 16, 21} noise types
    - N21 is the same as what we used in Experiment 1, and has the best noise diversity



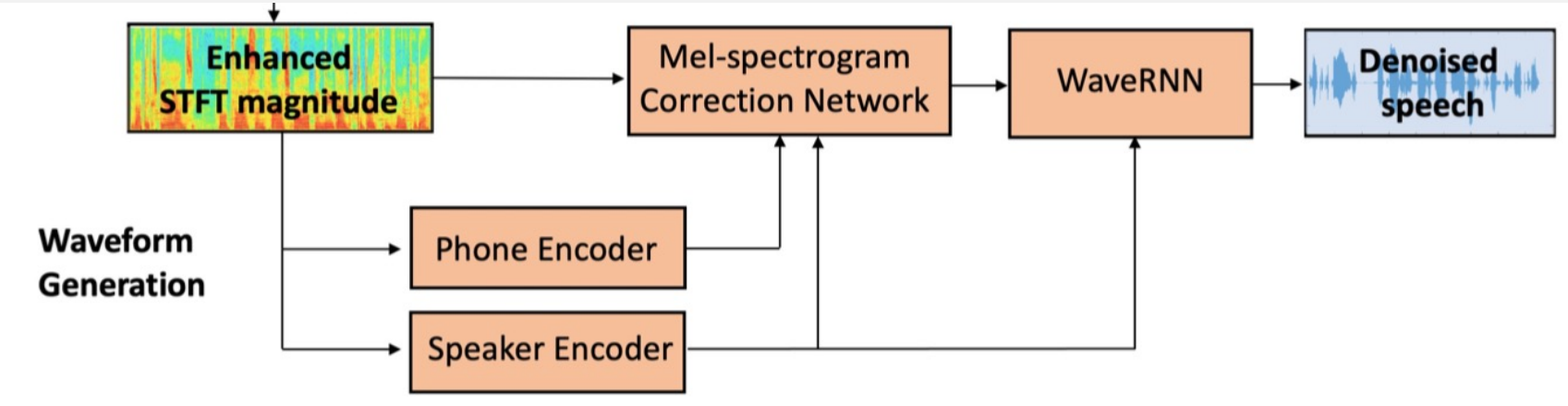
ISSUES 1&2: EXPERIMENT

- Result 2: Impact of noise diversity
  - Both used ISTFT for waveform synthesis
  - Noise tokens bring higher relative improvements on PESQ with increasing noise diversity
  - Noise tokens can effectively exploit multiple noises due to the modelling ability of their trainable templates

Noise corpus	BLSTM w/o NTs		BLSTM with NTs	
	PESQ	Relative imp.	PESQ	Relative imp.
N7	2.564	0.00%	**2.657**	**0.00%**
N12	2.639	2.94%	**2.786**	**4.86%**
N16	2.672	4.20%	**2.812**	**5.82%**
N21	2.686	4.71%	**2.858**	**7.54%**

ISSUES 1&2: EXPERIMENT

- Experiment 3: Vocoder-based waveform generation module
  - Apply WaveRNN [5] vocoder to synthesize waveform
    - Enhanced magnitude spectrogram is converted to 80-dim Mel-spectrogram via an additional NN to (1) further suppress residual noises; and (2) as input for WaveRNN
    - WaveRNN directly generates waveforms to avoid incorporating noisy phase
    - WaveRNN is pre-trained as a speaker-independent model using VCTK corpus



[5] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury,N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord,S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," arXiv preprint arXiv:1802.08435, 2018.

ISSUES 1&2: EXPERIMENT

- **Result 3: Objective evaluations for vocoder-based waveform generation module**
  - Vocoder-based generation is much worse than the traditional ISTFT, in terms of objective measures
  - Why? --> Probable reason: PESQ and STOI are not designed to evaluate neural vocoders

Methods	PESQ	STOI
Noisy	2.021	0.833
NT-ISTFT	**2.858**	**0.914**
NT-WG	2.509	0.867

Use ISTFT to synthesize → NT-ISTFT

Use WaveRNN to synthesize → NT-WG

  - Let's look into listening test results!

ISSUES 1&2: EXPERIMENT

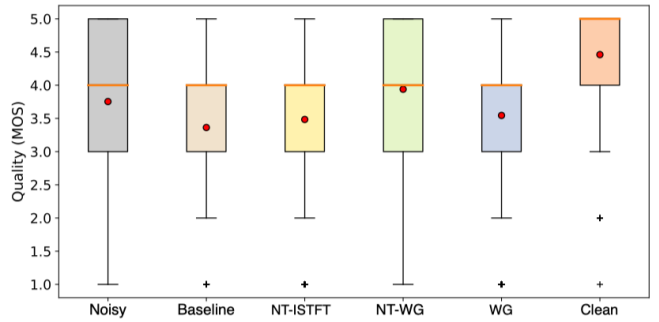
- Experiment 4: Listening tests
  - Six evaluated systems
    - Baseline: Standard BLSTM model; (Use ISTFT to generate waveform)
    - NT-ISTFT: BLSTM model + Noise tokens; Use ISTFT to generate waveform
    - NT-WG: BLSTM model + Noise tokens; Use WaveRNN to generate waveform -> expected to be the best
    - WG: Directly apply Waveform generation module to raw noisy speech; Use WaveRNN
    - Clean: Raw clean speech
    - Noisy: Unprocessed noisy speech

ISSUES 1&2: EXPERIMENT

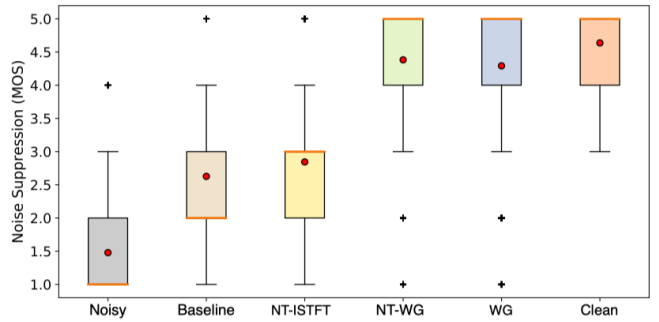
- **Experiment 4: Listening tests**
  - Six evaluated systems
    - Baseline: Standard BLSTM model; (Use ISTFT to generate waveform)
    - NT-ISTFT: BLSTM model + Noise tokens; Use ISTFT to generate waveform
    - NT-WG: BLSTM model + Noise tokens; Use WaveRNN to generate waveform
    - WG: Directly apply Waveform generation module to raw noisy speech; Use WaveRNN
    - Clean: Raw clean speech
    - Noisy: Unprocessed noisy speech
  - Rated in three aspects (521 listeners participated)
    - (1) speech quality; (2) noise suppression; and (3) overall performance

ISSUES 1&2: EXPERIMENT

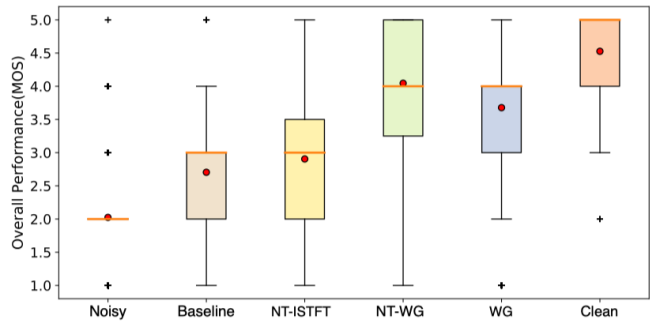
- **Result 4: Listening tests**
  - NT-ISTFT outperforms Baseline in all three scores
  - Compared to NT-ISTFT, NT-WG shows higher performances, especially on the noise suppression score (significant)



(a) Results on speech quality



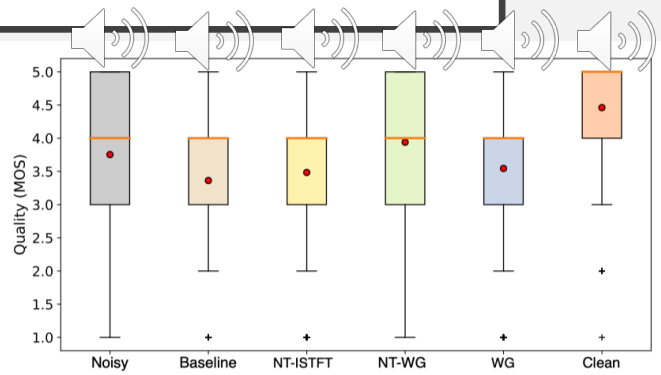
(b) Results on noise suppression



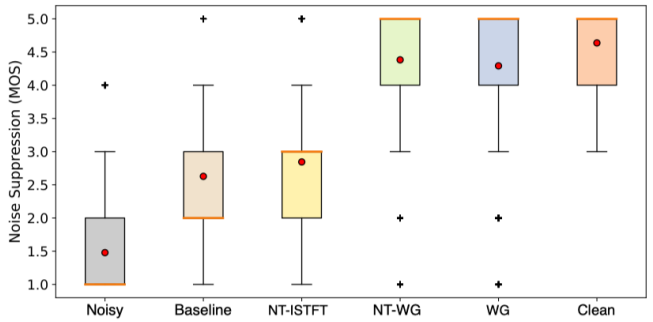
(c) Results on overall performance

ISSUES 1&2: EXPERIMENT

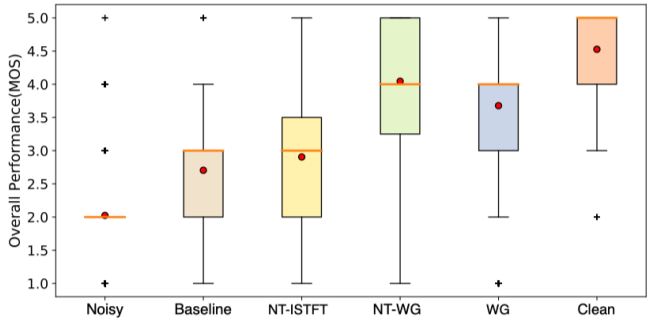
- **Result 4: Listening tests**
  - NT-ISTFT outperforms Baseline in all three scores
  - Compared to NT-ISTFT, NT-WG shows higher performances, especially on the noise suppression score (significant)
  - Surprisingly, WG module itself can even outperform NT-ISTFT
  - However, we found there are limited but some very bad-quality cases (mumbling) in vocoder-generated systems.



(a) Results on speech quality



(b) Results on noise suppression



(c) Results on overall performance

ISSUES 1&2: SUMMARY

- **Propose noise token model for issue 1**
  - To alleviate the noise mismatch problem of DNN-based noise reduction model
  - Noise token is effective across different network architectures and brings higher performance growth with increasing noise diversity
- **Propose waveform generation module for issue 2**
  - To synthesize the waveform using WaveRNN vocoder, instead of traditional ISTFT
  - Subjective listening tests show that the residual noise can be significantly reduced by the waveform generation module

CONTENTS

- Introduction
- Issues 1&2: Improve noise generalization and alleviate phase distortion
- Issue 3: Improve noise reduction for device-degraded speech
- Issues and methods
- Conclusion & Future Work
- Publication list

ISSUE 3: MOTIVATION

- **Transform device-degraded speech into high-quality ones**
  - In issues 1&2 (Chapter 3), only additive noise was considered
  - General device-degraded speech features: background noise, room reverb, and bad microphone response.

Degraded version iPhone-bedroom:  Studio: 

  - These factors are jointly considered. We collectively refer to as the ***channel factor***
  - Enhance these recordings by simultaneously removing noise, reverb, and also applying pleasing audio effect via a unified network

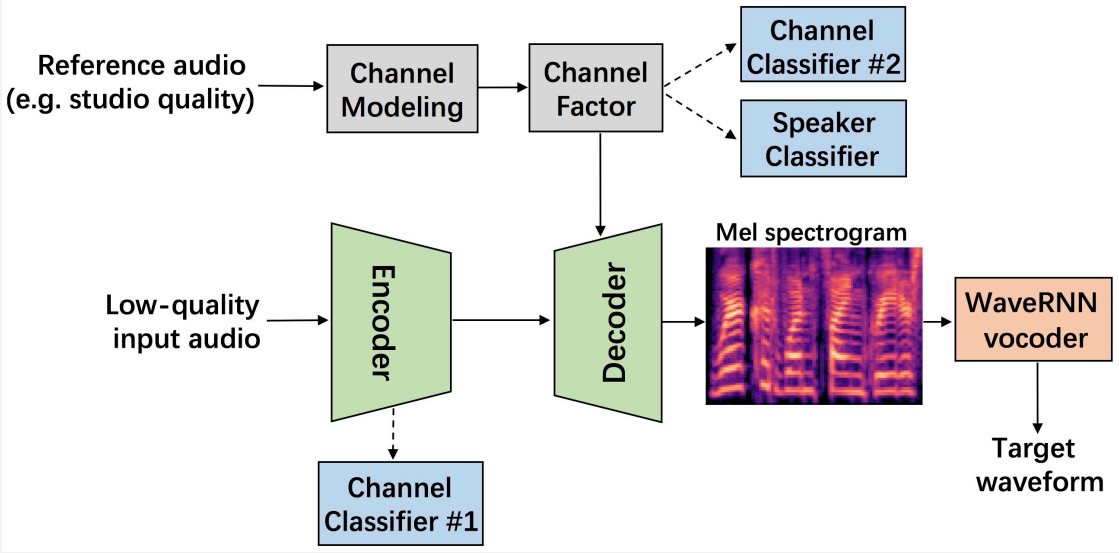
ISSUE 3: MOTIVATION

- **Transform device-degraded speech into high-quality ones**
  - In issues 1&2 (Chapter 3), only additive noise was considered
  - General device-degraded speech features: background noise, room reverb, and bad microphone response.
  - These factors are jointly considered. We collectively refer to as the *channel factor*
  - Enhance these recordings by simultaneously removing noise, reverb, and also applying pleasing audio effect via a unified network

- **Explore TTS techniques on noise reduction task**
  - Regard this task as a style transfer task, from low quality style to high quality
  - Apply neural waveform model to synthesize speech, instead of using ISTFT

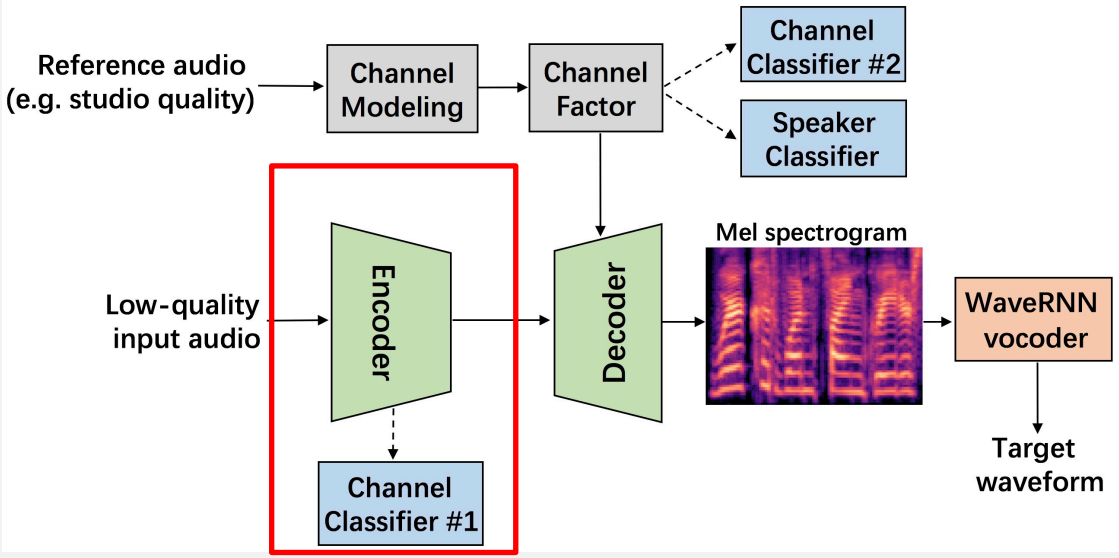
ISSUE 3: METHOD

- **Overview of system diagram**
  - Encoder
    - Filter out the channel characteristics from the input audio
  - Channel Modeling
    - Disentangle the channel factor from a reference audio
  - Decoder
    - Predict the target-style Mel spectrogram, conditioned on extracted channel factor
  - WaveRNN vocoder
    - Generate target-style waveform (professional high-quality recording)



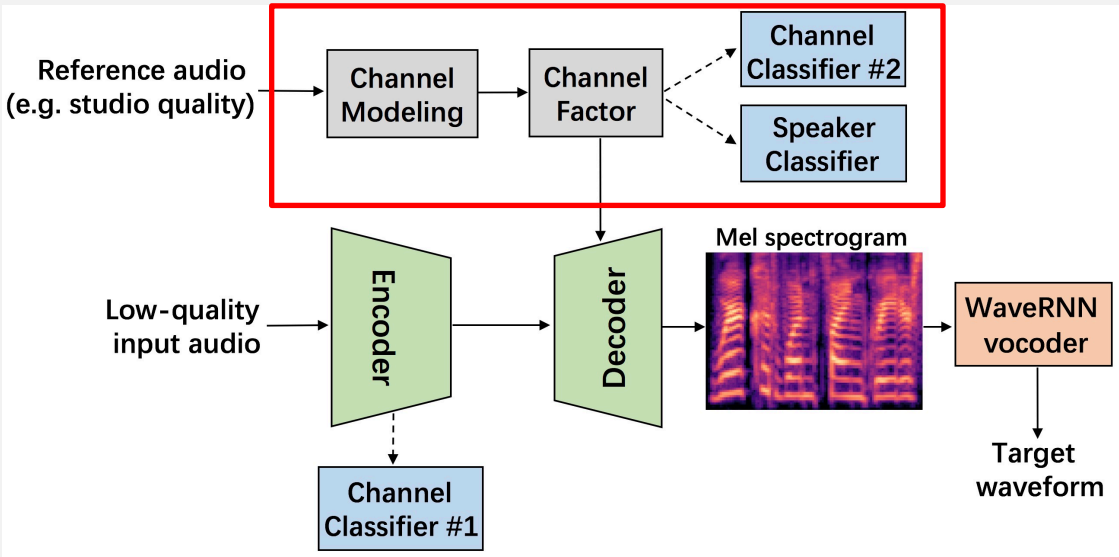
ISSUE 3: METHOD

- Component details
  - Encoder
    - Filter out the channel characteristics from the input audio
    - Consists of 2-D CNNs+BLSTM
  - Adversarial training
    - Add channel classifier #1 to encourage encoder to produce channel-invariant features



ISSUE 3: METHOD

- **Component details**
  - Channel modeling
    - Disentangle the channel factor from a reference audio

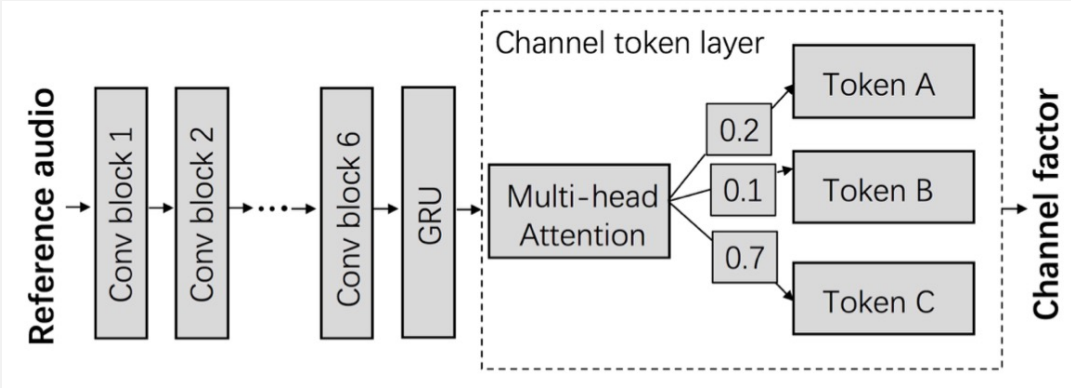


  - Additional classifiers
    - Channel classifier #2 used to encourage extracted channel factor to be more informative about channel information
    - Speaker classifier used for adversarial training, to filter out the remained speaker information from the channel factor

ISSUE 3: METHOD

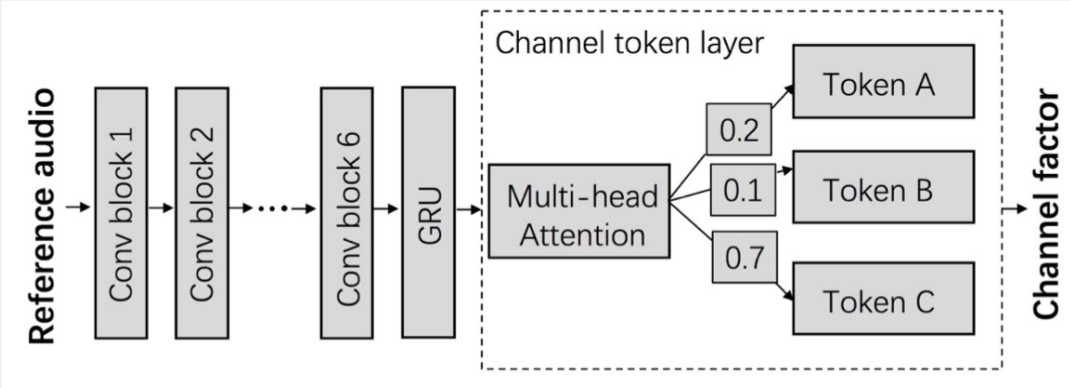
- **Component details**
  - Channel modeling
    - Shares a similar network structure with “Noise Token”
    - Design an interpretable and controllable channel modeling module. (e.g.,

      Token A might represent reverb level, Token B represents noise level, etc.)



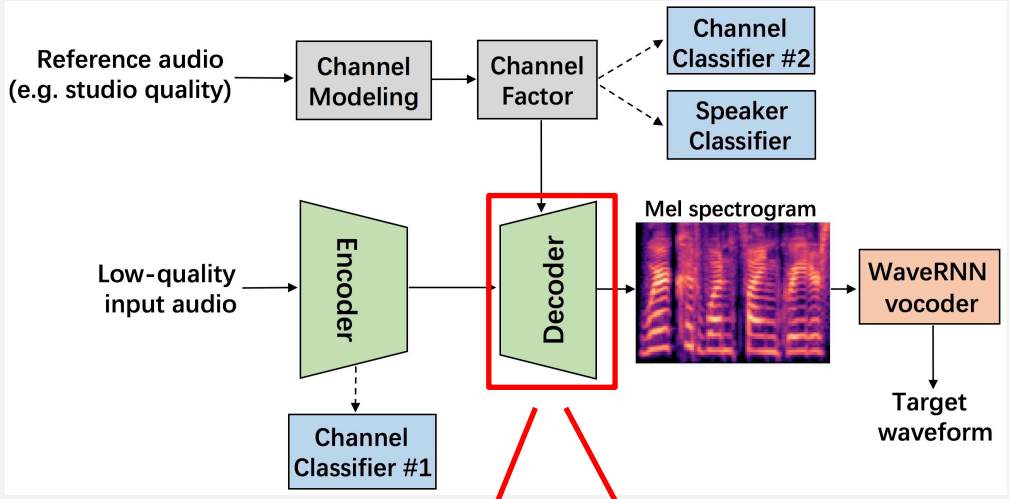
ISSUE 3: METHOD

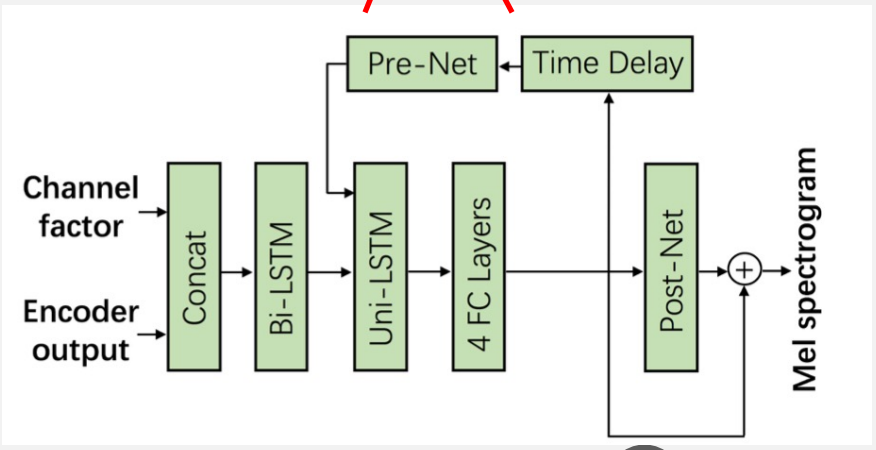
- **Component details**
  - Pros
    - Enables module to deal with the unseen channel condition and unlabeled reference audio
    - Controllable style transfer by adjusting weights of learned tokens
  - Cons
    - Need an additional provided reference audio
    - Bad performance if channel factor not accurate



ISSUE 3: METHOD

- **Component details**
  - Decoder
    - Predict the target-style Mel spectrogram, conditioned on extracted channel factor
    - Similar structure with Tacotron2-Decoder, including Prenet, Postnet, and auto-regressive generation
  - WaveRNN vocoder
    - A pre-trained universal WaveRNN vocoder



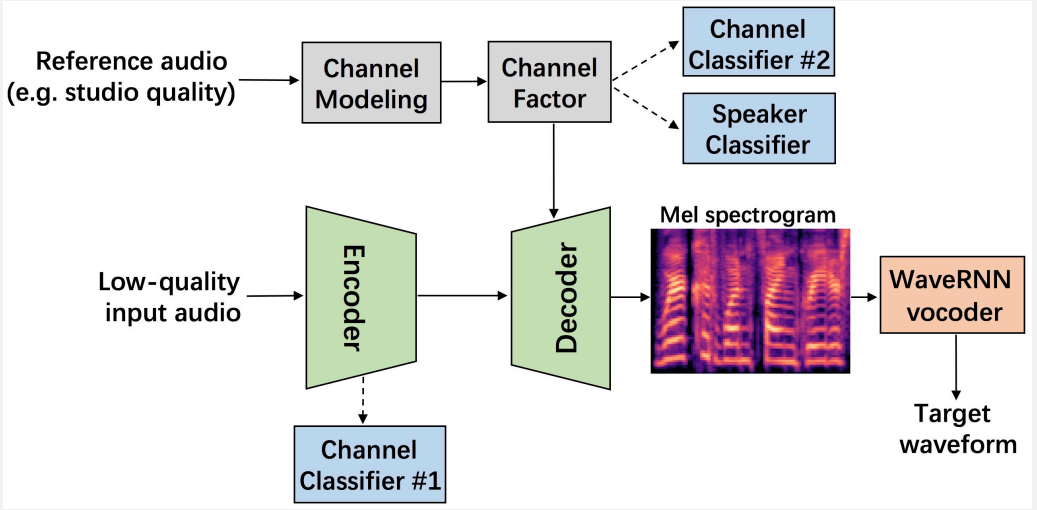


ISSUES 3: EXPERIMENTS

- Dataset
  - DAPS (device and produced speech) dataset
  - It provides aligned recordings of high-quality speech and a number of versions of low-quality speech, recorded in noisy environment with cheap device.
  - Two unseen speakers (1 male + 1 female), and three unseen channels are used for testing: (1) ipad_livingroom, (2) ipadflat_office, and (3) iphone_bedroom

ISSUES 3: EXPERIMENTS

- Ablation Study
  - ED: contains only encoder and decoder
  - ED+CM: contains encoder, decoder, and channel modelling
  - FULL (ED+CM+Classifiers): contains encoder, decoder, channel modelling, and 3 auxiliary classifiers
  - Linear+ISTFT: Same settings with FULL model, except the decoder output was linear spectrogram. Use ISTFT to synthesize waveform



ISSUES 3: EXPERIMENTS

- Other compared methods
  - Raw audio: lower bound
  - Studio audio: higher bound
  - WPE: signal-processing method for speech dereverberation
  - WPE+L: signal-processing method for speech dereverberation + LogMMSE for denoising
  - WaveNet [1]: Denoising-WaveNet model

[1] Jiaqi Su, Adam Finkelstein, and Zeyu Jin, "Perceptually-motivated environment-specific speech enhancement," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 7015–7019

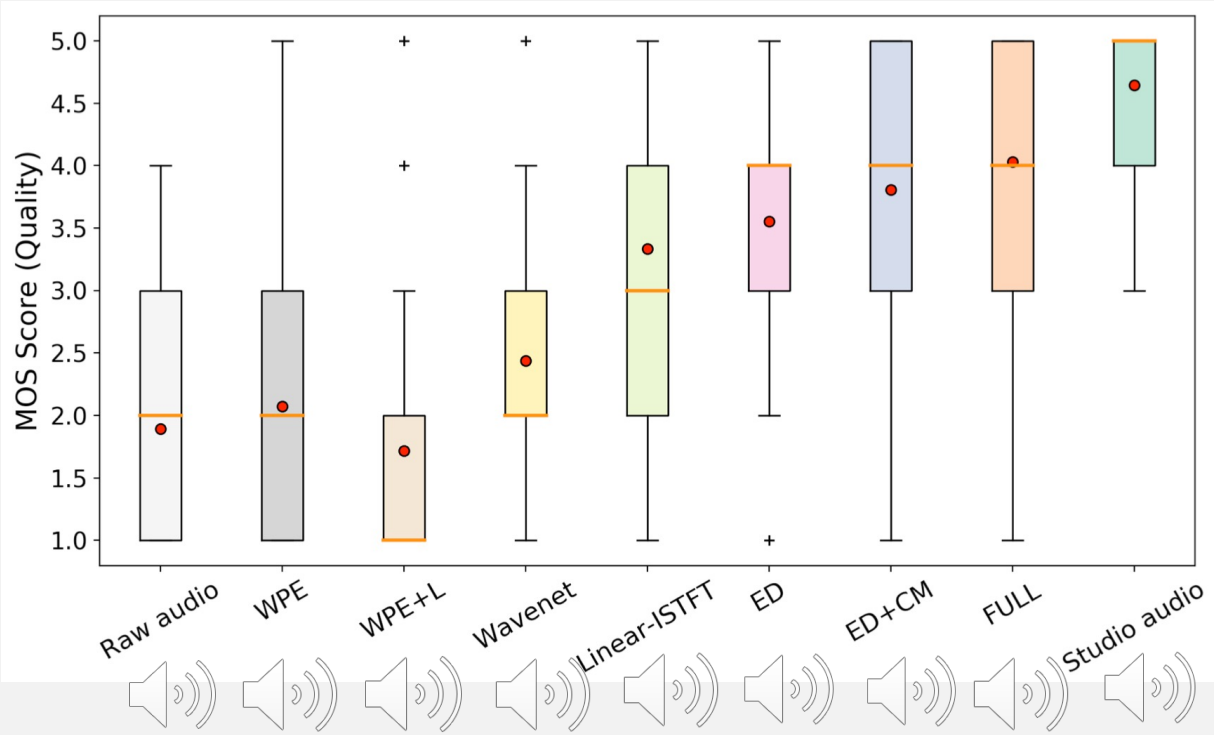
ISSUES 3: EXPERIMENTS

- Objective results
  - FULL consistently improves its two simplified versions, ED and ED+CM, and other compared methods (WPE, WPE+L, and WaveNet)
  - FULL system is worse than Linear-ISTFT in terms of CBAK and COVL
  - Objective metrics usually give lower scores to vocoder-generated waveform

System	CSIG	CBAK	COVL	STOI
Raw audio	3.05	2.23	2.60	0.869
WPE	3.16	2.41	2.75	0.888
WPE+L	2.81	2.33	2.52	0.811
Wavenet	3.67	2.42	3.08	0.904
Linear-ISTFT	3.94	2.61	3.37	0.905
ED	3.89	2.48	3.28	0.906
ED+CM	3.73	2.49	3.16	0.886
FULL	3.94	2.52	3.34	0.906

ISSUES 3: EXPERIMENTS

- Subjective results
  - Conducted crowdsourced listening tests, 165 individuals rated quality for given samples with 5-point MOS score
  - FULL gives the best performance
  - FULL > Linear-ISTFT, means WaveRNN improves the quality of the synthetic waveform, compared with ISTFT



  - Audio samples: https://nii-yamagishilab.github.io/hyli666-demos/evr-slt2021/

ISSUES 3: SUMMARY

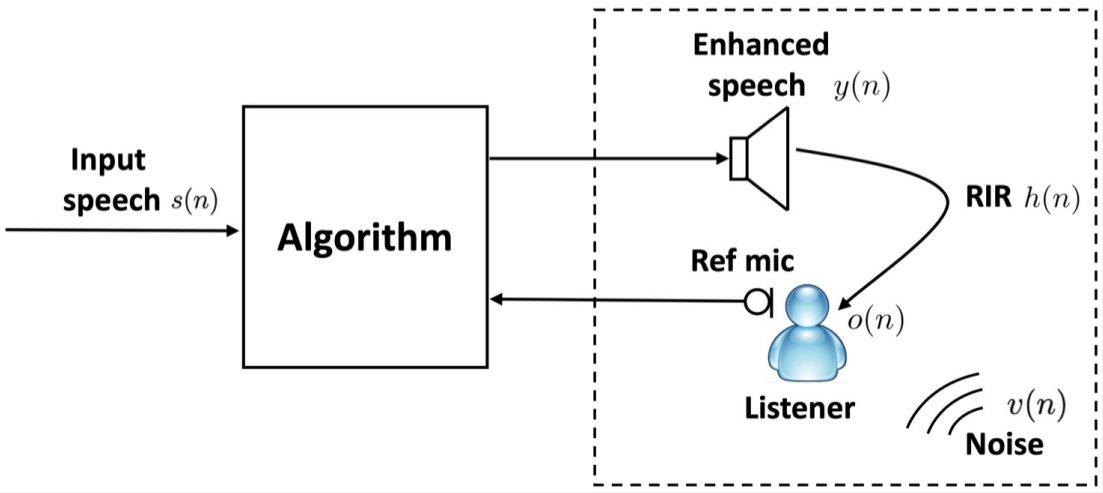
- Apply style transfer approach into noise reduction task
  - For device-degraded speech
  - Jointly consider denoising, dereverberation, and applying pleasing audio effect to low-quality recordings
  - Mel+WaveRNN waveform synthesis module outperforms Linear+ISTFT in subjective evaluations
  - Proposed system outperforms a time-domain model (WaveNet) and several signal-processing baselines

- However…
  - Still require expensive parallel recordings for training -> Expand to non-parallel style transfer?

CONTENTS

ISSUE 4: BACKGROUND

- Speech intelligibility degrades in noisy environment
  - Cause stressful listening; or even non-understanding for listener
  - Unlike noise reduction, we cannot suppress noise since it is physically present
- Solution
  - Can we simply increase voice volume? -> Yes, but loud voice causes uncomfortable listening
  - We seek methods to modify speech signal but without changing its energy (i.e., volume)
  - Applications: mobile telephony, public-address announcement, etc.



ISSUE 4: METHOD

- **How to modify speech?**
  - Reallocate the speech energy on time-frequency domain in such a way as to boost the acoustic cues that are perceptually crucial
  - But.. How do we know what kind of acoustic cue is crucial for human perception?

ISSUE 4: METHOD

- **How to modify speech?**
  - Reallocate the speech energy on time-frequency domain in such a way as to boost the acoustic cues that are perceptually crucial
  - But.. How do we know what kind of acoustic cue is crucial for human perception?
- **Rely on objective speech intelligibility metrics**
  - Many objective metrics have been proposed to predict the intelligibility of speech
  - Specifically, they require a clean speech signal as the reference to predict intelligibility for distorted (speech-in-noise) speech
  - Our goal is thus formulated as: Modifying speech to maximize its objective intelligibility scores

ISSUE 4: METHOD

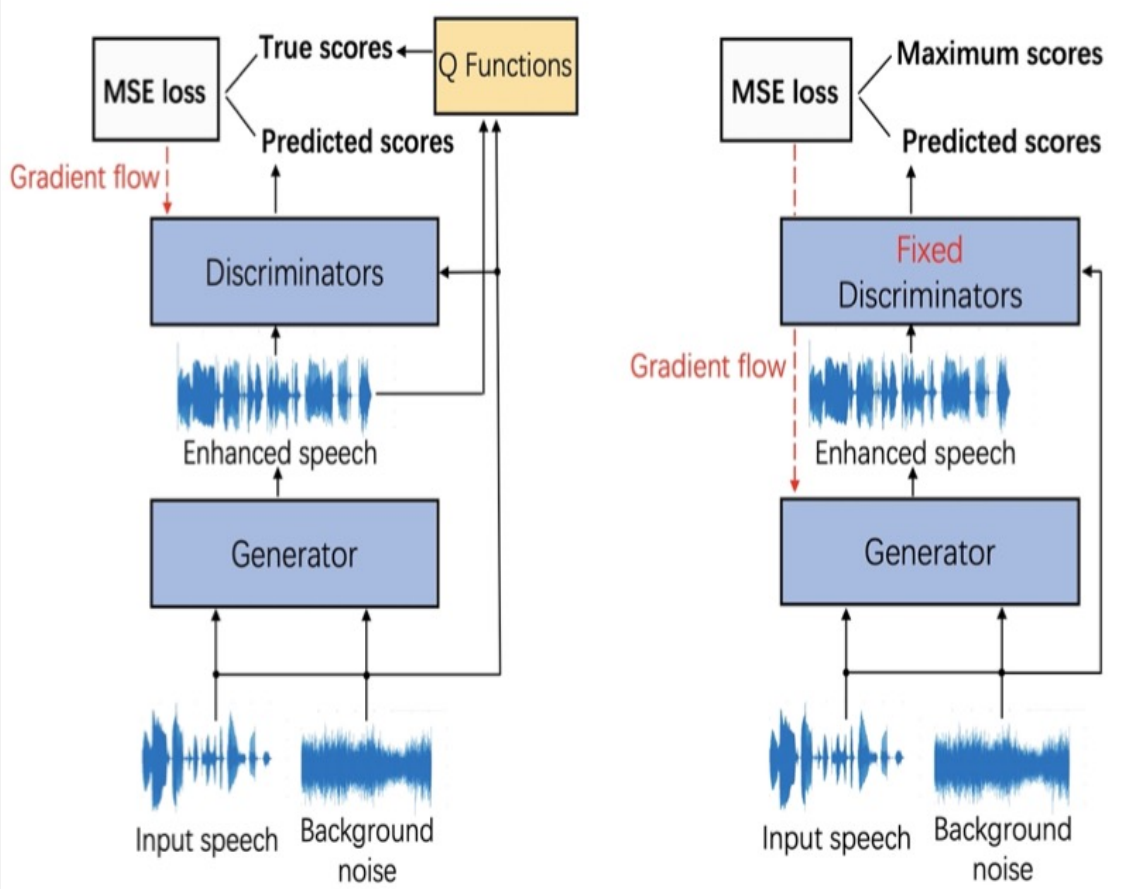
- **Difficulties**
  - Intelligibility metrics are usually quite complex and mathematically intractable
  - Hard to solve such an optimization problem in an efficient manner

ISSUE 4: METHOD

- **Difficulties**
  - Intelligibility metrics are usually quite complex and mathematically intractable
  - Hard to solve such an optimization problem in an efficient manner
- **Solution -> Try GAN**
  - We introduce GAN model into our system
  - Replace intractable objective metrics with differentiable neural network
  - Jointly optimize three intelligibility metrics (SIIB, HASPI, and ESTOI) for improved intelligibility
  - Also optimize two quality metrics (PESQ and ViSQOL) for improved speech quality
  -

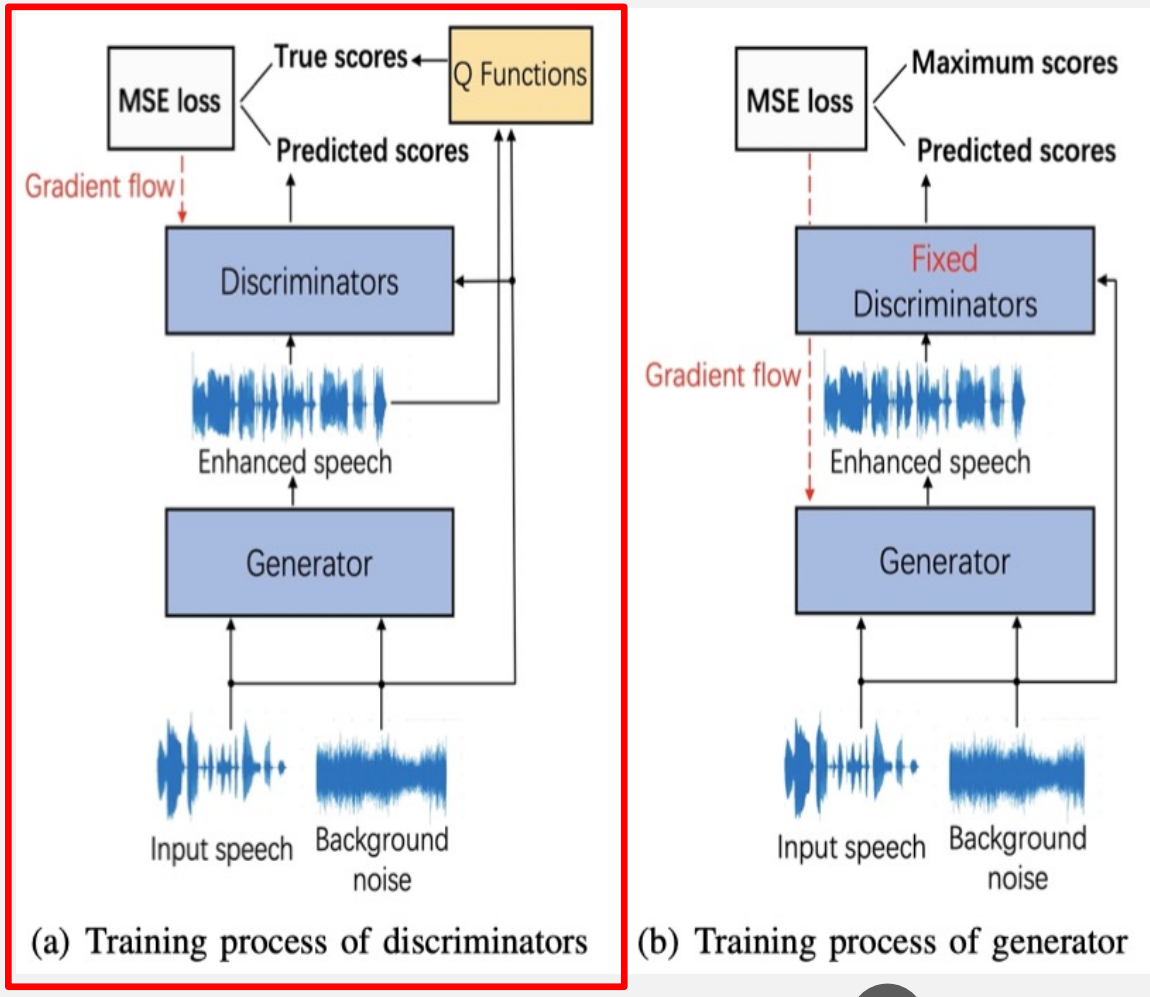
ISSUE 4: METHOD

- System overview
  - Generator (G) enhances the speech signal as the intelligibility enhancement module
  - Discriminator (D) learns to predict the intelligibility scores of modified speech
  - G is then trained with the guidance of D



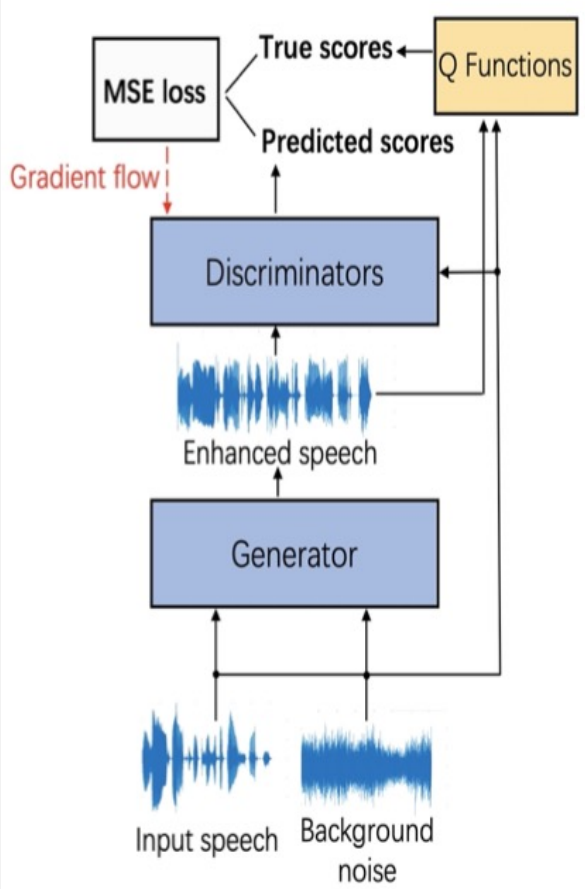
ISSUE 4: METHOD

- Discriminator
  - Used to approximate intractable metrics
  - Q function represents metrics to be modelled, i.e., SIIB, HASPI, and ESTOI
  - D tries to accurately predict the score of the modified speech by minimizing MSE between predicted and true scores
  - Optimize true intelligibility metrics (Q function) = Optimize output scores of D
  - This is possible since D is differentiable now

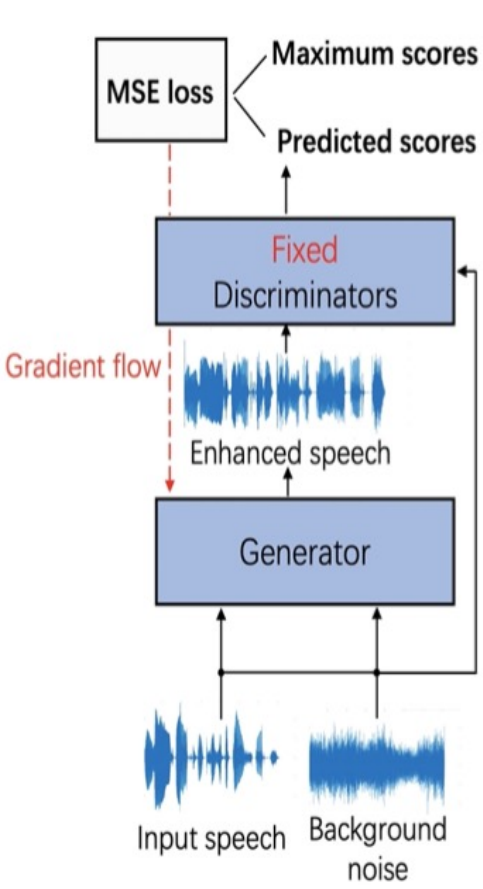


ISSUE 4: METHOD

- Generator
  - Used to modify natural speech as intelligibility enhancement module
  - D parameters are fixed, and G is trained to reach the intelligibility scores as high as possible
  - MSE between predicted and target maximum scores is set as G loss



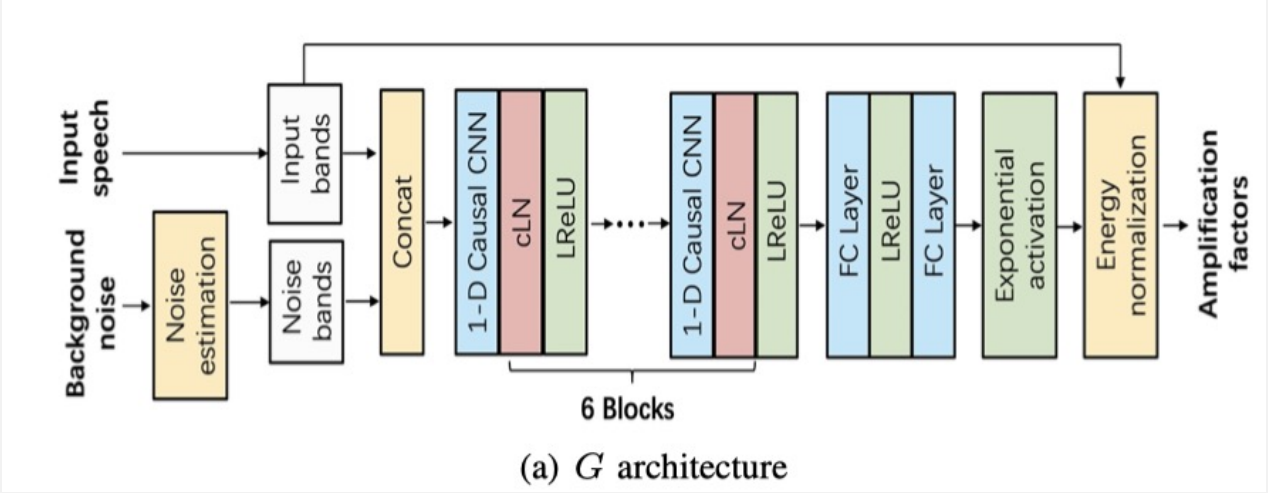
(a) Training process of discriminators



(b) Training process of generator

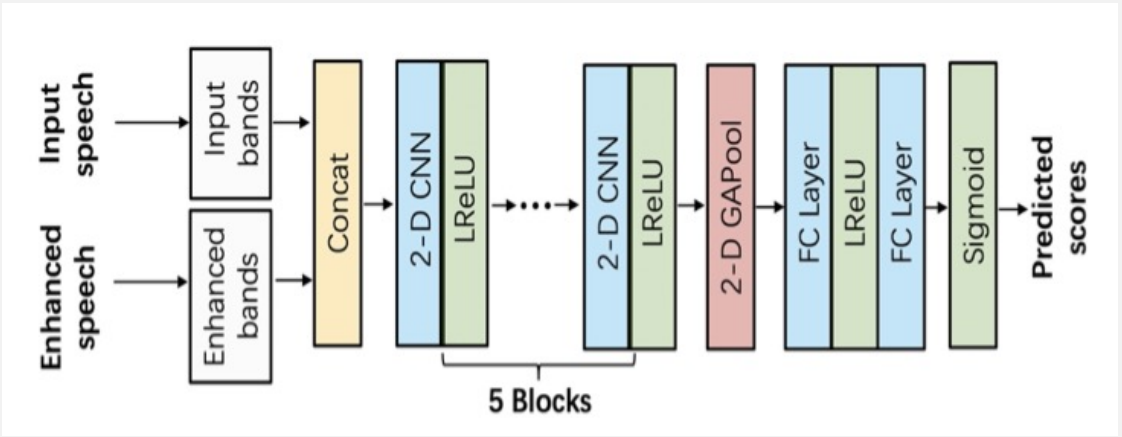
ISSUE 4: METHOD

- **Model architecture**
- **Generator:**
  - 6 blocks of 1-D causal CNNs



(a) G architecture

- **Discriminators:**
  - 5 blocks of 2-D CNNs
  - Outputs are predicted intelligibility or quality scores (range from 0 to 1)



ISSUE 4: EXPERIMENTS

- **Setup**
  - We used 1200 Harvard sentences for training; 120 for test
  - Each test sentence is mixed with 2 types of unseen noise at 3 SNRs, under 3 reverb conditions
  - Intelligibility metrics: SIIB, HASPI, ESTOI
  - Quality metrics: PESQ and ViSQOL

ISSUE 4: EXPERIMENTS

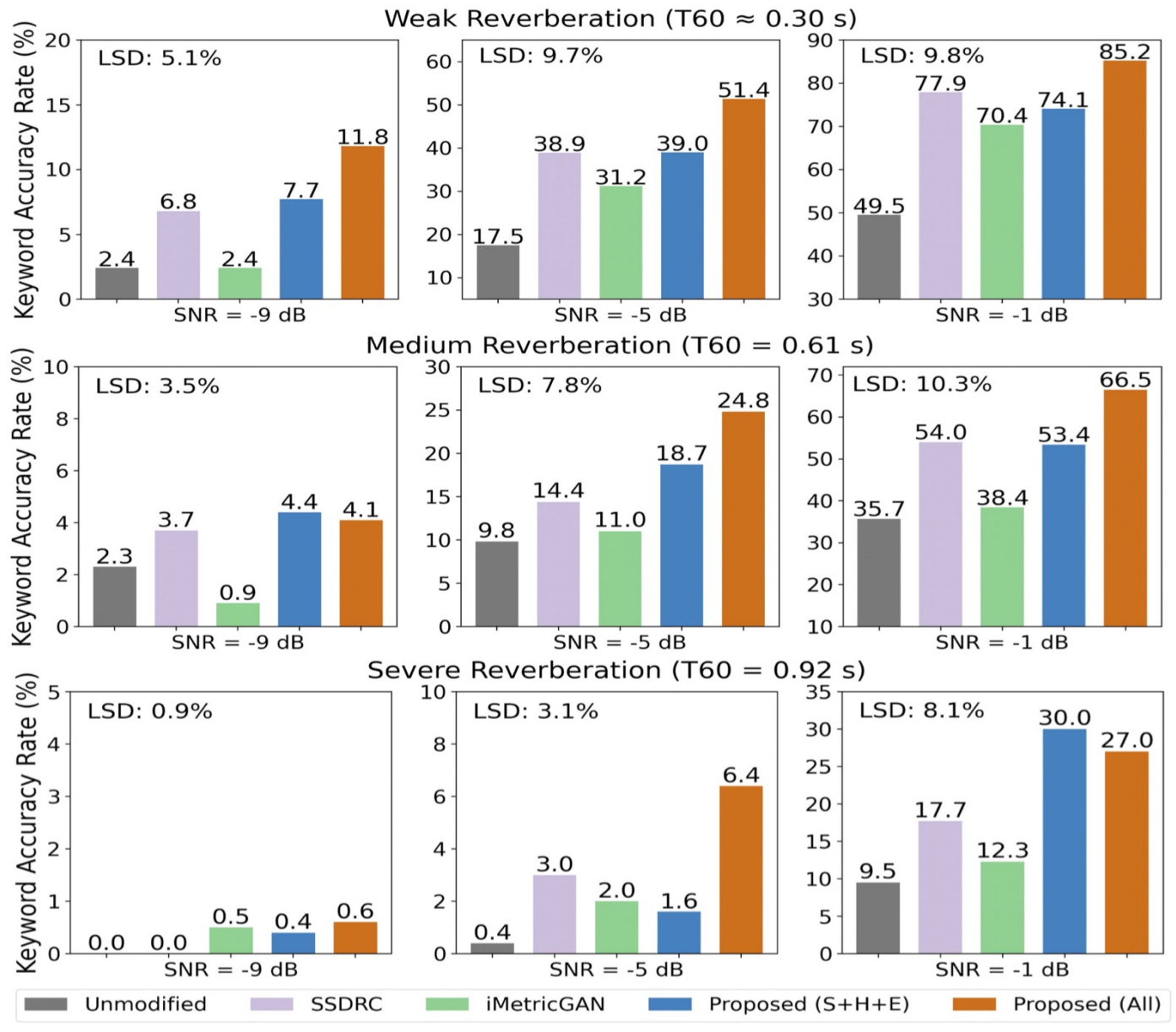
- 5 compared systems:
  - **Unmodified speech:** Plain speech without any modification
  - **SSDRC [6]:** State-of-the-art system, giving best performance as reported in previous study
  - **iMetricGAN [7]:** Our previously proposed model; Developed in 2020
  - **Proposed (S+H+E):** New developed, optimizing three intelligibility metrics
  - **Proposed (All):** New developed, jointly optimizing three intelligibility metrics and two quality metrics

[6] Zorila, Tudor-Catalin, Varvara Kandia, and Yannis Stylianou. "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression." Thirteenth Annual Conference of the International Speech Communication Association. 2012.
[7] Li, Haoyu, et al. "iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning." arXiv preprint arXiv:2004.00932 (2020).

ISSUE 4: EXPERIMENTS

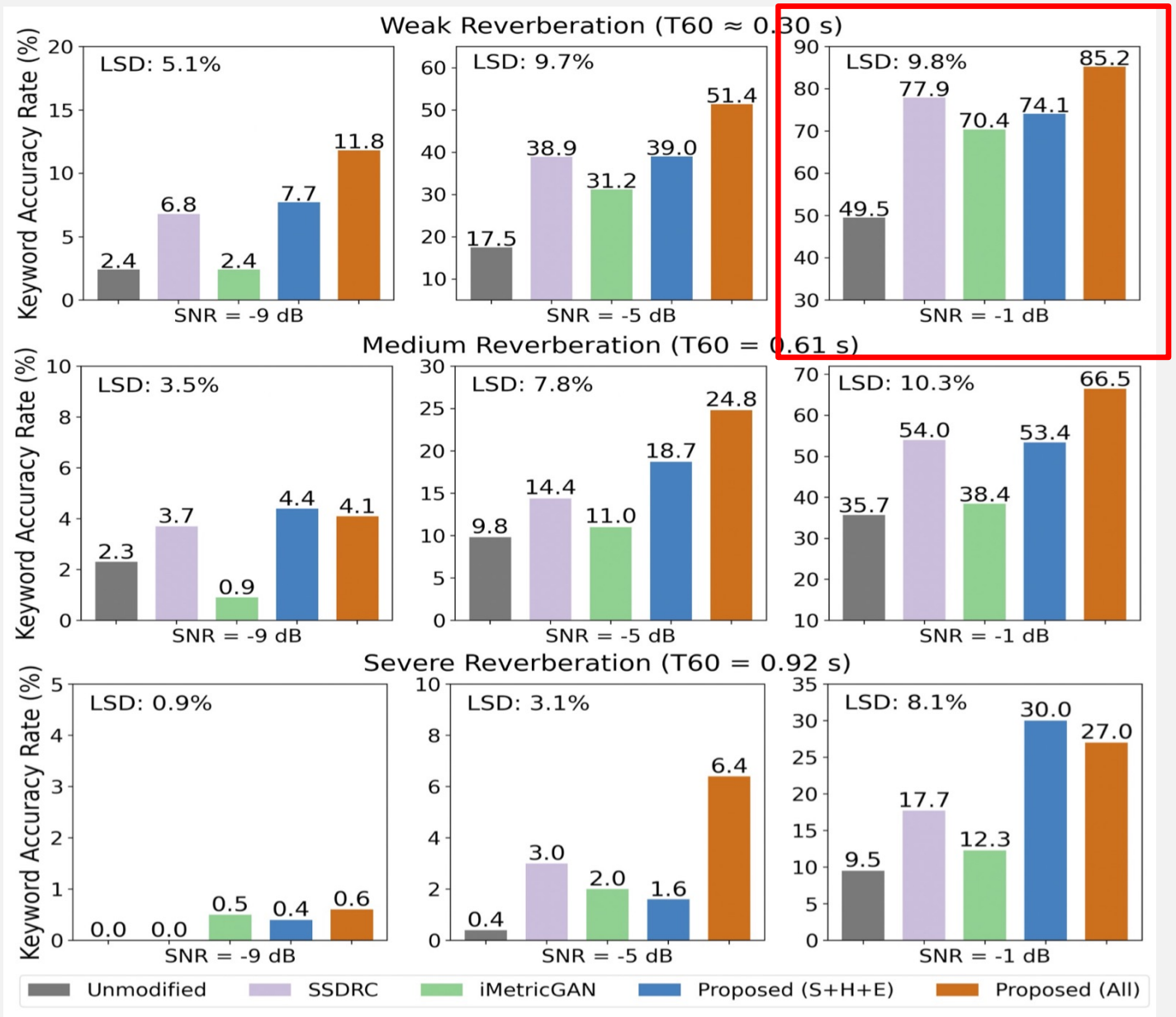
- **Subjective evaluations**
  - We conducted listening test, in which 90 native English speakers were asked to listen and type in what they heard
  - Under 2 noises, 3 reverberations and 3 SNR level conditions
  - Totally 18 different listening conditions investigated
  - Keyword Accuracy Rate as performance measure for intelligibility
    - How many correct content words (in percentage) listener can hear

ISSUE 4: EXPERIMENTS



- Under Cafeteria noise
  - SNR: the level of noise
  - T60: the level of reverberation
  - Condition with low SNR and high T60 is more challenging

ISSUE 4: EXPERIMENTS

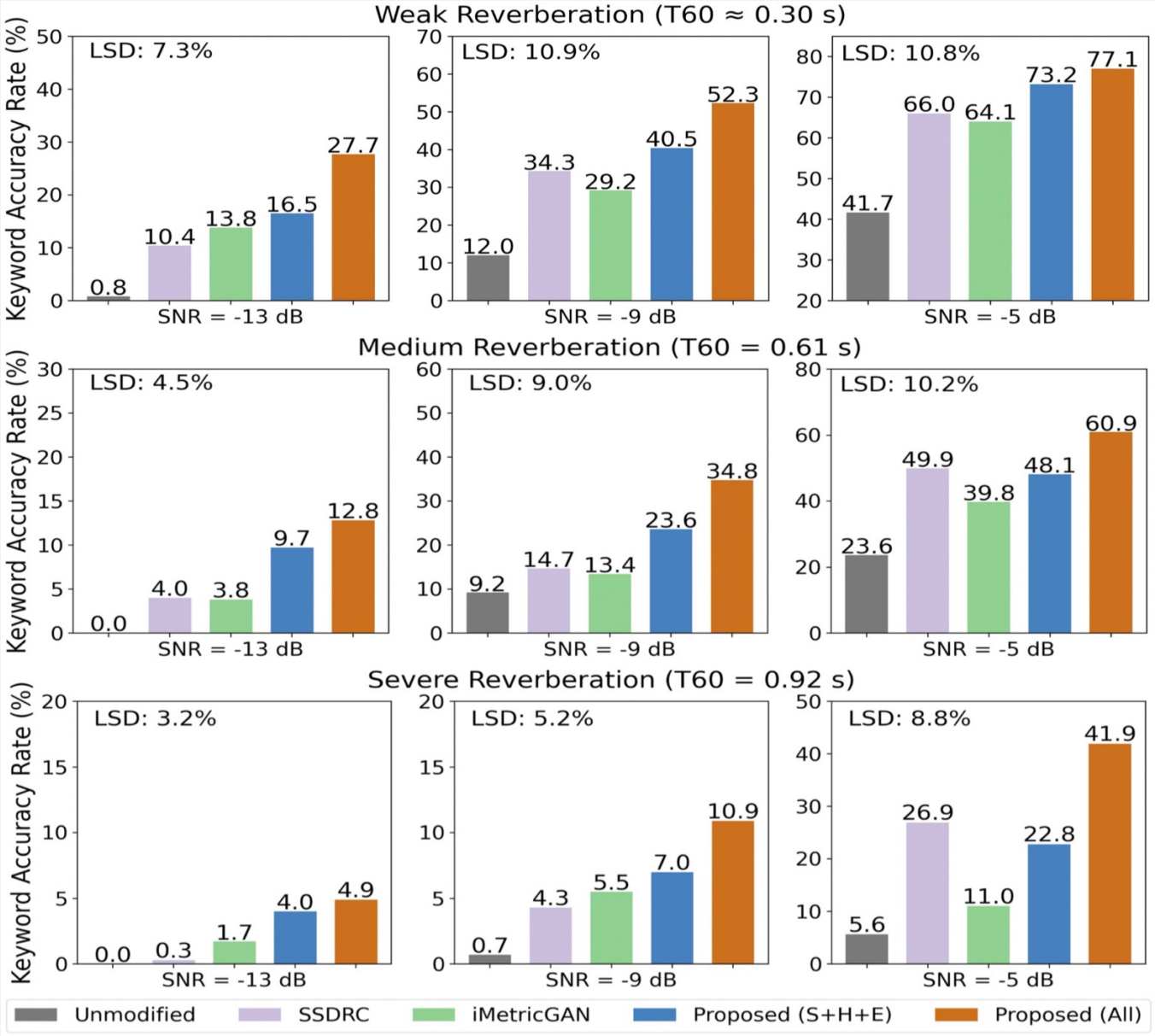


- Under Cafeteria noise
  - Proposed (All) performs best
  - Outperform SSDRC
  - Accuracy rate greatly improved: From 49.5% (unmodified) up to 85.2%

Demo: SNR=-5 dB; T60=0.30 s

- Clean speech in noise
- Modified speech in noise
- Text: "We don't like to admit our small faults"

ISSUE 4: EXPERIMENTS



- Under Airport Announcement noise

Demo: SNR=-5 dB; T60=0.92 s

- Clean speech in noise
- Modified speech in noise
- Text: "The bombs left most of the town in ruins"

ISSUE 4: EXPERIMENTS

- **Objective results:**
  - Under Cafeteria noise

System	Intelligibility in $T_{60} \approx 0.30$ s				Intelligibility in $T_{60} = 0.61$ s				Intelligibility in $T_{60} = 0.92$ s				Quality	
	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	PESQ	ViSQOL
Unmodified	15.90	1.92	0.228	6.70	15.76	1.77	0.220	6.61	9.26	1.42	0.134	5.89	4.50	5.00
SSDRC	30.98	2.74	0.314	7.03	24.72	2.27	0.273	6.77	15.24	1.83	0.199	6.04	3.52	**2.71**
iMetricGAN	35.61	2.85	0.302	7.16	26.90	2.34	0.256	6.88	16.44	1.89	0.193	6.14	3.20	2.56
S-GAN	37.89	2.77	0.239	7.31	30.57	2.35	0.208	7.04	17.91	1.79	0.154	6.20	2.08	2.02
H-GAN	35.12	**3.12**	0.242	**7.55**	27.58	2.61	0.205	7.13	16.57	1.99	0.149	6.28	2.07	2.08
E-GAN	34.20	2.71	**0.331**	7.21	28.17	2.36	**0.285**	6.94	16.03	1.81	0.207	6.15	3.07	2.38
Proposed (S+H+E)	**41.33**	3.11	0.313	7.53	**32.99**	**2.62**	0.268	**7.17**	**18.90**	**2.00**	0.194	**6.28**	2.63	2.17
Proposed (All)	37.97	2.95	0.324	7.44	31.05	2.52	0.277	7.11	18.48	1.96	**0.209**	6.26	**3.54**	2.69

ISSUE 4: EXPERIMENTS

- **Objective results**
  - Under Airport Announcement noise

System	Intelligibility in $T_{60} \approx 0.30$ s				Intelligibility in $T_{60} = 0.61$ s				Intelligibility in $T_{60} = 0.92$ s				Quality	
	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	SIIB	HASPI	ESTOI	sEPSM	PESQ	ViSQOL
Unmodified	16.25	2.20	0.191	6.63	16.12	2.07	0.190	6.61	9.43	1.58	0.115	5.79	4.50	5.00
SSDRC	32.49	3.38	0.286	7.24	25.80	2.71	0.261	6.85	16.37	2.17	0.203	6.06	3.52	**2.71**
iMetricGAN	35.68	3.44	0.280	7.37	27.72	2.73	0.250	6.95	17.98	2.23	0.204	6.18	3.22	2.58
S-GAN	42.34	3.54	0.214	7.82	34.21	2.85	0.195	7.26	21.75	2.25	0.160	6.30	2.12	2.04
H-GAN	39.19	**3.80**	0.226	7.89	31.50	3.03	0.201	7.34	20.25	**2.41**	0.165	6.37	2.08	2.10
E-GAN	35.04	3.36	0.283	7.39	28.88	2.82	**0.263**	7.03	18.09	2.23	0.205	6.17	3.07	2.40
Proposed (S+H+E)	**43.45**	3.75	0.279	**7.94**	**35.31**	**3.04**	0.250	**7.36**	**22.36**	2.40	0.206	**6.37**	2.71	2.19
Proposed (All)	42.54	3.72	**0.288**	7.87	34.30	3.00	0.257	7.30	22.03	2.38	**0.209**	6.36	**3.56**	2.67

  - Large intelligibility improvements in all tested noise and reverberations
  - Proposed (S+H+E) and Proposed (All) significantly outperform SSDRC baseline

ISSUE 4: SUMMARY

- The proposed system greatly improves the intelligibility of speech under various listening conditions
- Intelligibility boosting is achieved by optimizing multiple intelligibility and quality metrics through GAN model
- GAN helps avoid intractable optimization problem
- Our system outperforms the state-of-the-art methods

- **More demos**
  - https://nii-yamagishilab.github.io/hyli666-demos/intelligibility/index.html

CONTENTS

ISSUE 5: MOTIVATION

- **Background**
  - Noise exists in both far-end speaker and near-end listener environments
  - Noise reduction (NR) and intelligibility boosting (IB) should be accordingly carried out

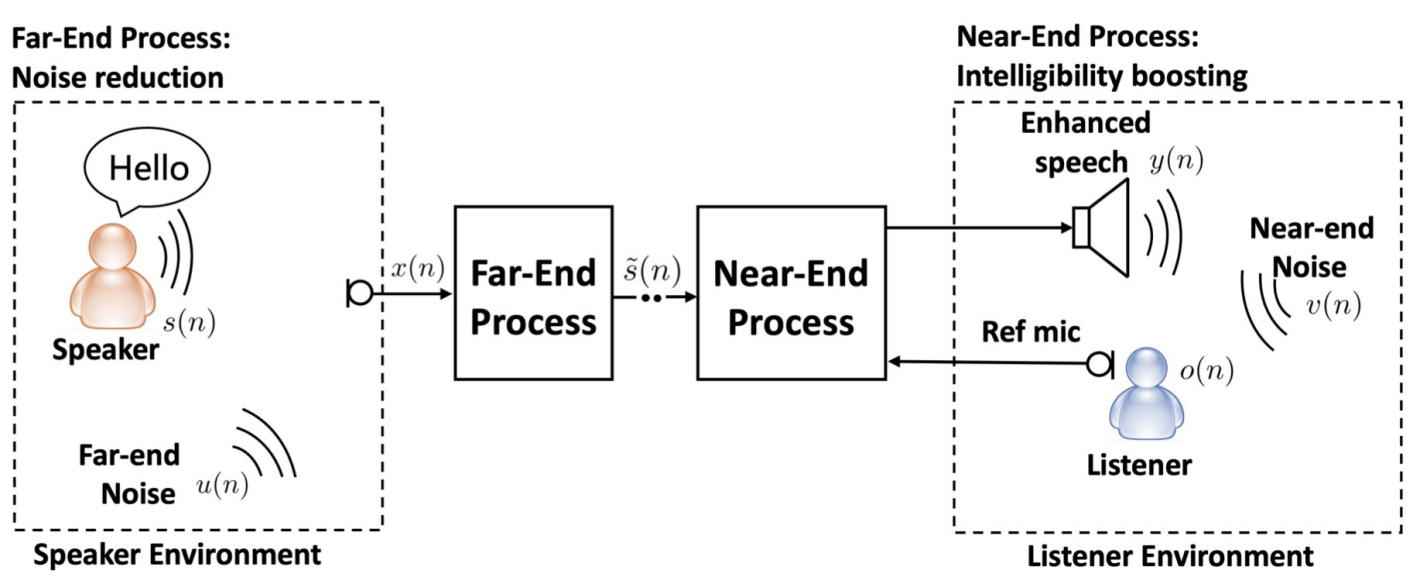
- **Solution**
  - Disjoint pipeline processing: sequentially apply NR and IB?
  - Jointly optimize NR and IB within a unified framework?

ISSUE 5: PROBLEM FORMULATION

- Signal model

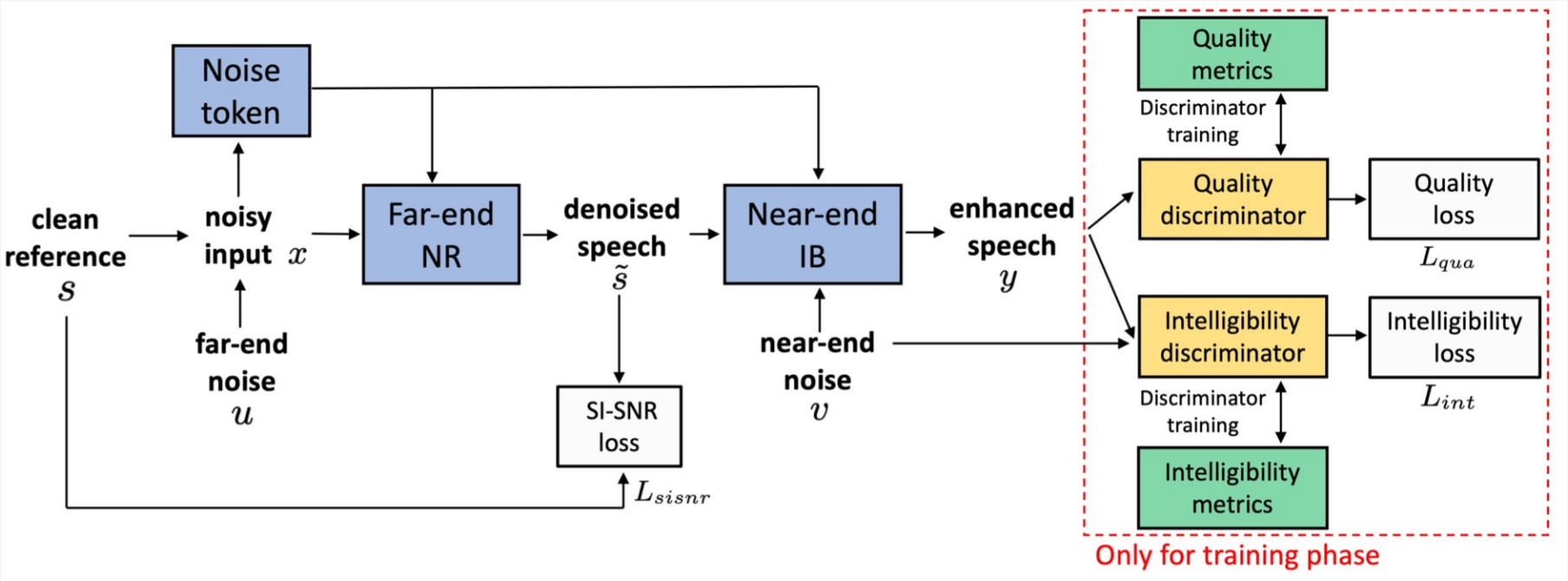
$$x = s + u, \quad \tilde{s} = NR(x), \quad y = IB(\tilde{s}|v), \quad o = y + v,$$

- Goal
  - Improve the listening experience for listeners, i.e., the intelligibility of *o(n)* and quality of *y(n)*, by designing effective NR and IB modules.



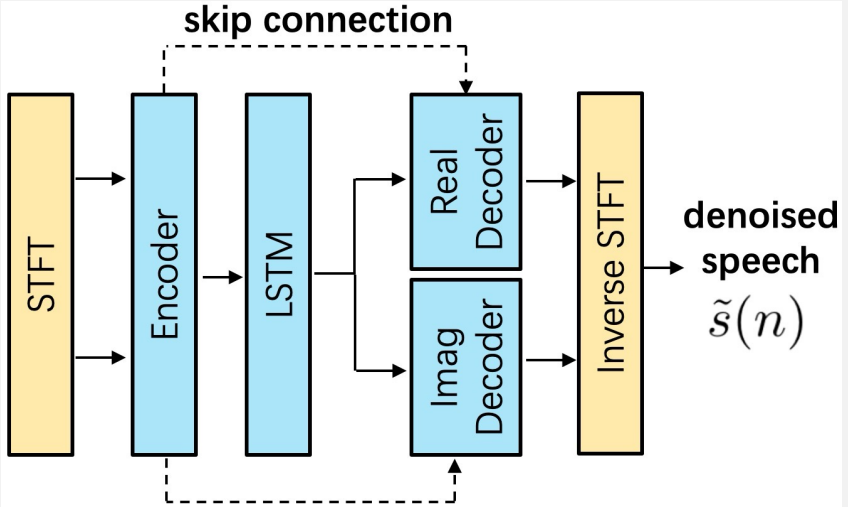
ISSUE 5: METHOD

- **Three main modules**
  - 1. NR module: suppress noise
  - 2. IB module: further improve the intelligibility of denoised speech
  - 3. Noise token module: extract noise embedding and informs other modules of far-end environmental information



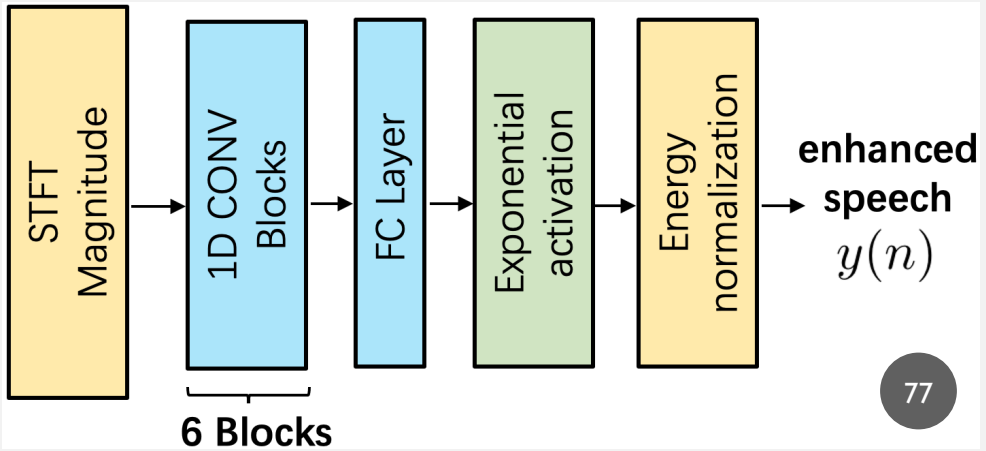
ISSUE 5: METHOD

- **Noise reduction module**
  - Adopt Convolutional recurrent network (CRN)
  - CNN-Encoder – LSTM – CNN-Decoder
  - Input: Real and Imag spectrogram (2 channels) of noisy speech
  - Real and Imag Decoders output real and imag spectrogram, respectively



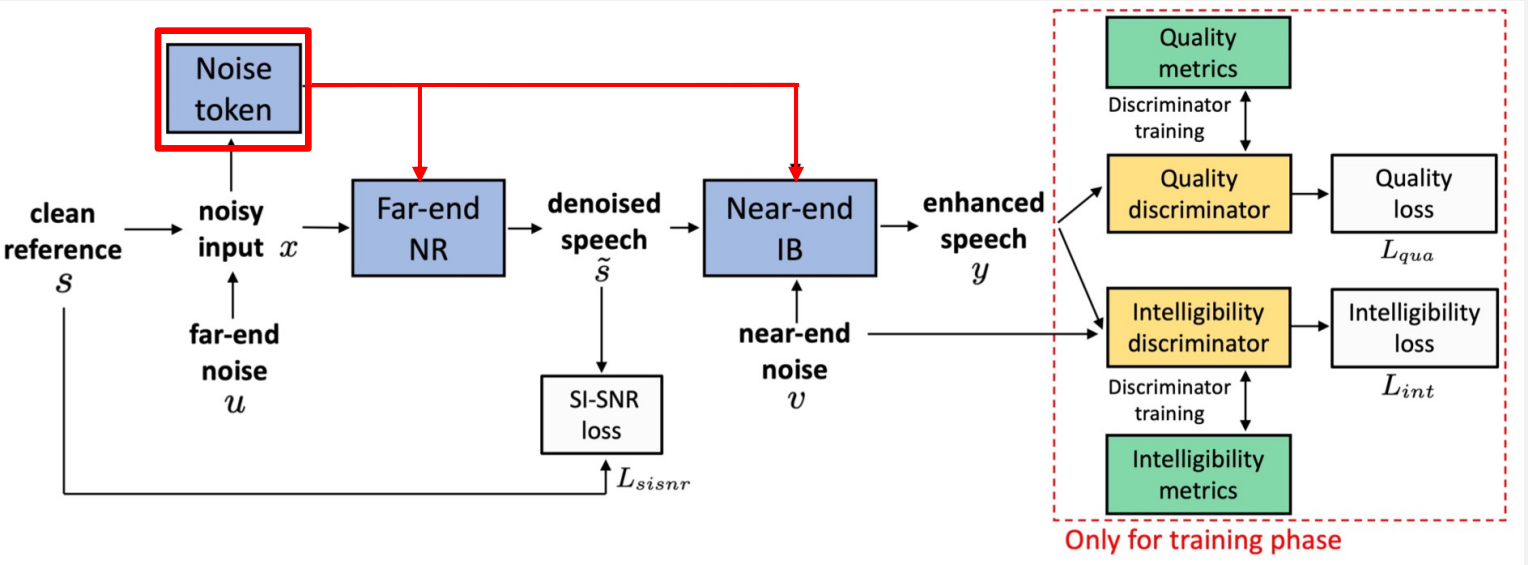
ISSUE 5: METHOD

- **Intelligibility boosting module**
  - GAN model; Similar to that in Chapter 5
  - Consists of 1D CNN blocks
  - Input: Spectrogram magnitude; Output: Amplification factors redistributing speech energy
  - Optimize 6 metrics:
    - SIIB, HASPI, and ESTOI as intelligibility metrics;
    - PESQ, ViSQOL, and HASQI as quality metrics
  - Discriminators are composed of 2D CNN



ISSUE 5: METHOD

- **Noise token module**
  - Encode the far-end environment information and generate noise embedding
  - We believe both NR and IB modules can benefit from additional noise information
  - For example, by exploiting far-end noise knowledge, IB module may avoid to amplify the time-frequency regions containing much residual noise.
  - Network design is similar to our previous "noise tokens" work



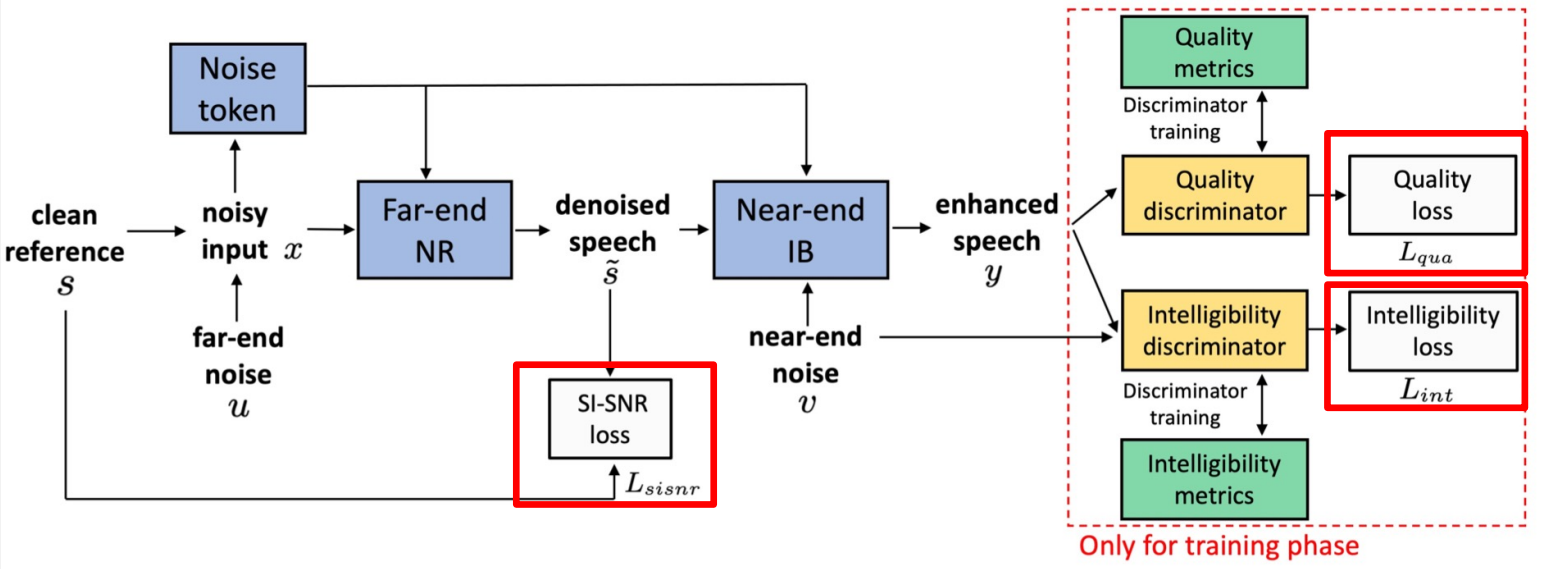
ISSUE 5: METHOD

- **Training objective**

$$L = L_{int} + \alpha * L_{qua} + \beta * L_{sisnr},$$

Note: t_{int} is the maximum scores of the selected intelligibility metrics

- 1. Intelligibility loss: calculated by intelligibility discriminator $L_{int} = ||D_{int}(y) - t_{int}||^2$
- 2. Quality loss: by quality discriminator
- 3. Speech denoising loss: Scale-Invariant SNR (SI-SNR) calculated by comparing denoised speech with clean reference



ISSUE 5: EXPERIMENTS

- **Setup**
  - We used 1200 Harvard sentences for training; 120 for test
  - Far-end noise type in test set is Cafeteria noise at {6, 10, 14} dB
  - Near-end noise type in test set is Airport Announcement noise at {-9, -5, -1} dB

  - Intelligibility metrics: SIIB, HASPI, ESTOI
  - Quality metrics: PESQ, ViSQOL, and HASQI

ISSUE 5: EXPERIMENTS

- 7 compared systems:
  - **Noisy:** The far-end noisy speech is played under the near end noise without any modification
  - **Noisy+NR:** Processed by only NR module
  - **Noisy+IB:** Processed by only IB module

ISSUE 5: EXPERIMENTS

- **7 compared systems:**
  - **Noisy:** The far-end noisy speech is played under the near end noise without any modification
  - **Noisy+NR:** Processed by only NR module
  - **Noisy+IB:** Processed by only IB module
  - **DSPPipeline:** Processed by the signal-processing-based disjoint pipeline, consisting of Wiener filter (for NR) and SSDRC (for IB)
  - **NeuralPipeline:** Processed by the neural-network-based disjoint pipeline, consisting of CRN-based NR and GAN-based IB

ISSUE 5: EXPERIMENTS

- **7 compared systems:**
  - **Noisy:** The far-end noisy speech is played under the near end noise without any modification
  - **Noisy+NR:** Processed by only NR module
  - **Noisy+IB:** Processed by only IB module
  - **DSPPipeline:** Processed by the signal-processing-based disjoint pipeline, consisting of Wiener filter (for NR) and SSDRC (for IB)
  - **NeuralPipeline:** Processed by the neural-network-based disjoint pipeline, consisting of CRN-based NR and GAN-based IB
  - **Joint:** Processed by the partial joint model (without noise token); NR and IB are jointly optimized
  - **Joint+NT:** Processed by the full joint model (with noise token); NR and IB are jointly optimized

ISSUE 5: EXPERIMENTS

- Intelligibility scores
  - Applying only NR or IB does not increase the intelligibility
  - Joint methods consistently outperform the pipeline
  - Benefiting from noise token, Joint+NT achieves the overall best performance

System	Far-end SNR = 6 dB			Far-end SNR = 10 dB			Far-end SNR = 14 dB		
	SIIB	HASPI	ESTOI	SIIB	HASPI	ESTOI	SIIB	HASPI	ESTOI
Noisy	17.98	2.20	0.221	19.72	2.31	0.237	21.07	2.41	0.249
Noisy+NR	19.52	2.24	0.250	20.73	2.32	0.259	21.65	2.39	0.266
Noisy+IB	15.79	2.09	0.180	18.76	2.28	0.206	21.91	2.47	0.232
DSPPipeline	15.58	1.96	0.208	18.22	2.10	0.229	21.06	2.24	0.251
NeuralPipeline	24.47	2.67	0.302	27.34	2.85	0.319	30.09	**3.00**	0.333
Joint	26.16	2.70	0.305	28.65	2.84	0.319	30.77	2.96	0.330
Joint+NT	**28.48**	**2.73**	**0.320**	**31.45**	**2.87**	**0.334**	**33.79**	2.99	**0.344**

ISSUE 5: EXPERIMENTS

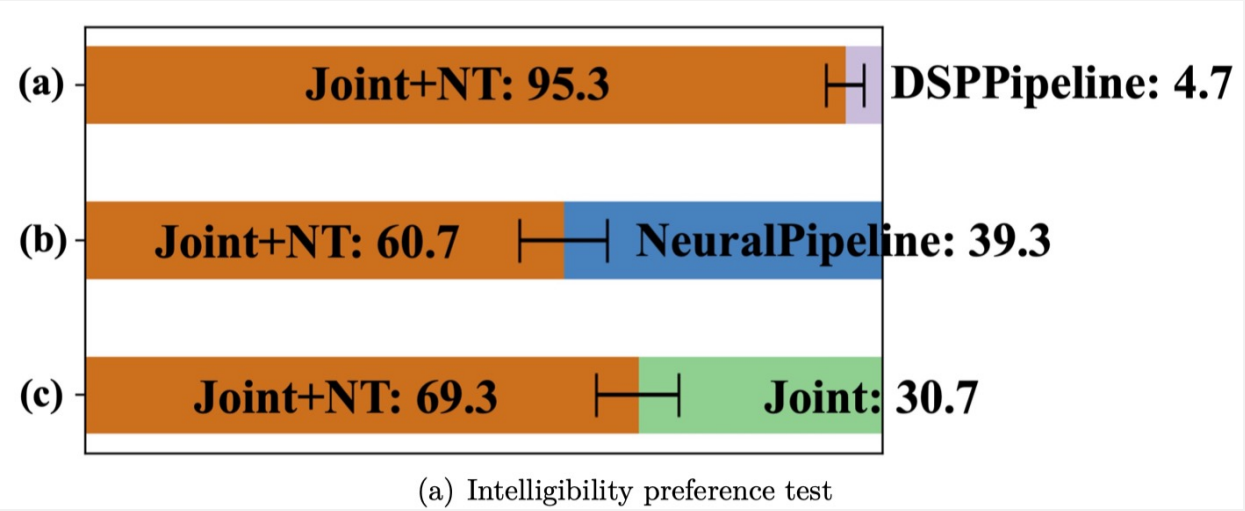
- Quality scores
  - Intelligibility-enhancing modifications degrade the speech quality at the cost of increasing intelligibility
  - Joint methods preserve speech quality well compared to pipeline

System	Far-end SNR = 6 dB			Far-end SNR = 10 dB			Far-end SNR = 14 dB		
	PESQ	HASQI	ViSQOL	PESQ	HASQI	ViSQOL	PESQ	HASQI	ViSQOL
Noisy	1.41	0.15	1.83	1.55	0.18	1.94	1.69	0.21	2.09
Noisy+NR	**2.33**	0.28	**2.48**	**2.52**	0.32	**2.69**	**2.70**	**0.36**	**2.91**
Noisy+IB	1.24	0.10	1.66	1.32	0.12	1.71	1.41	0.14	1.78
DSPPipeline	1.32	0.10	1.68	1.43	0.12	1.74	1.54	0.14	1.81
NeuralPipeline	2.01	0.23	2.14	2.19	0.26	2.25	2.35	0.28	2.35
Joint	2.14	0.28	2.20	2.30	0.30	2.32	2.43	0.33	2.43
Joint+NT	2.26	**0.30**	2.32	2.45	**0.32**	2.43	2.58	0.35	2.52

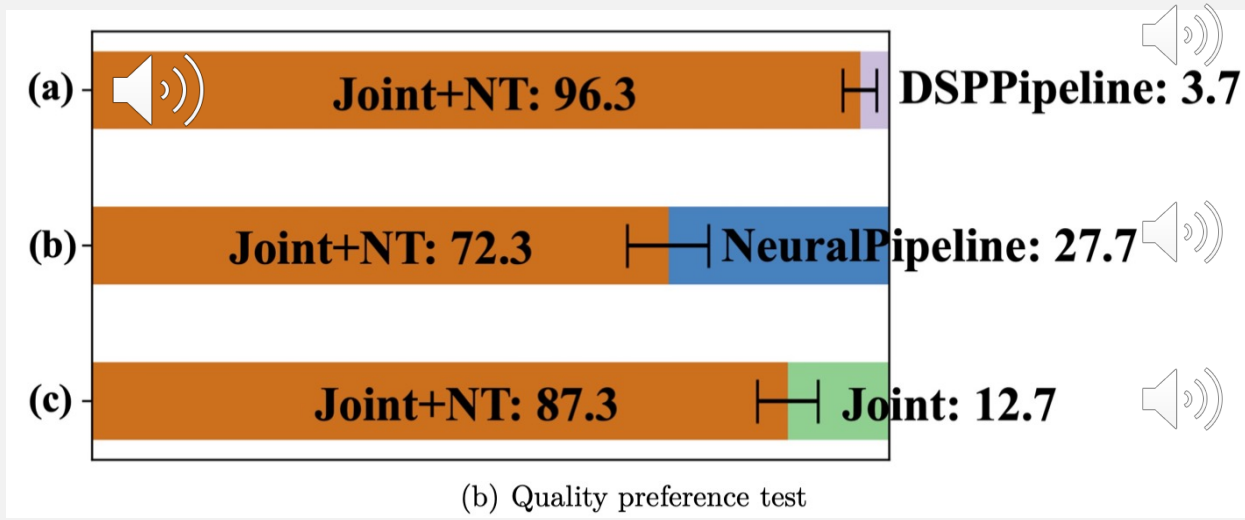
ISSUE 5: EXPERIMENTS

- Subjective preference test
  - 20 native speakers were recruited
  - Listen to a pair of samples and select the one sounds more clear or better
  - Still, proposed Joint+NT method performs best

Plain noisy:



(a) Intelligibility preference test



(b) Quality preference test

ISSUE 5: SUMMARY

- We proposed a joint framework integrating noise reduction with intelligibility boosting to address the full-end SE task
- Under this joint framework, these two modules can be jointly optimized
- It achieves significant intelligibility gain while preserving speech quality well

CONTENTS

CONCLUSION

- **We improved noise reduction model:**
  - Improving its noise generalization
  - Using WaveRNN vocoder to synthesize waveform to alleviate phase distortion
  - Considering and enhancing general device degradation
- **We proposed a novel neural intelligibility boosting model:**
  - Introducing GAN into intelligibility boosting model
  - It outperforms state-of-the-art baseline
- **We integrated noise reduction with intelligibility boosting:**
  - Addressing full-end speech enhancement task where noises exist in both speaker and listener sides
  - It outperforms disjoint pipeline methods

FUTURE DIRECTION

- In Chapter 3, try non-autoregressive vocoder such as HiFi-GAN
- In Chapter 4, try mutual information (MI) minimization to filter out channel characteristic
- In Chapter 5, try new reverberation modeling method
- In Chapter 6, try end-to-end method that directly maps noisy speech into clean intelligibility-boosted speech

- Extend single-channel approach to the use of microphone array

CONTENTS

PUBLICATIONS

- First-authored
  - H. Li, Y. Liu and J. Yamagishi, "Joint Noise Reduction and Listening Enhancement for Full-End Speech Enhancement," Submitted to SLT 2022. (Chapter 6)
  - H. Li and J. Yamagishi, "DDS: A new device-degraded speech dataset for speech enhancement," Accepted to Interspeech 2022. (Appendix 1)
  - H. Li and J. Yamagishi, "Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3000-3011, 2021, doi: 10.1109/TASLP.2021.3111566. (Chapter 5)
  - H. Li, S. Fu, Y. Tsao and J. Yamagishi, 2020 ``iMetricGAN: Intelligibility Enhancement for Speech-in-Noise Using Generative Adversarial Network-Based Metric Learning'', Proc. Interspeech 2020, 1336-1340. (Chapter 5)
  - H. Li, Y. Ai and J. Yamagishi, ``Enhancing Low-Quality Voice Recordings Using Disentangled Channel Factor and Neural Waveform Model," 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 2021, pp. 734-741. (Chapter 4)
  - H. Li and J. Yamagishi, 2020 ``Noise Tokens: Learning Neural Noise Templates for Environment-Aware Speech Enhancement'', Proc. Interspeech 2020, 2452-2456. (Chapter 3)

PUBLICATIONS

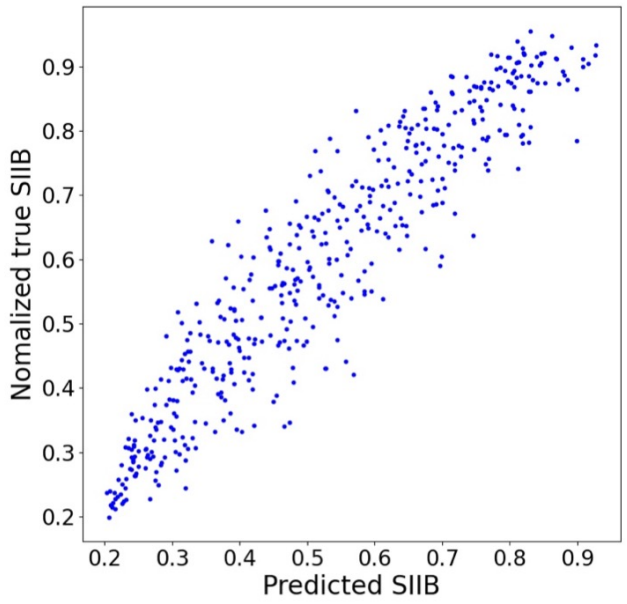
- Co-authored
  - Y. Ai, H. Li, X. Wang, J. Yamagishi and Z. Ling, ``Denoising-and-Dereverberation Hierarchical Neural Vocoder for Robust Waveform Generation'' 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 2021, pp. 477-484.
  - Y. Zhao, H. Li, C. Lai, J. Williams, E. Cooper and J. Yamagishi, 2020 ``Improved Prosody from Learned F0 Codebook Representations for VQ-VAE Speech Waveform Reconstruction,'' Proc. Interspeech 2020, 4417-4421.

Thanks for attention

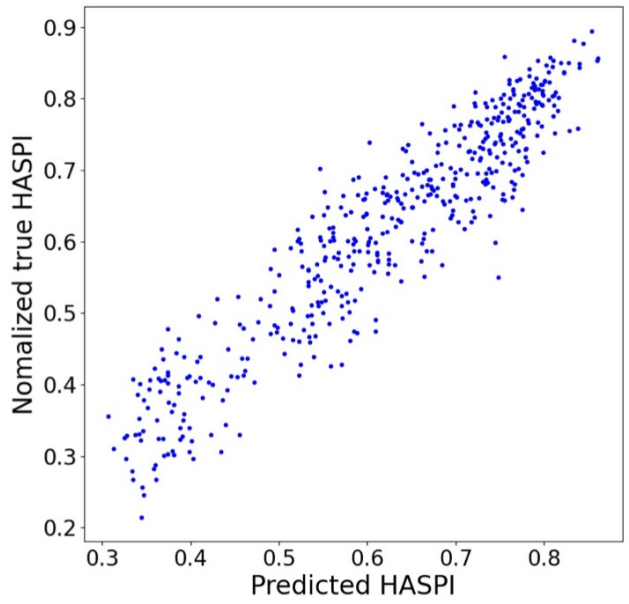
Appendix

CORRELATIONS OF DISCRIMINATOR

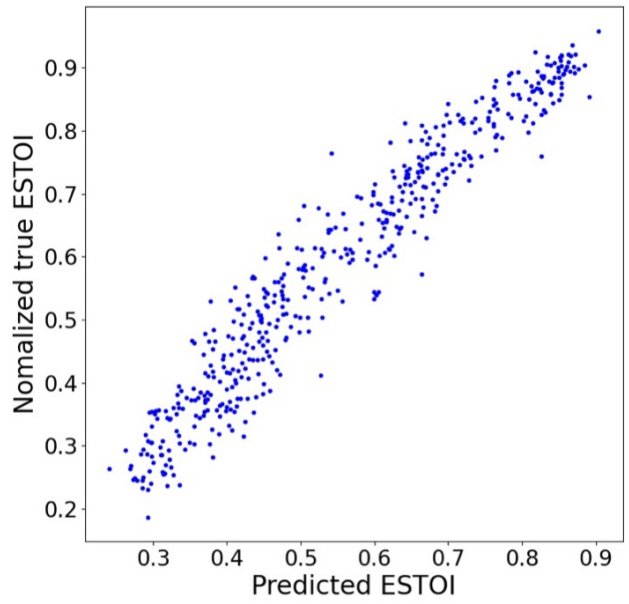
- Correlations between the predicted metric scores and true labels in Chapter 5



(a) SIIB, $r = 0.94$



(b) HASPI , $r = 0.93$



(c) ESTOI, $r = 0.97$