Paper ID: 5347

Direct Noisy Speech Modeling for Noisy-to-Noisy Voice Conversion

Chao Xie, Yi-Chiao Wu, Patrick Lumban Tobing, Wen-Chin Huang and Tomoki Toda

Nagoya University, Japan

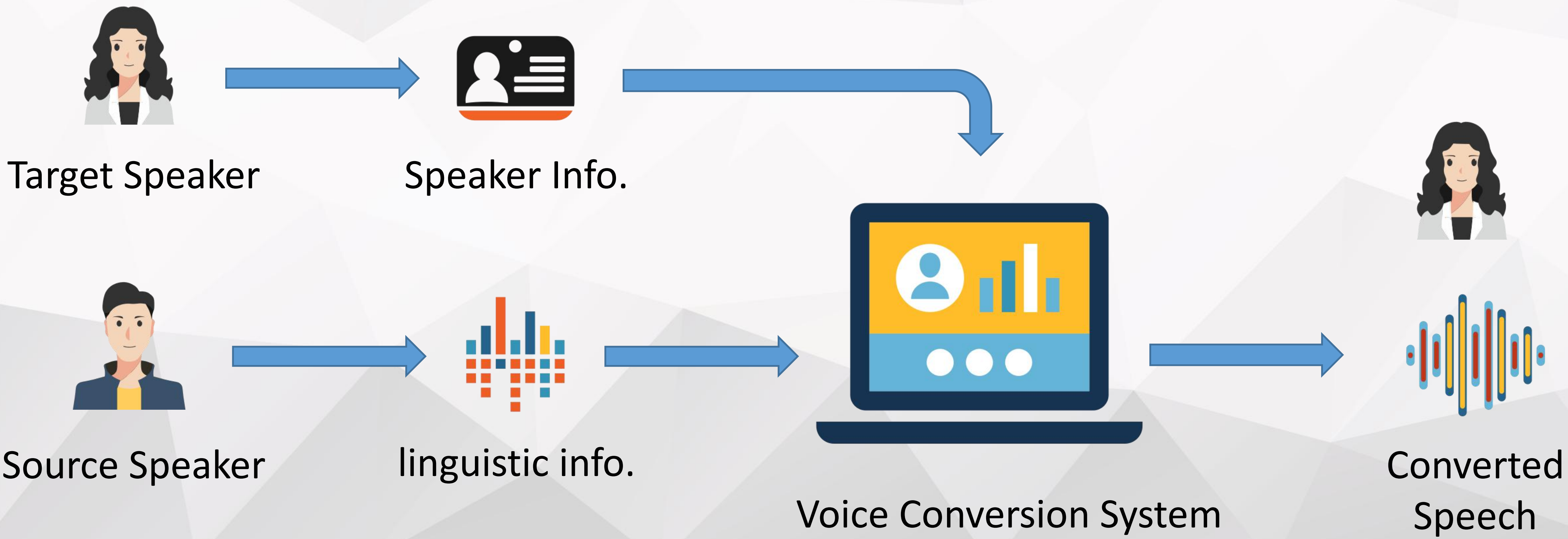
22-27 MAY 2022

ICASSP 2022
SINGAPORE

NAGOYA UNIVERSITY

Voice Conversion

Voice Conversion (VC) is a technique that modifies the speaker's identity to the target speaker without changing the linguistic information.



Target Speaker

Speaker Info.

Source Speaker

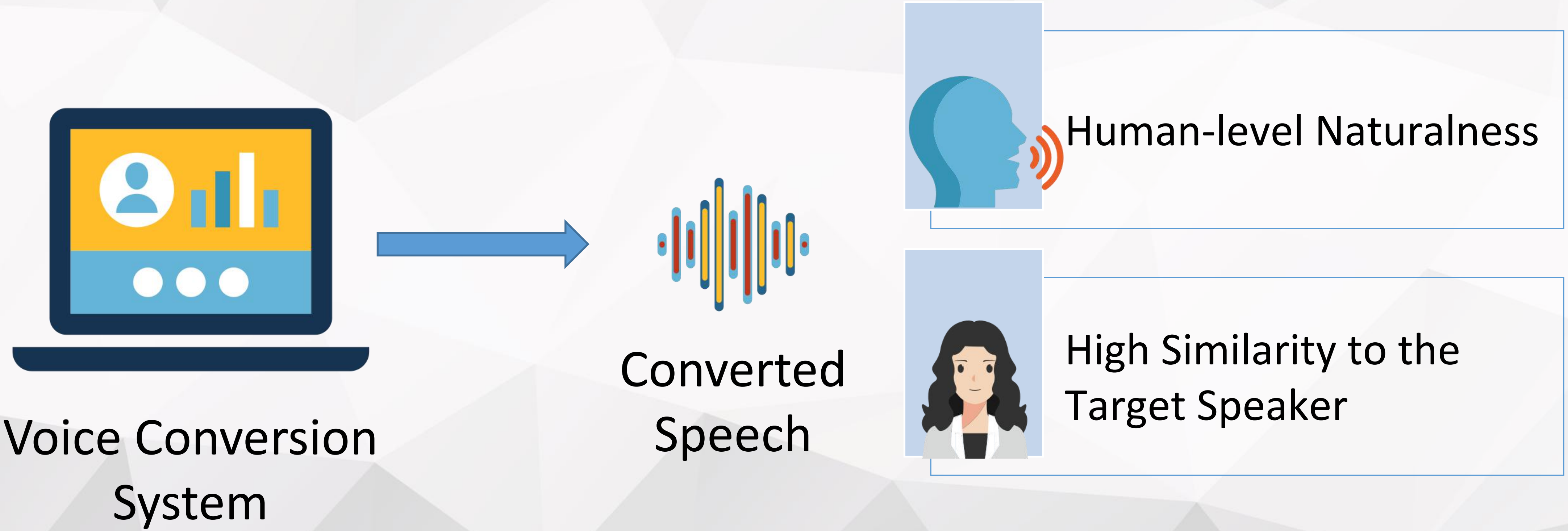
linguistic info.

Voice Conversion System

Converted Speech

About Voice Conversion Challenge* (VCC)

Goal: Reach human-level naturalness and high similarity to the target speaker.



Voice Conversion System

Converted Speech

Human-level Naturalness

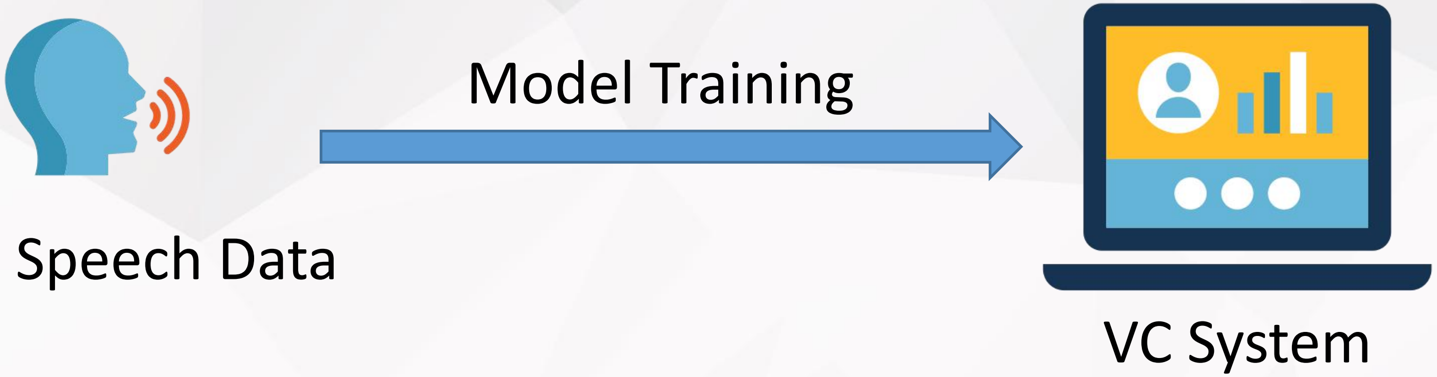
High Similarity to the Target Speaker

* Zhao Y, Huang WC, Tian X, Yamagishi J, Das RK, Kinnunen T, Ling Z, Toda T. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. arXiv preprint arXiv:2008.12527. 2020 Aug 28.

VC in Real-World (Noisy Environment)

Well-trained VC system deployed in noisy environment:

Training:

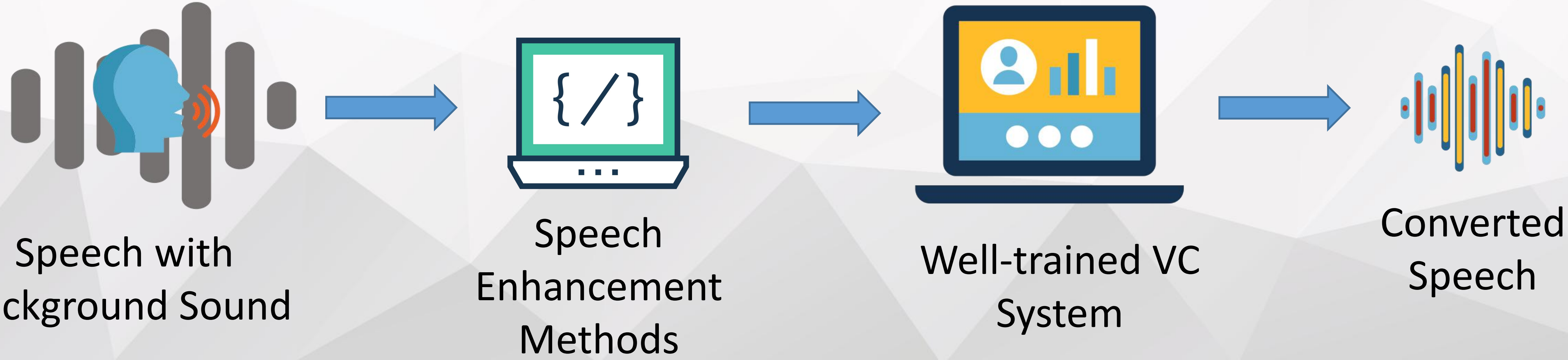


Model Training

Clean Speech Data

VC System

Converting:



Speech with Background Sound

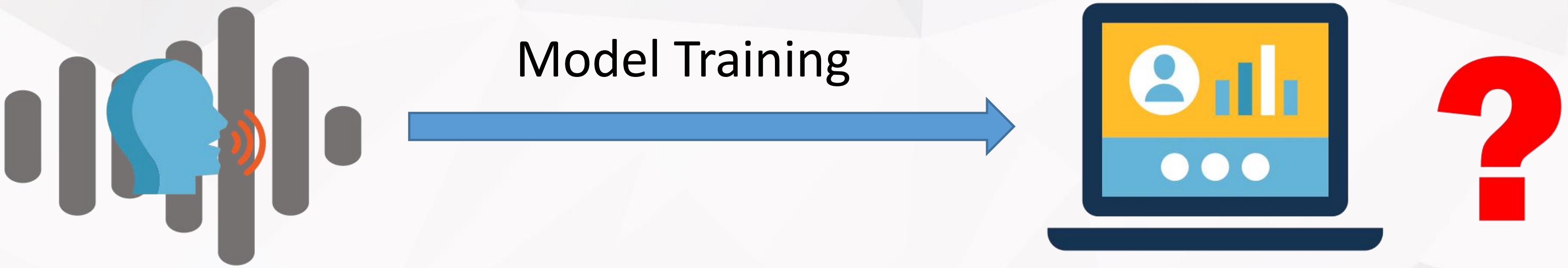
Speech Enhancement Methods

Well-trained VC System

Converted Speech

VC in Real-World (More Challengeable)

Only NOISY source/target speech data are available in the training stage:



Source/Target Speech
with Background Sound

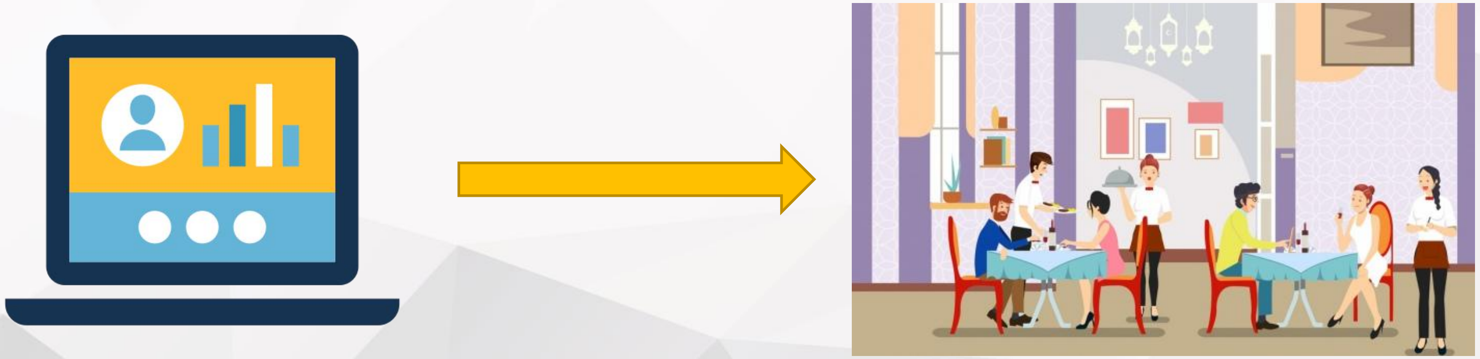
VC System

- High-quality source/target speech data are costly to collect;
- Directly training on the noisy dataset can not guarantee the performance of the VC model.

Flexible Dealing with Background Sound

Background sound is **Annoying**, but **Not Useless.**
Depending on different scenarios, the background sound should be **suppressed** or **maintained.**

Noise-Robust VC: Background sound is surppressed to reduce the interference.



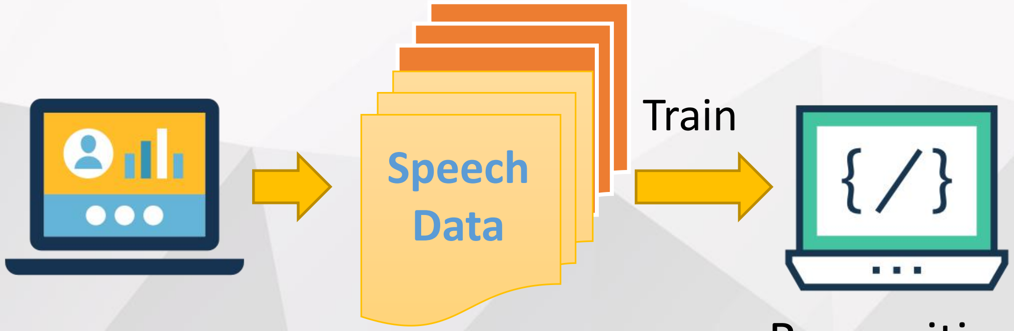
VC System in Noisy Environment

Flexible Dealing with Background Sound

Noisy-to-Noisy VC: Retain the background noise/voice while converting the voice.

VC System in Movies/Video

Only the speech is converted without changing the information of background voice/music.



VC System for Data Augmentation

Recognition System

The background sound in the dataset is also kind of **‘Resources’:**
It is desired that such noise can be preserved to improve the robustness of the recognition system.

Research Target

Noisy-to-Noisy (N2N) Voice Conversion Framework

The **First “Noisy”** means:

- We can only get noisy source/target speech data to train the VC model.

y: Noisy speech **s:** Clean speech

h: Room impulse response **n:** Noise signal

The real-world noisy speech can be represented as: **y = s*h + n.**

Our current research focus on the noisy speech: y = s + n

The **Second “Noisy”** means:

- We convert the speaker information but retain the background sound.
- We can either keep the background sound or suppress it, according to individual applications.

Final Goal: VC with flexible controlling the background noise (SNR levels, noise category).

Related Work

Most previous researches focus on noise-robust VC: the background sounds are considered as interference to be discarded.

Hsu *et al.** proposed a text-to-speech(TTS) based VC method using factorized latent variables to
control the noise in the converted speech:

- The clean speech data are augmented with a noisy copy to train a VAE to learn the disentangled representations of the speaker identity and background noise.
- Augmentation-adversarial training is utilized to further increase the degree of the disentanglement.

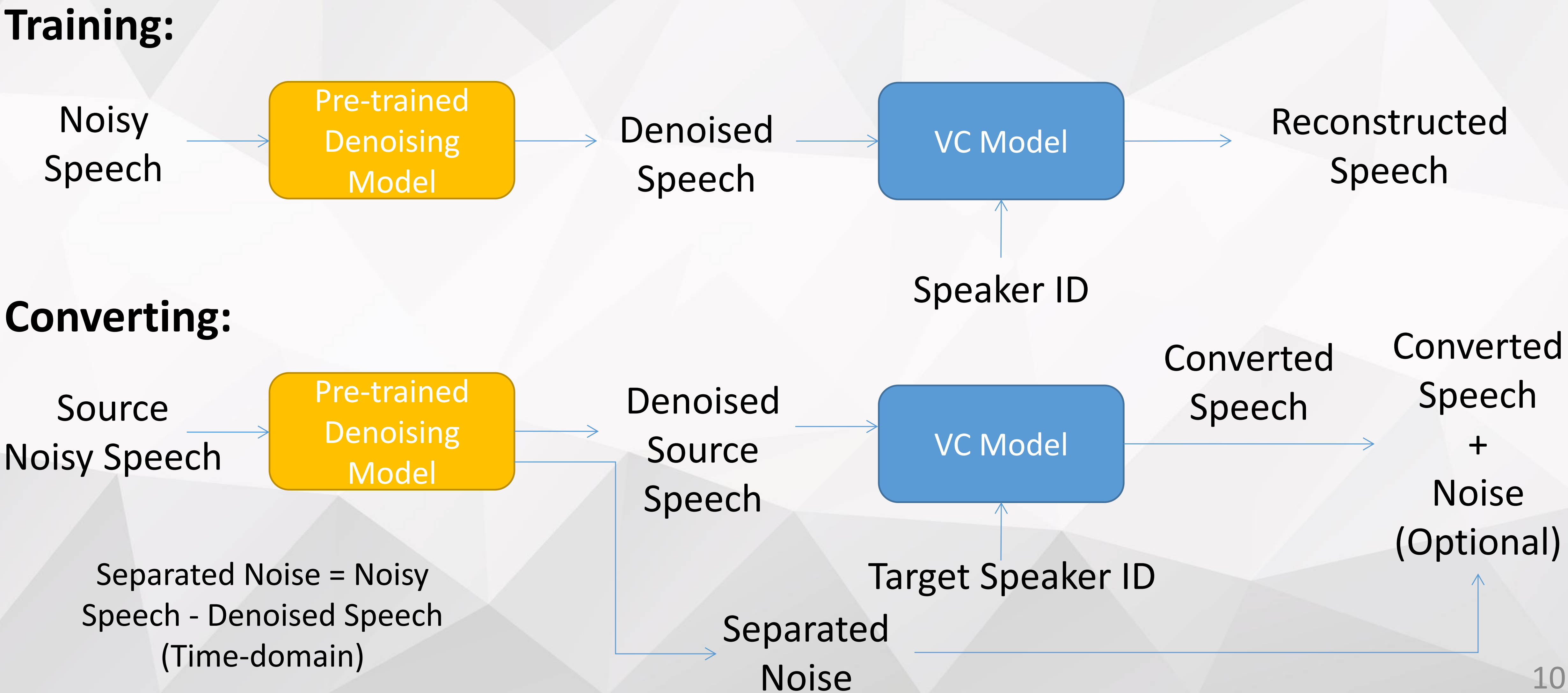
However,

- The quality of the background noise in the converted speech is quite limited.
- The clean source/target data for VC is still necessary.

* “Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization,” in Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5901–5905

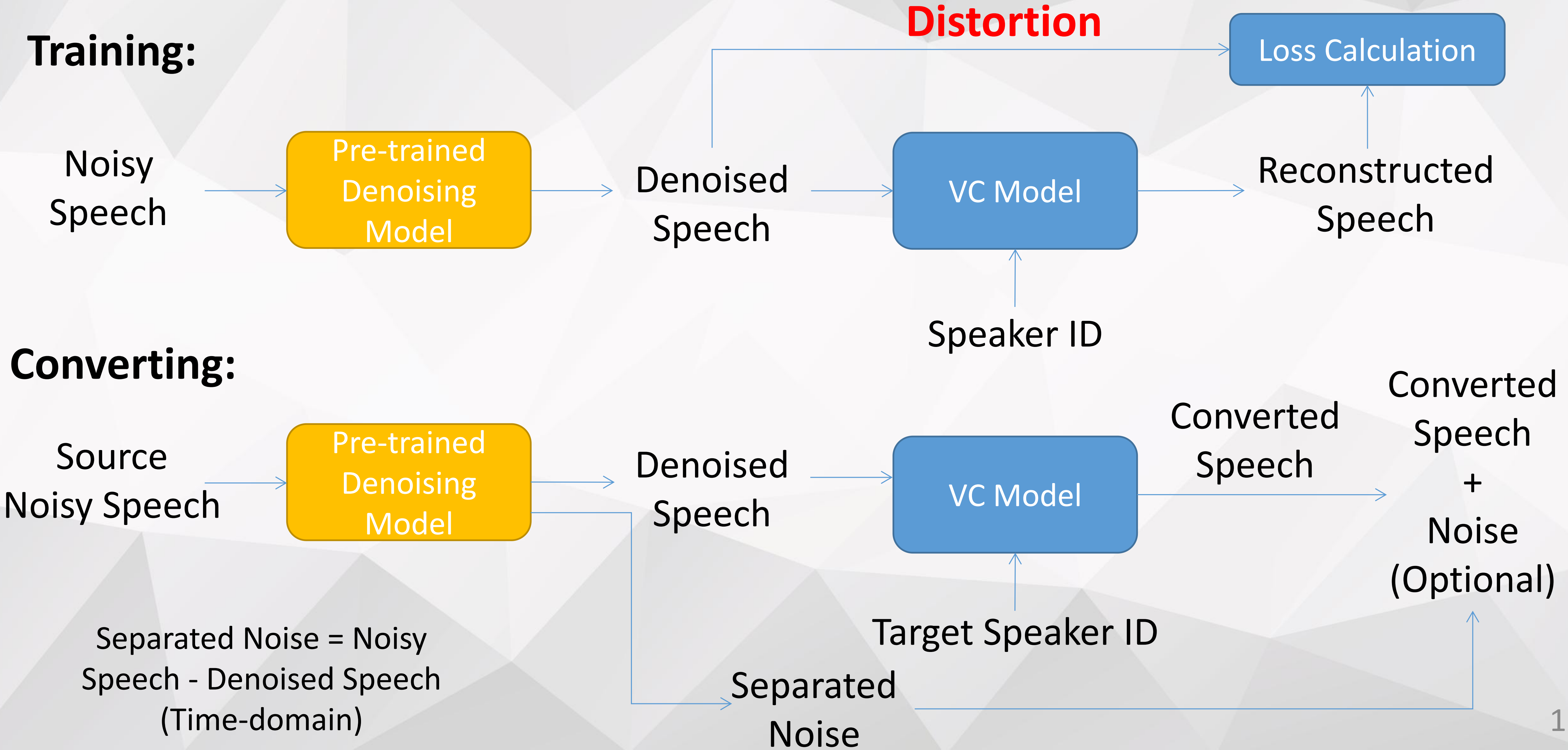
Baseline N2N VC Framework

The framework consists of a pre-trained denoising model and a VC model:



Separated Noise = Noisy Speech - Denoised Speech (Time-domain)

Baseline N2N VC Framework



Training:

Noisy Speech

Pre-trained Denoising Model

Denoised Speech

Distortion

Loss Calculation

VC Model

Reconstructed Speech

Speaker ID

Converting:

Source Noisy Speech

Pre-trained Denoising Model

Denoised Speech

VC Model

Converted Speech

Converted Speech + Noise (Optional)

Target Speaker ID

Separated Noise

Separated Noise = Noisy Speech - Denoised Speech (Time-domain)

Problem Analysis

How to solve the distortion problem caused by the denoising model?
Re-think what data we can get:

Separated Noise (**Distortion**) = Noisy Speech (**Non-Distortion**) - Denoised Speech (**Distortion**)

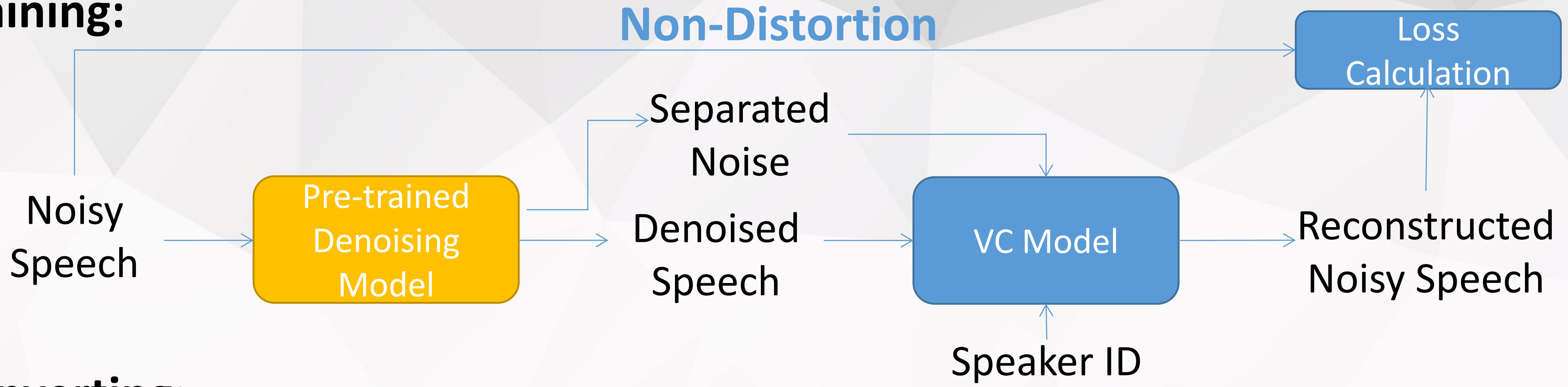
Only the noisy speech is **Non-Distortion.**
However, **Directly** modeling noisy speech is **DIFFICULT.**

IDEA:

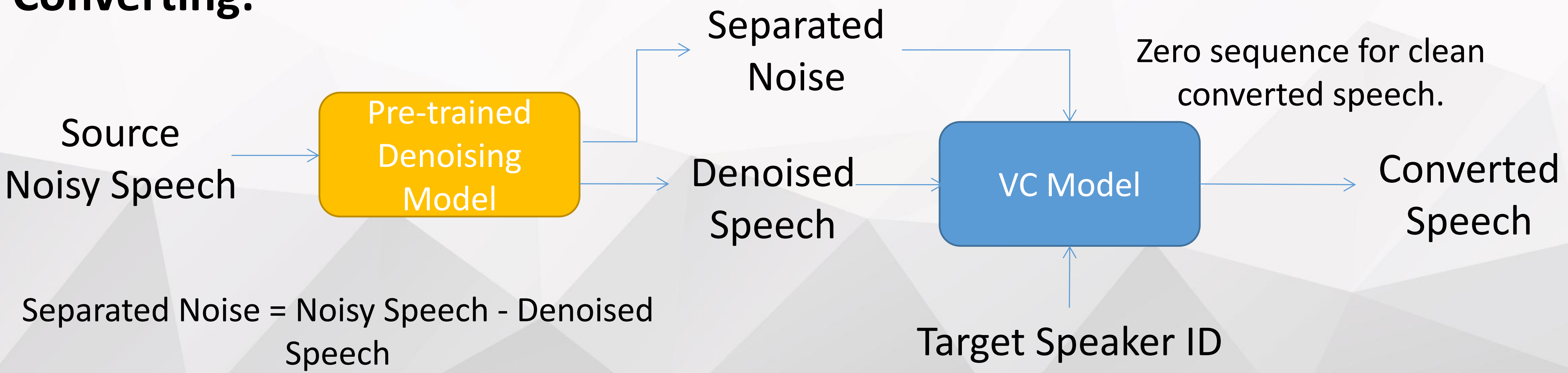
- Noisy speech is used as the training target in the VC model;
- The separated noise signal is provided as condition to the VC model to assist the difficult noisy speech modeling.

Improved N2N VC Framework

Training:



Converting:

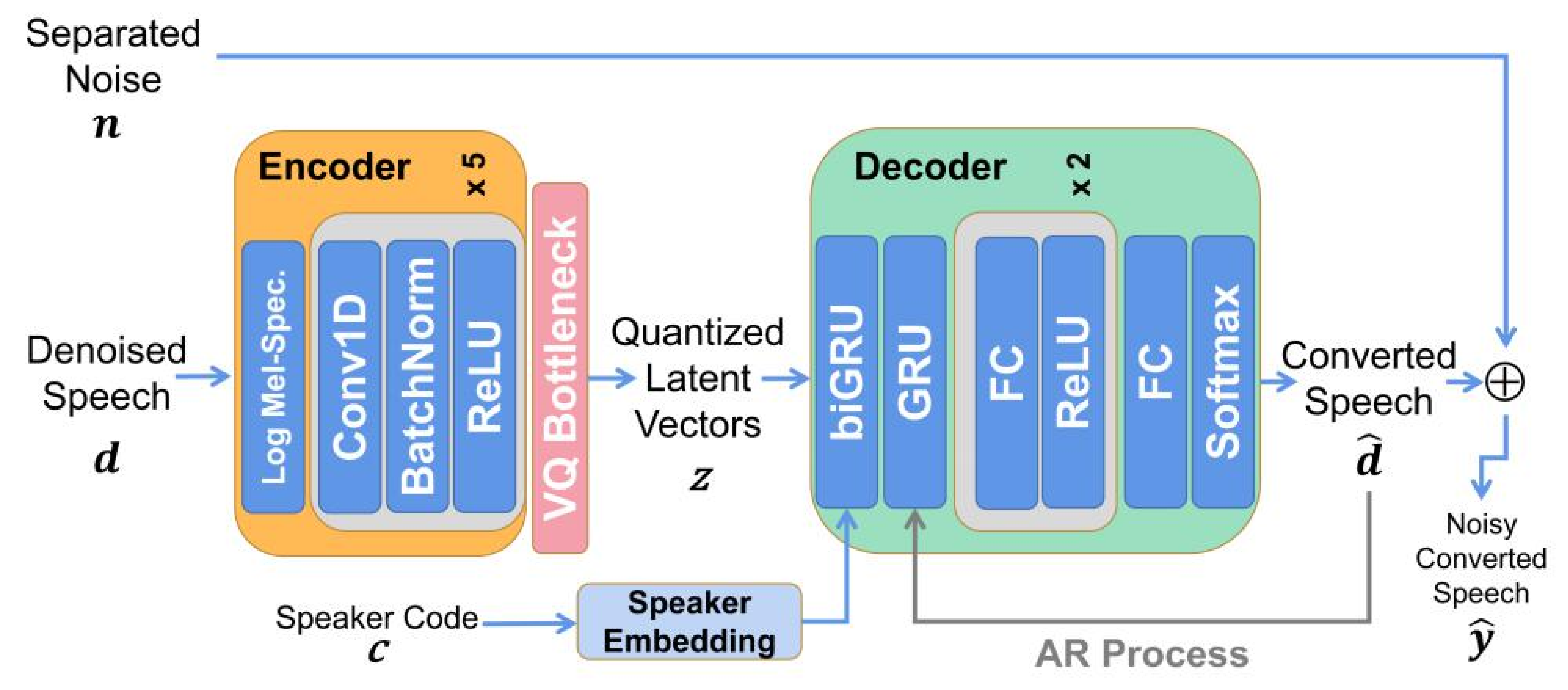


Zero sequence for clean converted speech.

Separated Noise = Noisy Speech - Denoised Speech
(Time-domain)

VC Model in the Baseline

VC Model: vector-quantized variational autoencoder (VQ-VAE)



Separated Noise $\boldsymbol{n}$
Encoder
x 5
Log Mel-Spec.
Conv1D
BatchNorm
ReLU
VQ Bottleneck
Denoised Speech $\boldsymbol{d}$
Quantized Latent Vectors $\boldsymbol{z}$
Decoder
x 2
biGRU
GRU
FC
ReLU
FC
Softmax
Converted Speech $\hat{\boldsymbol{d}}$
$\oplus$
Noisy Converted Speech $\hat{\boldsymbol{y}}$
Speaker Code $\boldsymbol{c}$
Speaker Embedding
AR Process

VC Model in the Baseline

VC Model: vector-quantized variational autoencoder (VQ-VAE)

y: Noisy speech

d: Denosied speech estimated by denoising model

n: Separated noise signal: **n** = **y** - **d** (Time domain)

z: Latent representation from VQ-bottleneck (Input of the encoder is **d**)

c: Speaker code

Decoder (Autoregressive):

$$p(\mathbf{d} \mid \mathbf{c}, \mathbf{z}) = \prod_{t=1}^{T} p(d_t \mid d_1, \ldots, d_{t-1}, \mathbf{c}, \mathbf{z})$$

Modified VC Model

y: Noisy speech
d: Denosied speech estimated by denoising model
n: Separated noise signal: **n** = **y** - **d** (Time domain)
z: Latent representation from VQ-bottleneck (Input of the encoder is **d**)
c: Speaker code

Decoder (Baseline):

$$p(\mathbf{d} \mid \mathbf{c}, \mathbf{z}) = \prod_{t=1}^{T} p(d_t \mid d_1, \ldots, d_{t-1}, \mathbf{c}, \mathbf{z}) \quad (1)$$

We hope to utilize the **Noisy Speech y** **(Non-distortion)** as the optimization target.
Considering **y** = **d** + **n** (Time domain)，the (1) is changed to:

$$p(\mathbf{y} \mid \mathbf{c}, \mathbf{z}) = \prod_{t=1}^{T} p(y_t \mid d_1 + n_1, \ldots, d_{t-1} + n_{t-1}, \mathbf{c}, \mathbf{z}), \quad (2)$$

However:

- This will force the decoder to learn the distribution of the noise.
- The decoder can not know the relationship of **y** = **d** + **n** (Time domain); which loses the controllability of the background noise.

Modified VC Model

y: Noisy speech
d: Denosied speech estimated by denoising model
n: Separated noise signal: **n** = **y** - **d** (Time domain)
z: Latent representation from VQ-bottleneck (Input of the encoder is **d**)
c: Speaker code

Decoder (Baseline):

$$p\left(\mathbf{d} \mid \mathbf{c}, \mathbf{z}\right) = \prod_{t=1}^{T} p\left(d_t \mid d_1, \ldots, d_{t-1}, \mathbf{c}, \mathbf{z}\right) \quad (1)$$

The history of the noise **n** is provided to the decoder as an assistant to model the noisy speech. We let the decoder learn the relationship of **y** = **d** + **n** (Time domain)

Proposed (Indrect):

$$p\left(\mathbf{y} \mid \mathbf{n}, \mathbf{z}\right) = \prod_{t=2}^{T} p\left(y_t \mid d_1, \ldots, d_{t-1}; n_2, \ldots, n_t; \mathbf{z}\right)$$

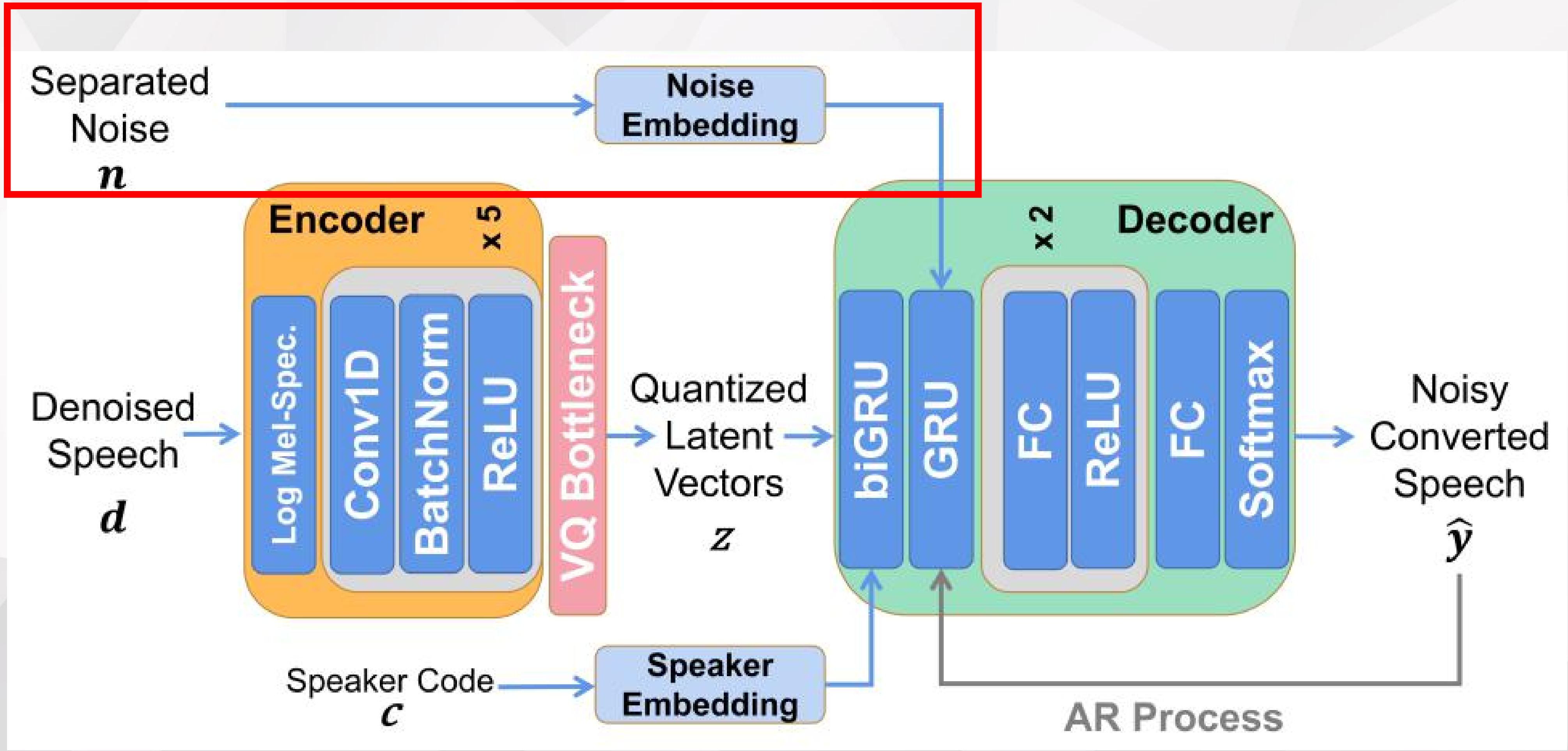
To further reduce the usage of destorted data **d** as the ground-truth in the teacher-forcing:

Proposed (Direct):

$$p\left(\mathbf{y} \mid \mathbf{n}, \mathbf{z}\right) = \prod_{t=2}^{T} p\left(y_t \mid y_1, \ldots, y_{t-1}; n_2, \ldots, n_t; \mathbf{z}\right)$$

Modified VC Model

The separated noise is used as condition to the decoder to assist the noisy speech modeling.



Denoising Model Settings

Denoising Model:

- Deep Complex Convolution Recurrent Network (DCCRN);
- Ranked 1st for the real-time-track in Deep Noise Suppression (DNS) Challenge 2020.

Separated Noise = Noisy Speech — Denoised Speech (Time-domain)

The power of the denoised speech should be matched to the clean reference speech. Hence, the original scale-invariant signal-to-noise ratio (SI-SNR) loss is substituted by the scale-dependent signal-to-distortion (SD-SDR) loss.

Dataset:

Deep Noise Suppression (DNS) Challenge 2020 dataset:

- 500 hours of speech from 2,150 speakers in various languages; 65,000 background clips.
- SNR levels: 5 dB to 20 dB.

VC Model (VQ-VAE) Settings

VC Model: Vector-quantized variational autoencoder (VQ-VAE) based VC model.

Noisy VC Dataset:

- Speech data: VCC 2018 dataset (12 speakers; 972 utterances for training; 420 utterances for evaluation)
- Noise data: PNL 100 Nonspeech Sounds (100 clips in 20 categories):

  For training set: N1 to N85 (85 clips in 9 categories)

  For testing set: N86 to N100 (15 clips in 11 categories)
- SNR levels:

  For training set: 6, 8, 10, 12, 14, 16, 18, 20 (dB)

The noisy VC dataset is unseen for the denoising model.

We evaluated the VC systems denoted as:

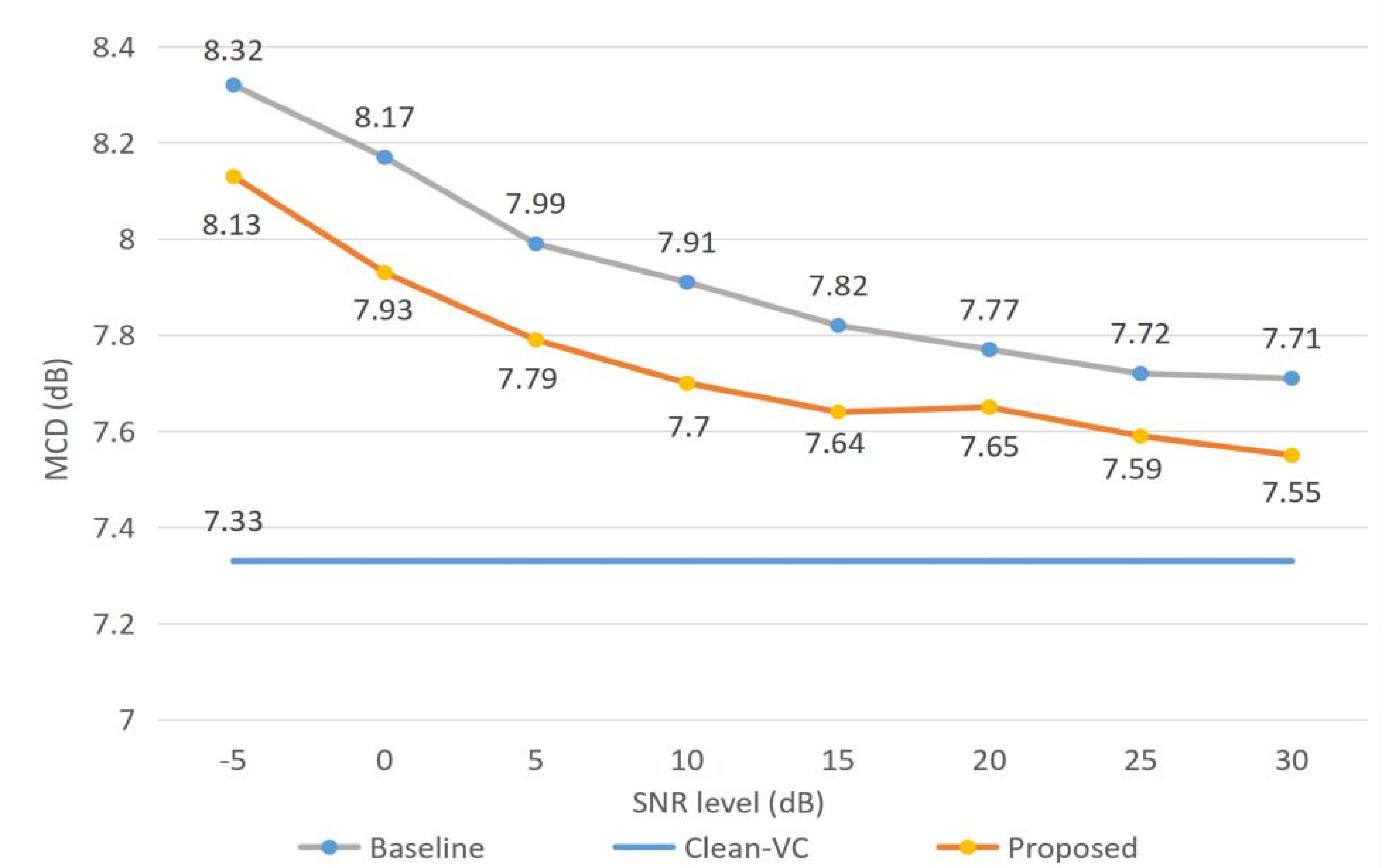
- **Clean-VC**: The VQ-VAE trained on the original VC dataset (The **upper bound** of the framework)
- **Baseline**: The VQ-VAE trained on the denoised noisy VC dataset.
- **Proposed**: The nois-conditioned VQ-VAE trained with denoised speech, separated noise and noisy speech:

  Indirect: the clean converted speeches were generated first and then superimposed with the separated noise.

  Direct: capable of synthesizing the noisy converted speech directly.

Objective Evaluation Results

Mel cepstral distortion (MCD) was employed as the objective measurement. (Lower is better)
Clean evaluation reference was leveraged.



Subjective Evaluation Settings

Mean opinion score (MOS) by an opinion test was applied to measure the naturalness of the converted noisy samples (Naturalness of the speech and the background sounds).

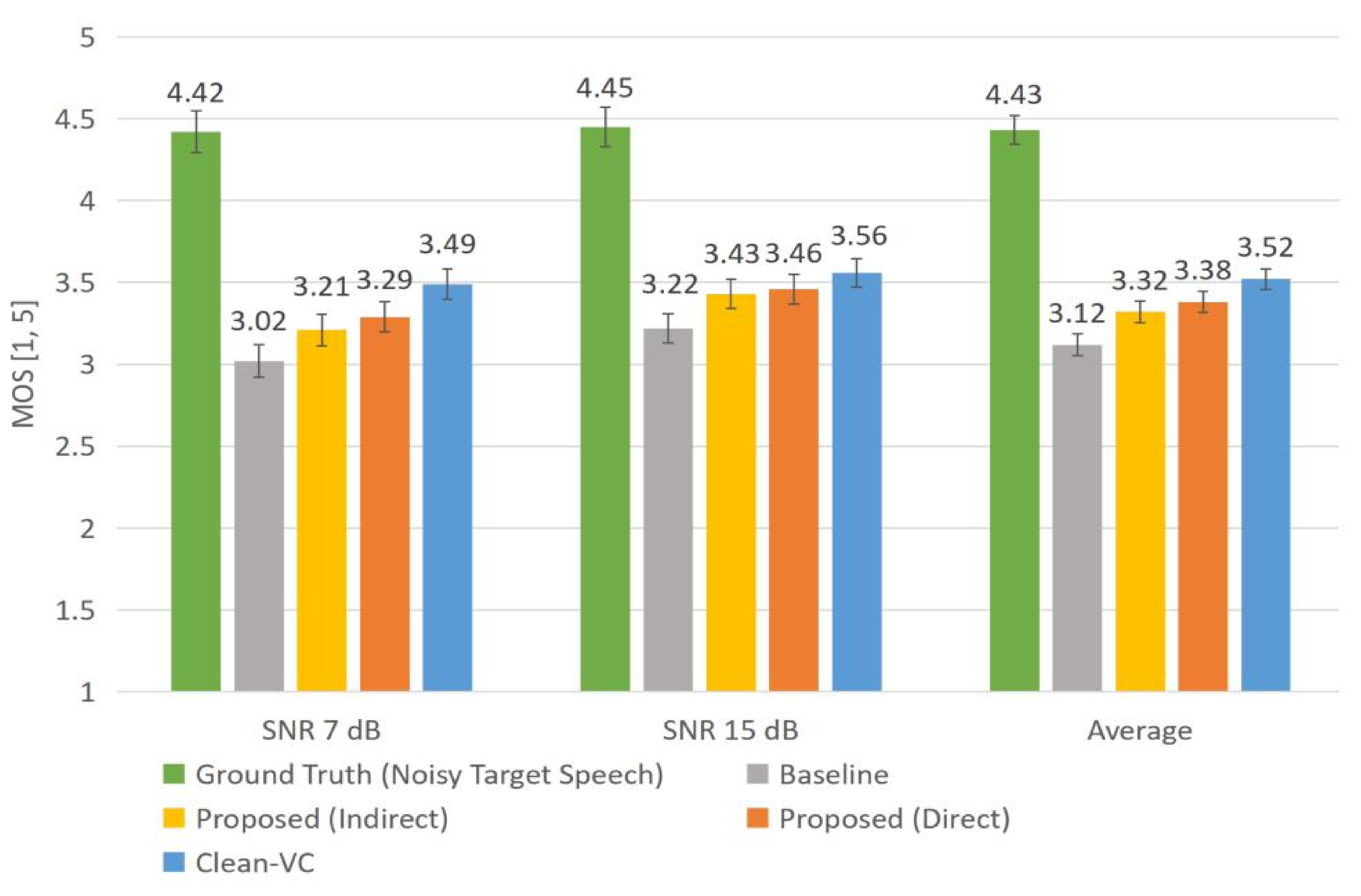
XAB test was conducted to evaluate the similarity quality of the converted samples.

- 15 participants.
- A total number of 340 audio samples : 80 audio samples per system and 20 samples from noisy ground-truth target speech.
- 10 samples under two SNR conditions (7 dB and 15 dB) for each conversion pair (8 conver).

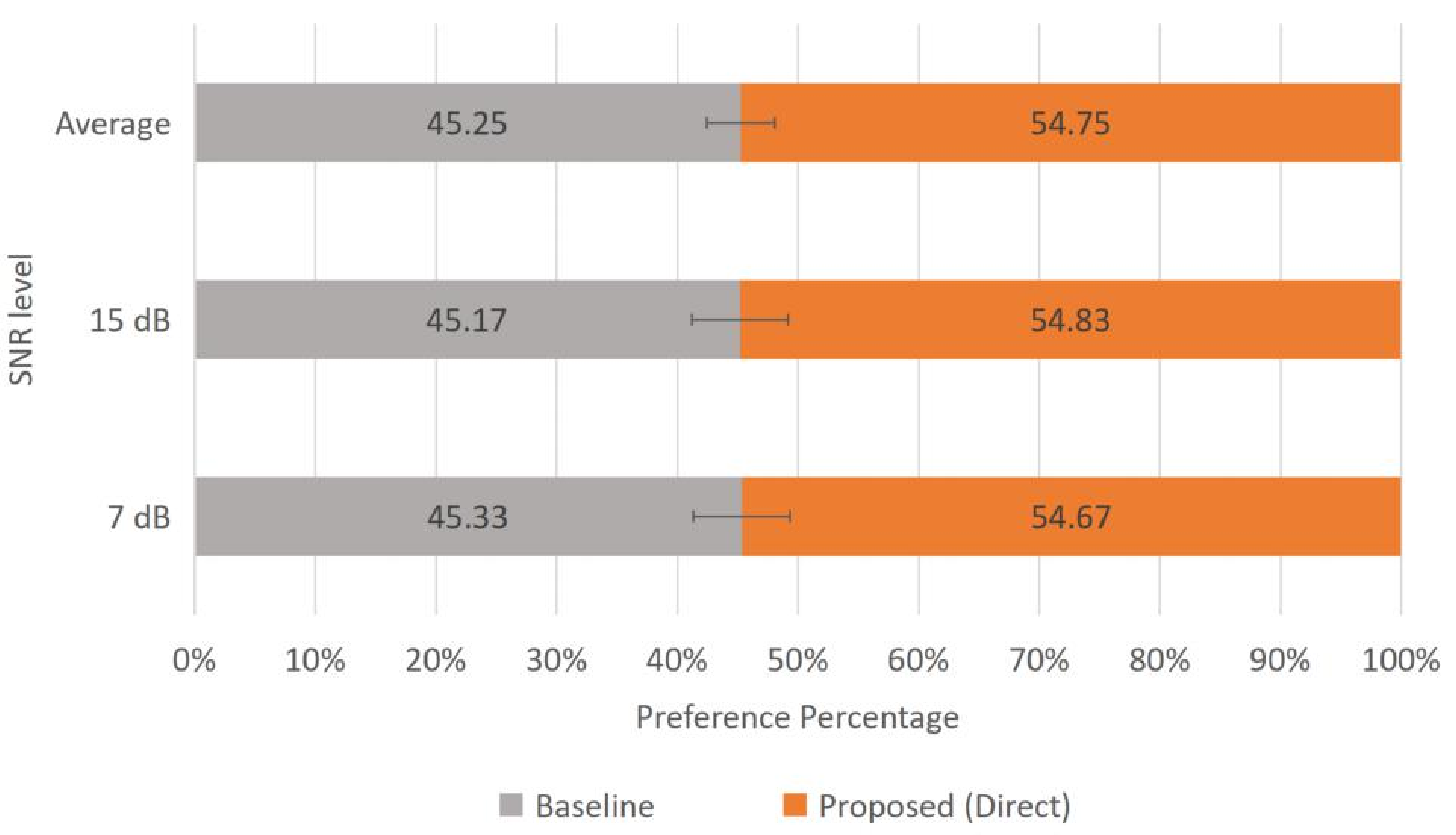
As our goal is N2N VC, the subjective evaluation was conducted on **noisy VCC evaluation dataset:**

- In MOS test, the naturalness of the background sounds was also taken into consideration.
- The categories of the background sounds and its original clip were provided to the participants during the evaluation.
- In XAB test, the participants were asked to ignore the background sounds.

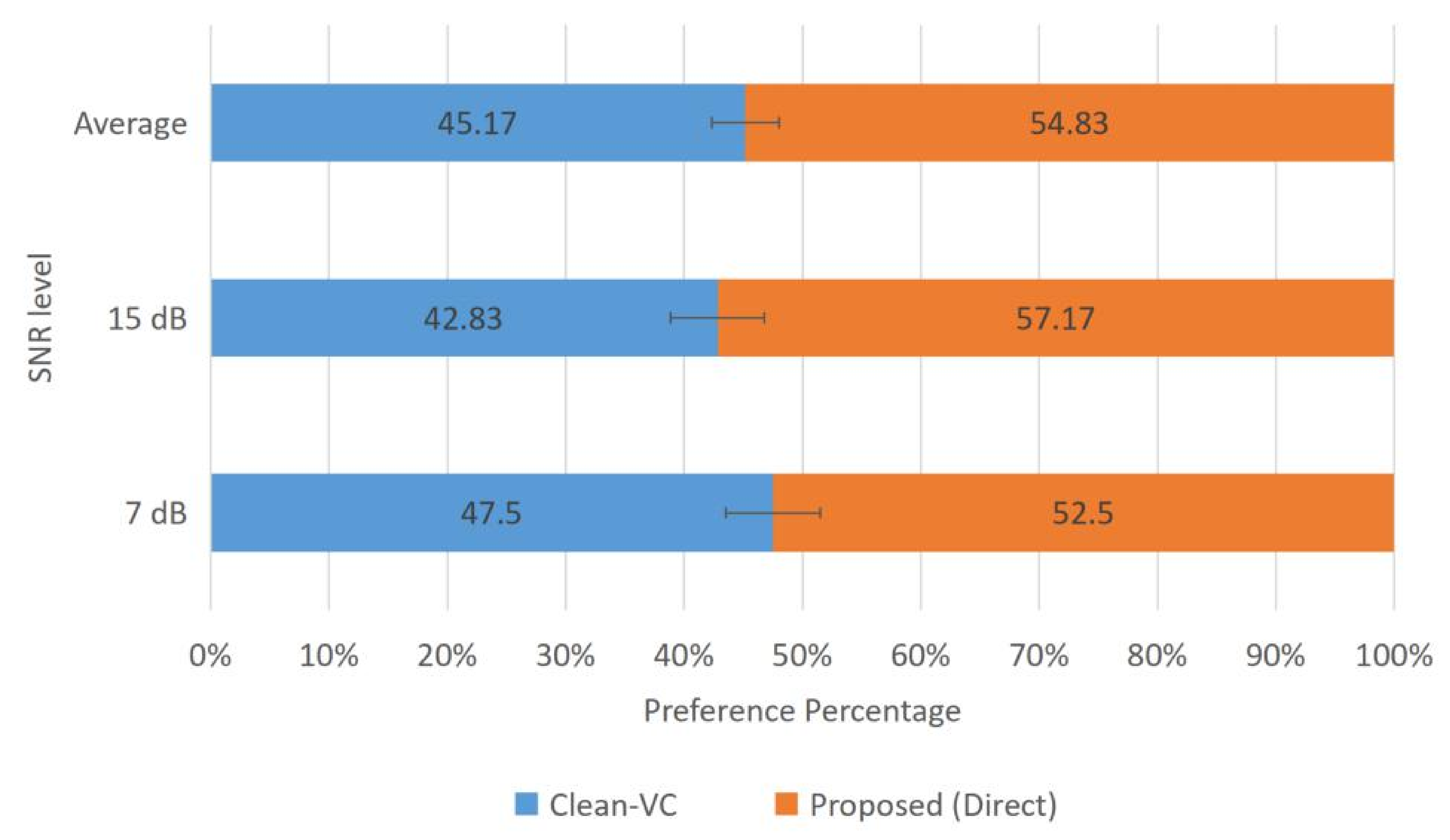
Subjective Evaluation MOS (Naturalness)



Subjective Evaluation XAB (Baseline v.s. Proposed)



Subjective Evaluation XAB (Clean-VC v.s. Proposed)



Conclusion

The proposed (Direct) outperforms the baseline under all SNR levels.

The method is straightforward but effective: Shorten the margin (from 0.4 to 0.14) in the average MOS score between the baseline and the upper bound by 65%.

The proposed method has minor effects on the speaker identity.

One-step noisy speech generation proposed (Direct) still maintains the high quality of the background sounds.

Future Work

Further improve the performance by shortening the gap from the upper bound;

Further investigate the impact of the extreme noise condition on VC task;

Demo: https://github.com/chaoxiefs/n2nvc