DIRECT NOISY SPEECH MODELING FOR NOISY-TO-NOISY VOICE CONVERSION

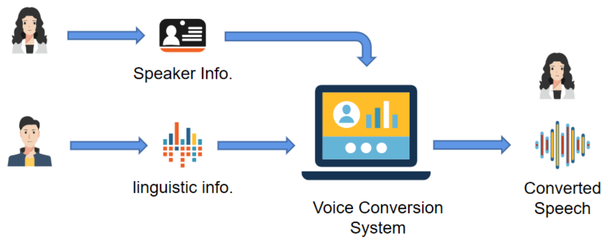
ID: 5347

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

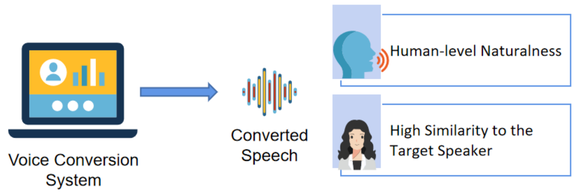
Nagoya University, Japan

Introduction

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



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



However, in the real world:

- High-quality source/target speech data are costly to collect;
- Directly training on the noisy dataset will significantly degrade the naturalness and similarity.

Noisy-to-Noisy (N2N) Voice Conversion

The **First "Noisy"** means:

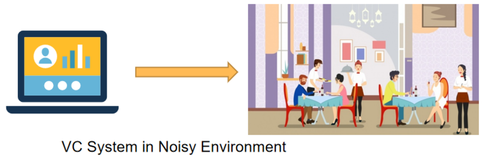
- We can only get noisy source/target speech data to train the VC model.
  y: Noisy speech **s**: Clean speech
  h: Room impulse response **n**: Noise signal
  The real-world noisy speech can be represented as: **y** = **s***h** + **n**.
  Our current research focuses on the noisy speech: y = s + n

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.

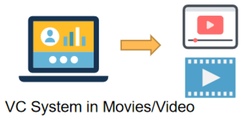
Application Scenarios

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



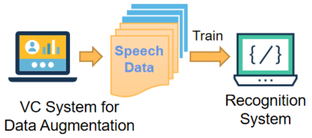
VC System in Noisy Environment

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.



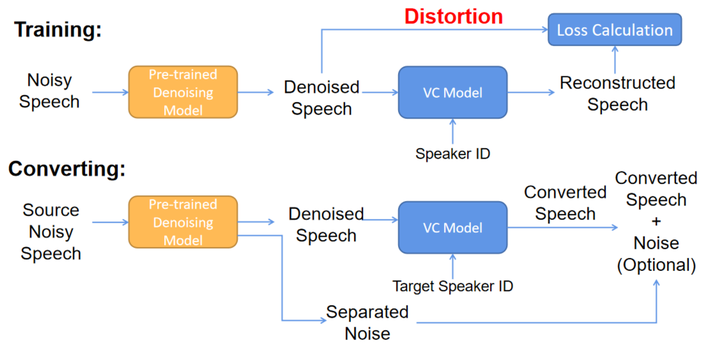
The background sound in the dataset is also a kind of **'Resource':** It is desired that such noise can be preserved to improve the robustness of the recognition system.

Proposed Method

The **Baseline N2N VC Framework** consists of a pre-trained denoising model and a VC model.

The denoising model is utilized to separate the speech and noise:

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



However, using denoised speech as the optimization target in VC model training will degrade the VC performance, for the data has **distortion** introduced by the denoising model.

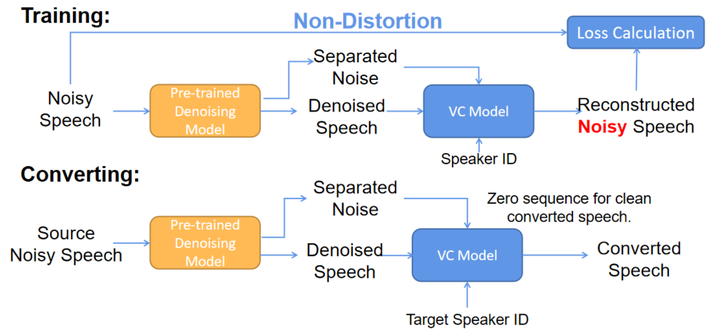
Re-think what data we have:

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

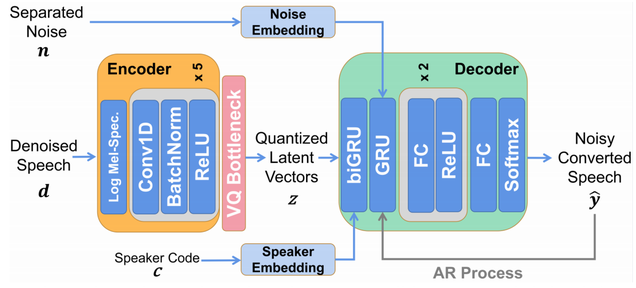
IDEA:

- 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.

The **Improved N2N VC Framework** uses **noisy** speech **(Non-distortion)** as the training target.



The modified VC model: noise-conditioned vector-quantized variational autoencoder (VQ-VAE)



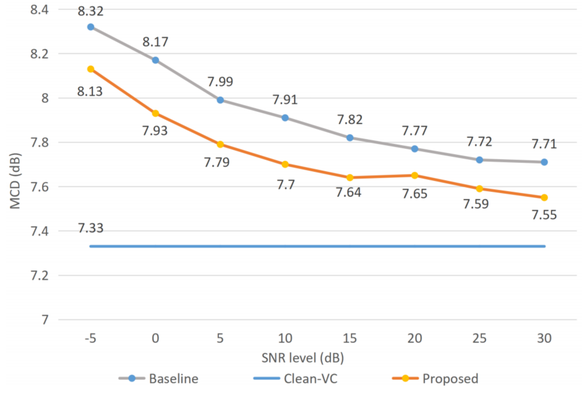
The decoder takes the noise signal as a condition to model the joint probability distribution of the noisy speech:

$$p(\mathbf{y} \mid \mathbf{n}, \mathbf{c}, \mathbf{z}) = \prod_{t=1}^{T} p(y_t \mid y_1, \ldots, y_{t-1}, n_1, \ldots, n_t, \mathbf{c}, \mathbf{z})$$

Experimental Results

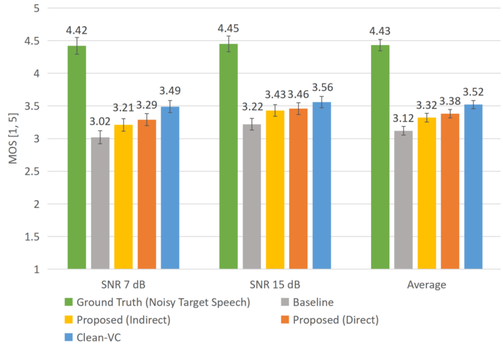
Objective Evaluation:

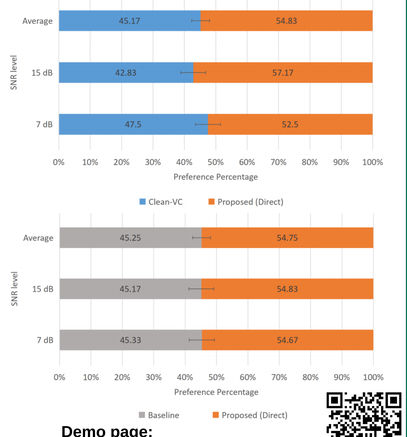
Mel cepstral distortion (MCD) was employed as the objective measurement. (Lower is better)



Subjective Evaluation:

Mean opinion score (MOS) to measure the naturalness (Left; Higher is better);
XAB test to compare the similarity (Right; Higher is better).





Conclusion:

- The proposed method significantly improves the naturalness of the baseline;
- The proposed method has minor effects on the speaker's identity.

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