**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.
- For example:
  - y: Noisy speech
  - s: Clean speech
  - h: Room impulse response
  - n: Noise signal

The real-world noisy speech can be represented as: \( y = s \ast h + n \).

Our current research focuses on the noisy speech: \( y = s \ast h \).

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 suppressed to reduce the interference.
- **Noisy-to-Noisy VC:** Retain the background noise/voice while converting the voice.
- **VC System for Movies/Video:** Only the speech is converted without changing the information of background voice/music.

**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:

\[
\text{Separated Noise} = \text{Noisy Speech} - \text{Denoised Speech (Time-domain)}
\]

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:

\[
\text{Separated Noise (Distortion)} = \text{Noisy Speech (Non-Distortion)} - \text{Denoised Speech (Distortion)}
\]

**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