Enhancing Low-Quality Voice Recordings Using Disentangled Channel Factor and Neural Waveform Model

Haoyu Li[1], Yang Ai[2], and Junichi Yamagishi[1]
[1]National Institute of Informatics, Japan
[2]University of Science and Technology of China, P.R.China

IEEE SLT 2021

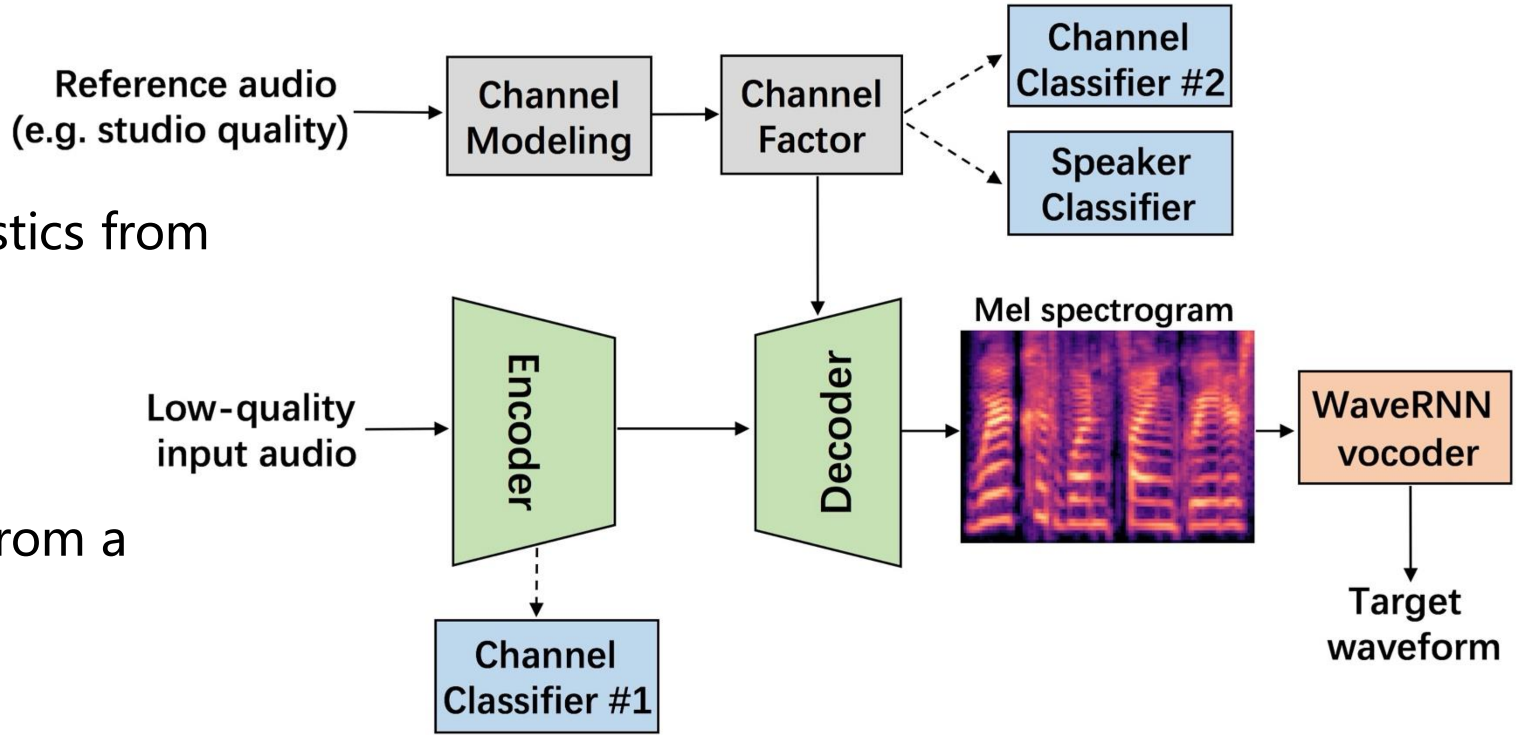
Goal of this paper

- Transform low-quality speech into high-quality ones (Speech Enhancement)
  - Low-quality recording 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 speech enhancement task
  - Regard SE as a style transfer task, from low quality style to high quality
  - Apply neural waveform model to synthesize speech, instead of using ISTFT

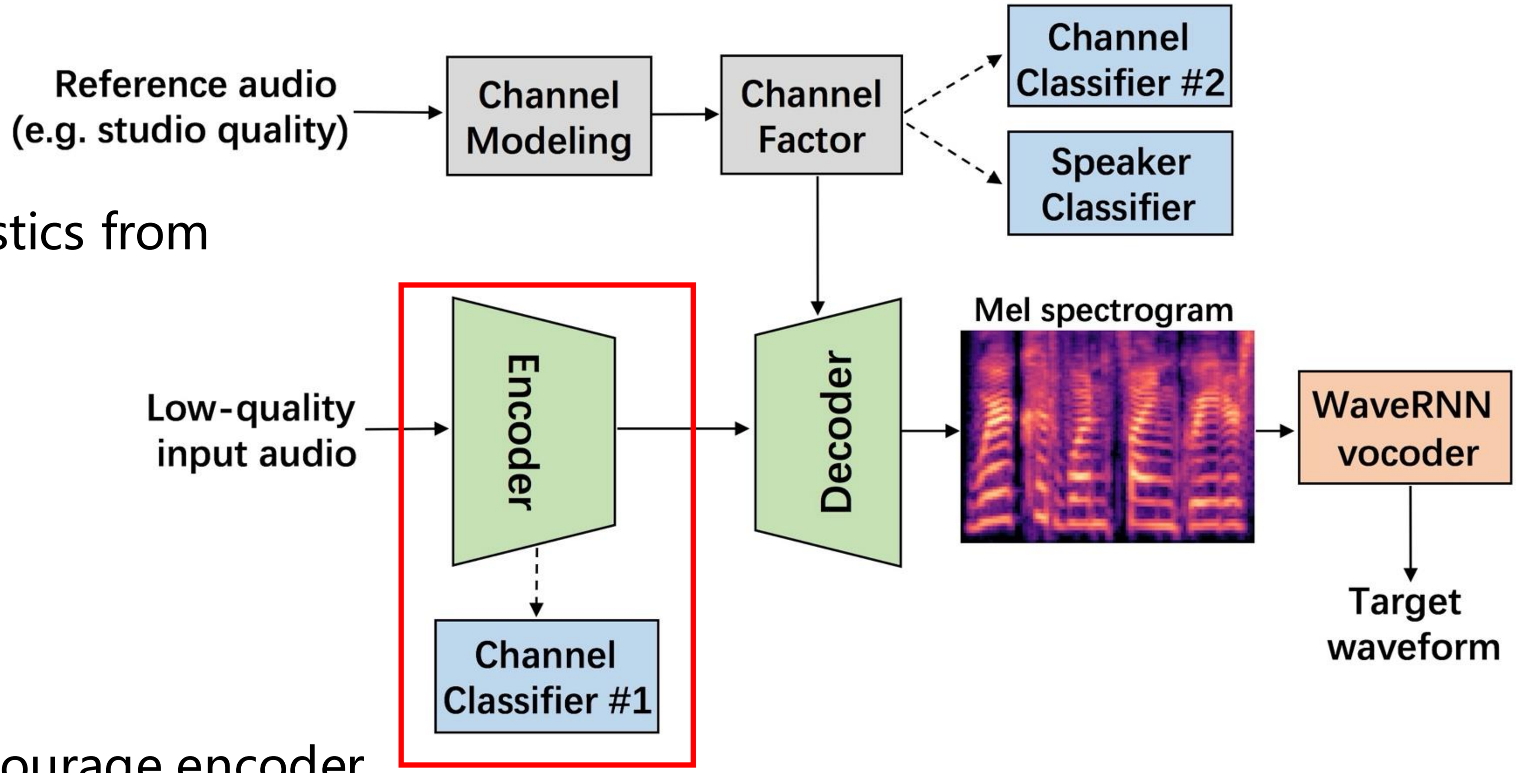
Overview of system diagram

- Encoder
  - Filter out the channel characteristics from the original 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)



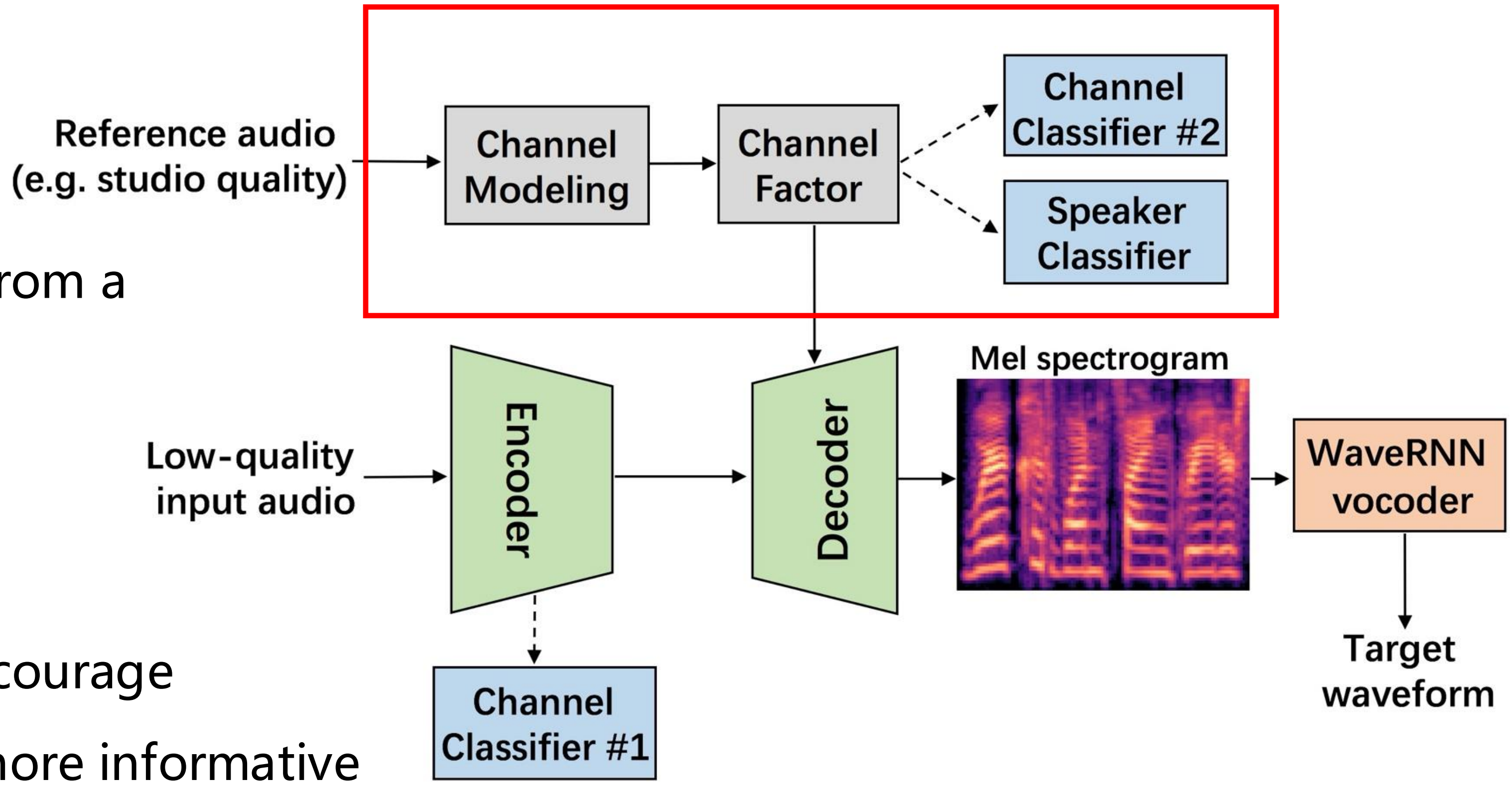
Component details

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



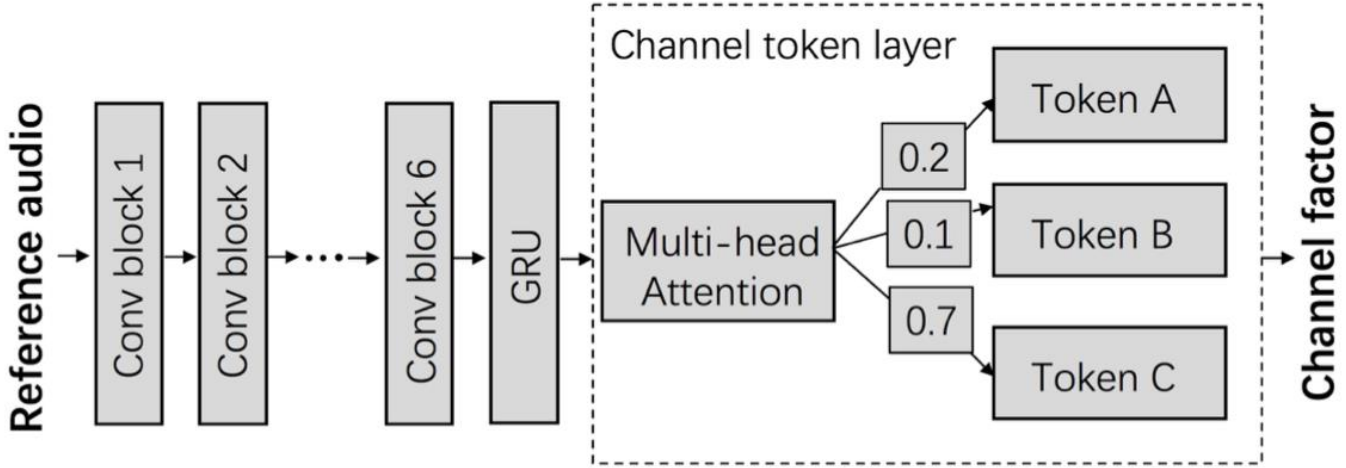
Component details

- 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

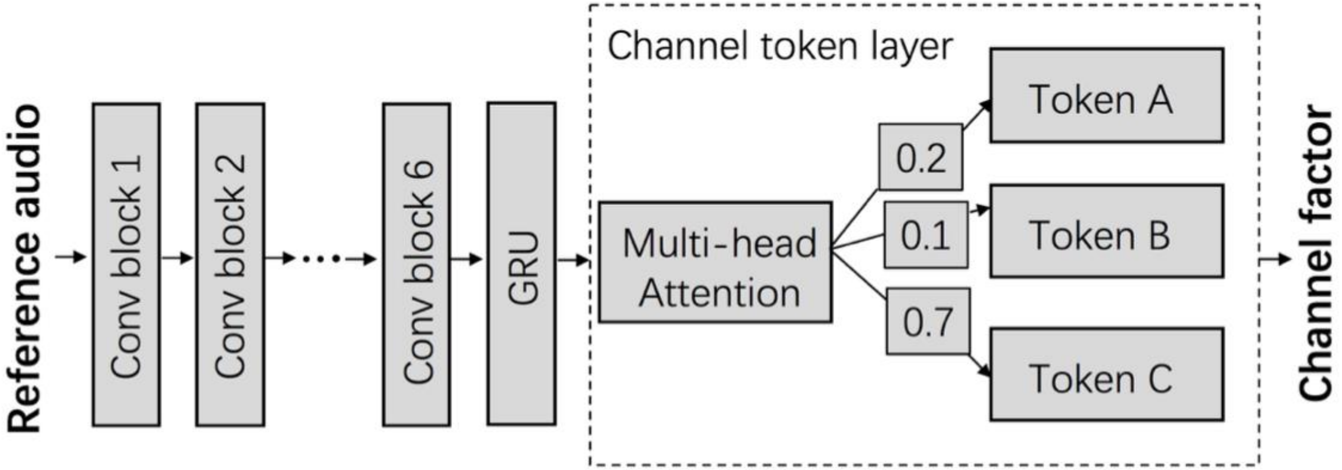


Component details

- Channel modeling
  - Shares a similar network structure with "Global Style Tokens"
  - Design an interpretable and controllable channel modeling module. (e.g., Token A might represent reverb level, Token B represents noise level, etc.)



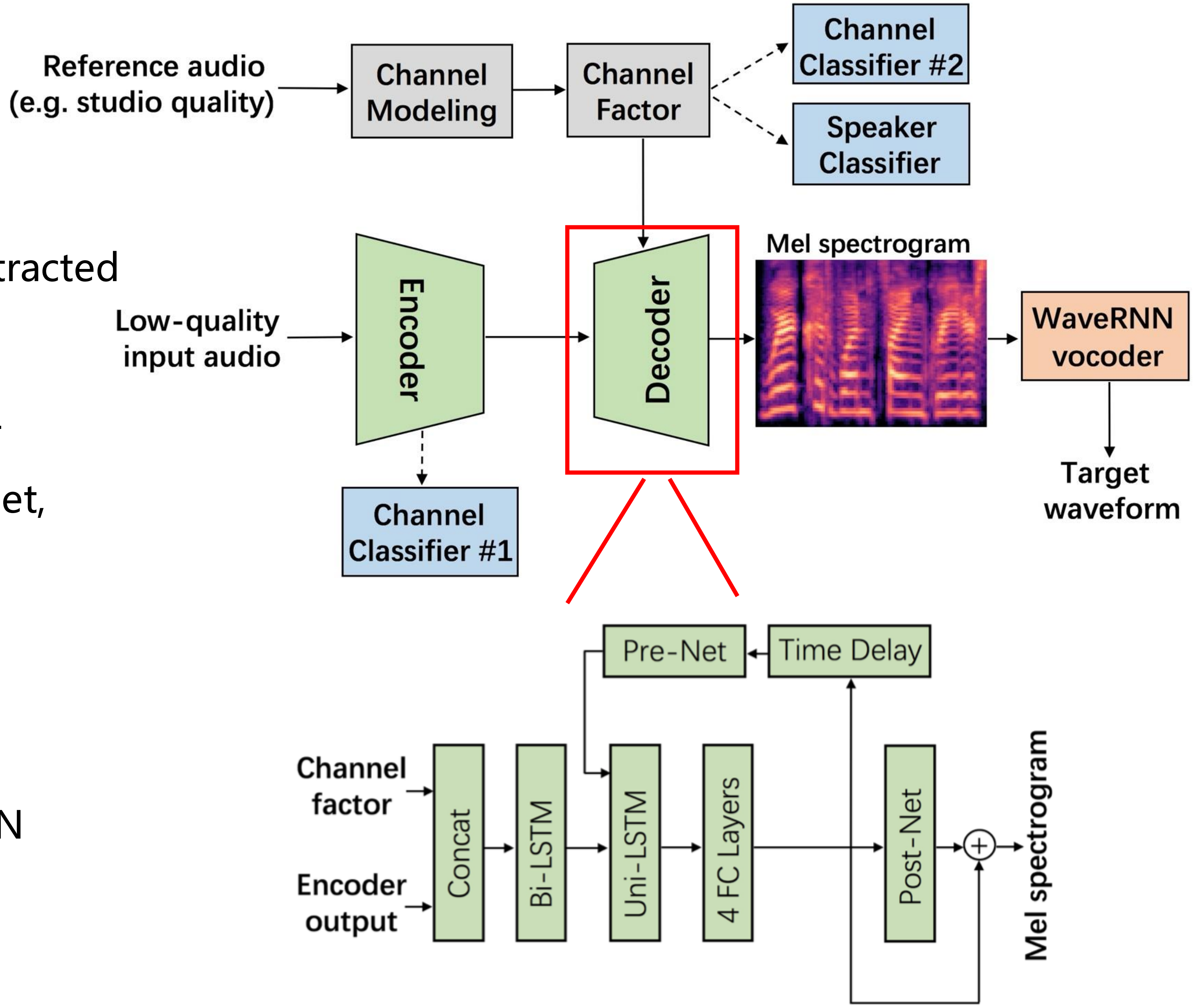
Component details



- Channel modeling
  - Shares a similar network structure with "Global Style Tokens"
  - Design an interpretable and controllable channel modeling module. (e.g., Token A might represent reverb level, Token B represents noise level, etc.)
- 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

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

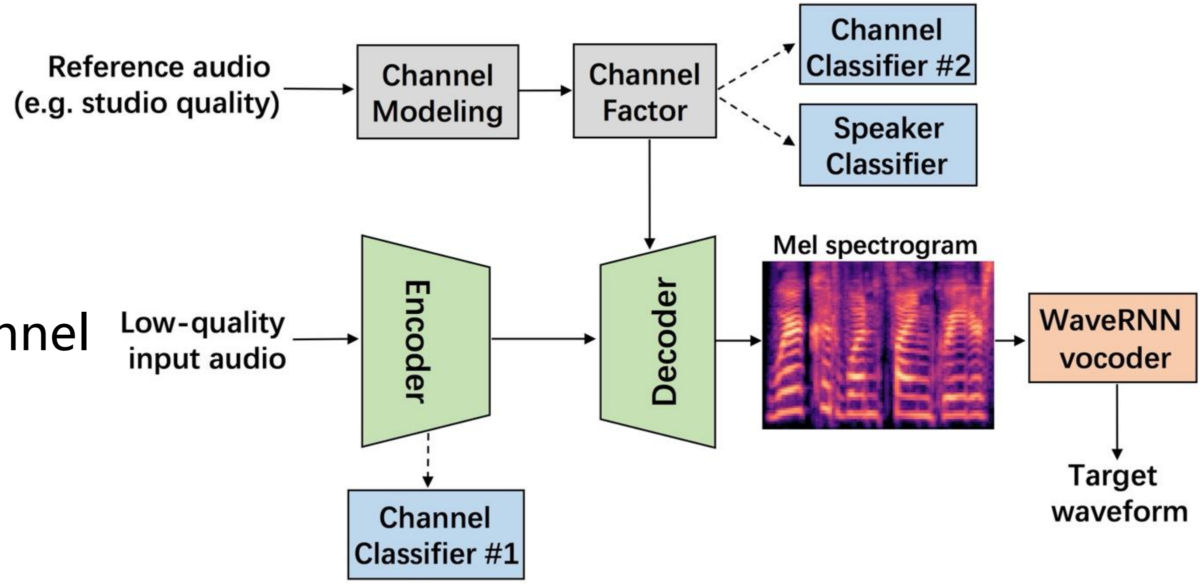


Experiments

- 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

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



Experiments

- Other compared methods
  - **Raw audio**: lower bound
  - **Studio audio**: higher bound
  - **WPE**: signal-processing method for speech dereverberation
  - **WPE+LogMMSE**: signal-processing method for speech dereverberation + 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

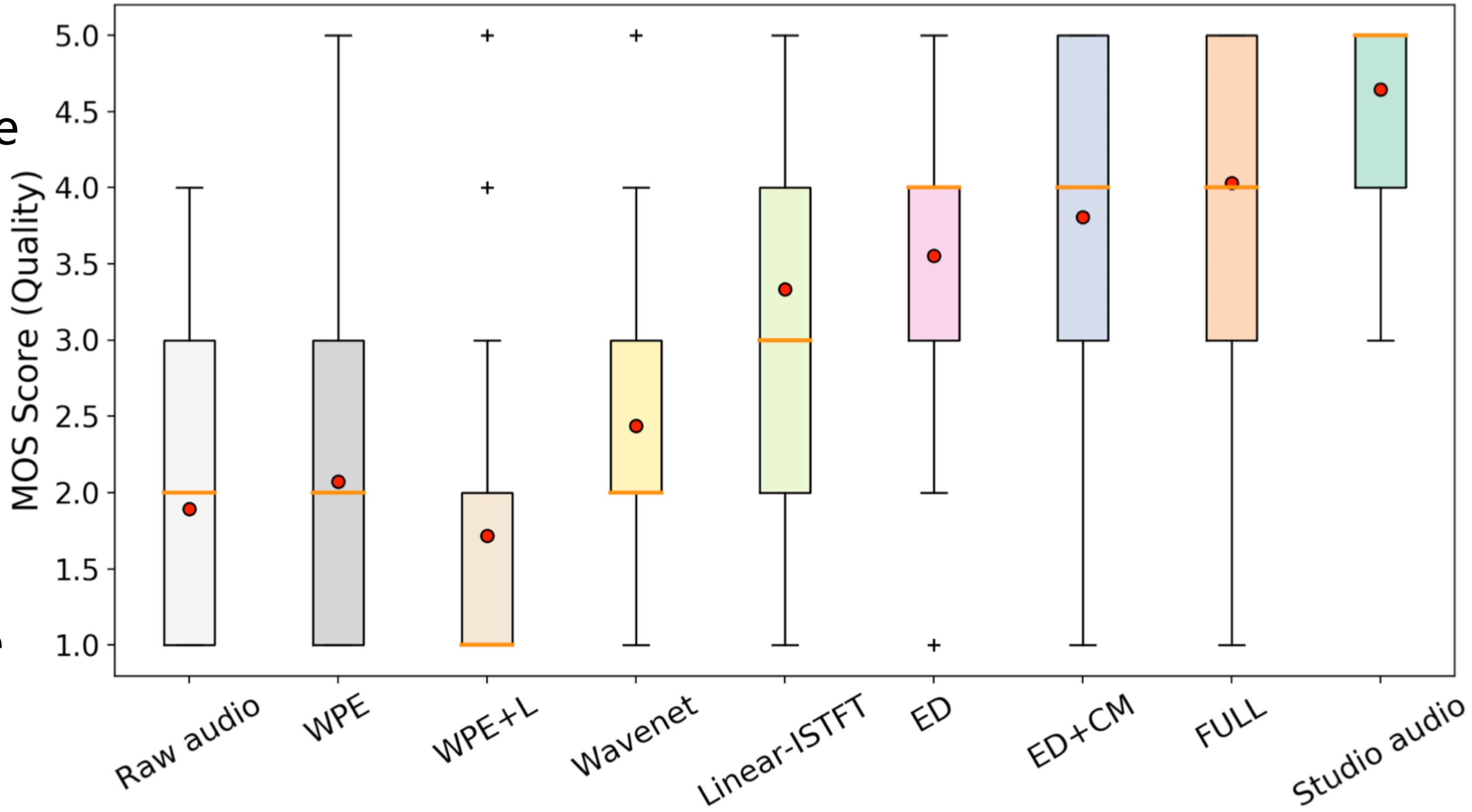
Objective results

- **FULL** consistently improves its two simplified versions, **ED** and **ED+CM**, and other compared methods (**WPE**, **WPE+L**, and **WaveNet**)
- **FULL** system 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

Subjective results

- Conducted crowdsourced listening tests, 165 individuals rated quality for given samples with 5-point MOS score



- **FULL** gives 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/

Beyond enhancement: Audio effect transfer

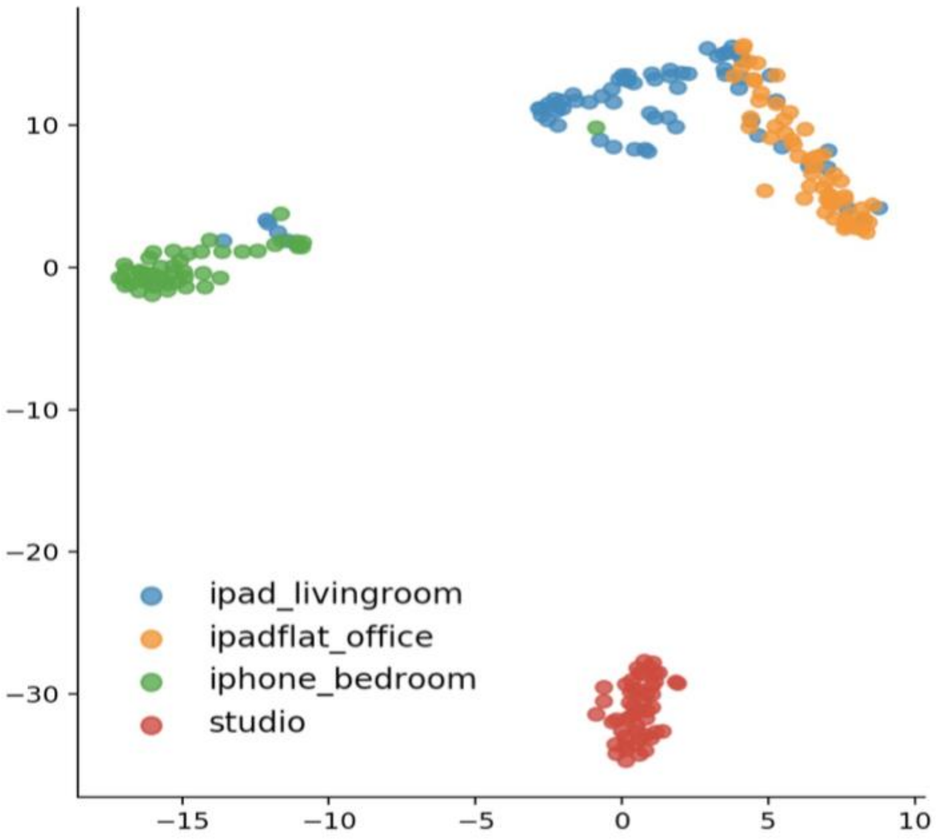
- Speech enhancement: Transfer low-quality to high-quality style

- Can we transfer speech into arbitrary style by designating a corresponding reference audio?

Visualization of learned channel factors

- Channel Modeling module extracts channel factors from 3 **unseen** recording (channel) conditions

- Can discriminate unseen reference audios and produce representative factors



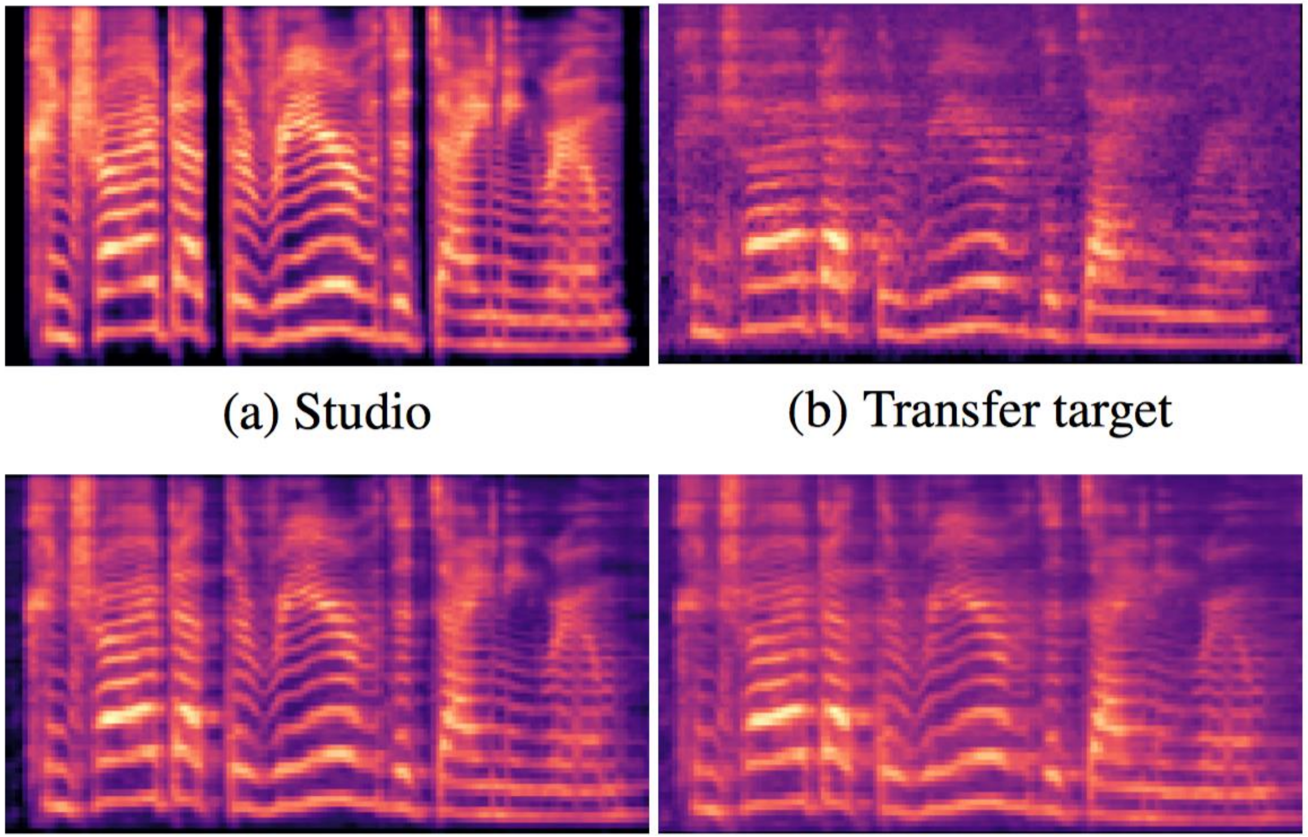
An example of flexible control on transferred style

- Control transferred effect from less reverberant to more reverberant by linear interpolation of two pre-computed channel factors:

$$\hat{\boldsymbol{z}}_c = (1-\alpha) * \boldsymbol{z}_c^{pro} + \alpha * \boldsymbol{z}_c^{iph}$$

- $\boldsymbol{z}_c^{pro}$ and $\boldsymbol{z}_c^{iph}$ denote the channel factors extracted from a professional studio recording and iphone bedroom recording

- α is the scale value that ranges from 0 to 1



(a) Studio (b) Transfer target

(c) Transferred at $\alpha = 0.6$ (d) Transferred at $\alpha = 1.0$

Conclusion

- Apply style transfer approach into speech enhancement task, in which we jointly denoising, dereverberation, and applying pleasing audio effect to low-quality recordings
  - System outperforms one time-domain model (Denoising-WaveNet) and several signal-processing baselines.
  - **Mel+WaveRNN** waveform synthesis module outperforms **Linear+ISTFT** in subjective evaluations
- However…
  - Still require expensive parallel recordings for training -> Expanded to non-parallel style transfer?
  - Although we can transfer any channel characteristics within this framework, but in practice people most commonly want clean channel characteristics only.

Thanks!