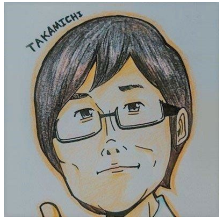
Mon-A7-O1 INTERSPEECH2024 Oral Session
(Speech Synthesis: Voice Conversion 1)

Spatial Voice Conversion:

Voice Conversion Preserving Spatial Information and Non-target Signals




Kentaro Seki[1]   Shinnosuke Takamichi[1,2]   Norihiro Takamune[1]   Yuki Saito[1]   Kanami Imamura[1]   Hiroshi Saruwatari[1]

1: The University of Tokyo, Japan / 2: Keio University, Japan

Overview of This Talk

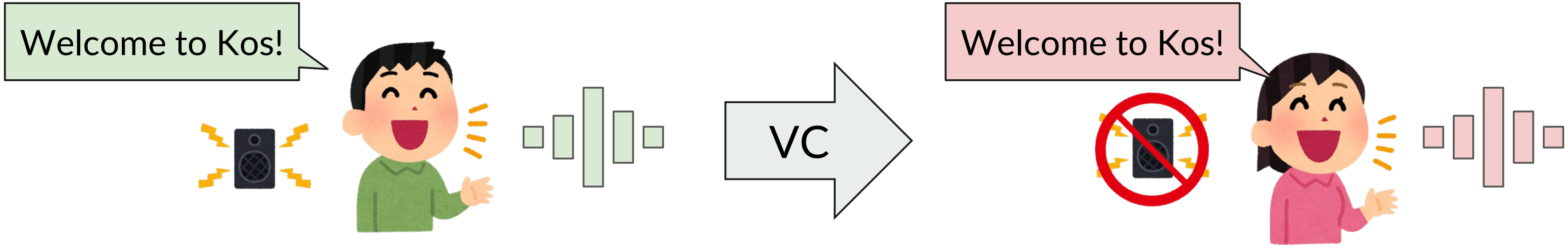
- We propose a new task: **Spatial Voice Conversion (Spatial VC)**
  - VC preserving spatial Information and non-target signals
- We propose a baseline method for spatia VC
  - **Combining BSS and VC** (BSS: blind source separation)
- We identify key challenges inherent in Spatial VC
  - Preserving spatial information may degrade audio quality

Outline

1. **Background**
2. Problem description
3. Method
4. Experimental results and Analysis

Background: Voice Conversion (VC)

- Convert speech to another speaker’s speech (same linguistic content)
- Conventional studies tend to ...
    - Deal with monaural signal (single channel)
    - Remove non-target audio (e.g., background noise) [1]



Background: Human Hearing

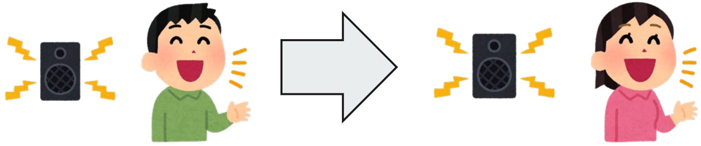
- Stereo hearing (not monaural)
    - Recognize spatial information (e.g., direction of the speaker)
- Even while focusing on a specific speech, other signals are still processed
    - For example, if an accident happens, we can recognize it
    - Non-target signals should not always be removed

Recognize spatial information

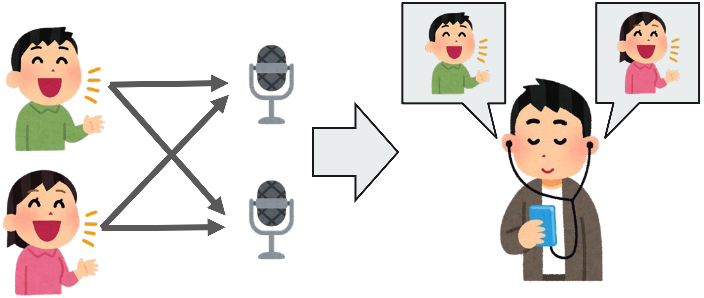
Process non-target signals

Related works

- **Example 1:** Noisy-to-noisy VC [2]
  - Preserve noise in input signal (not removed)
- **Example 2:** Stereo speech enhancement preserving spatial-cue [3]
  - Apply speech enhancement to two speakers's mixed speech
  - Output left speaker's speech to left-ch (maintain spatial cue)
- **These studied are based on spatial information and non-target signals**



Noisy-to-noisy VC



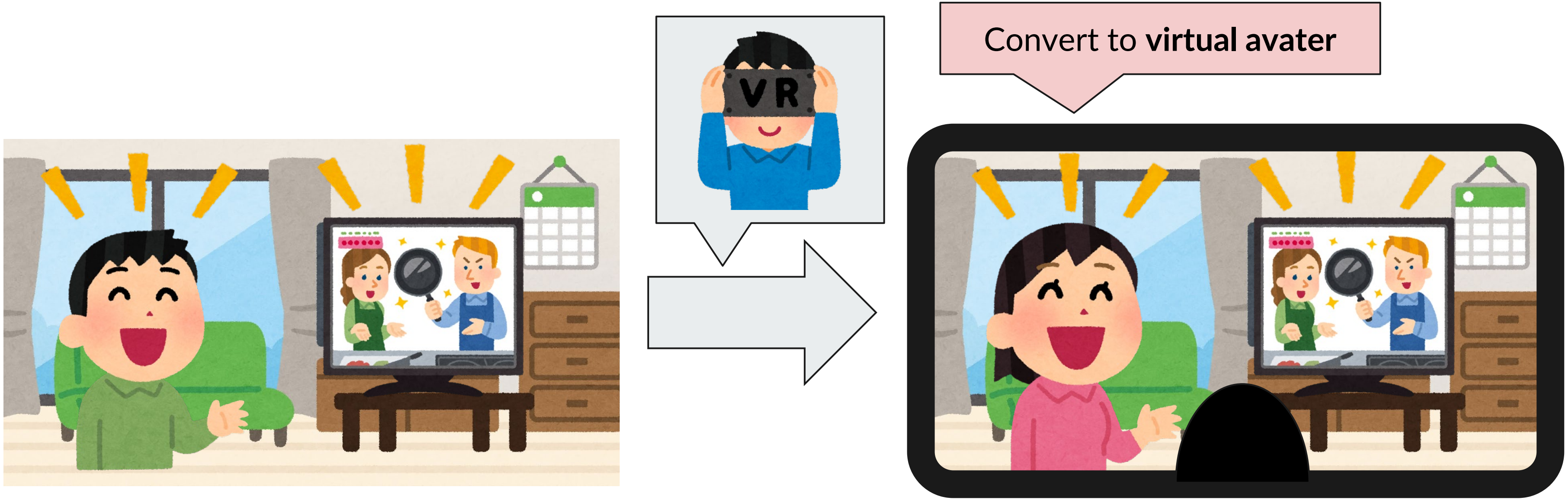
Stereo speech enhacement

Outline

1. Background
2. **Problem description**
3. Method
4. Experimental results and Analysis

Goal: Spatial VC enables

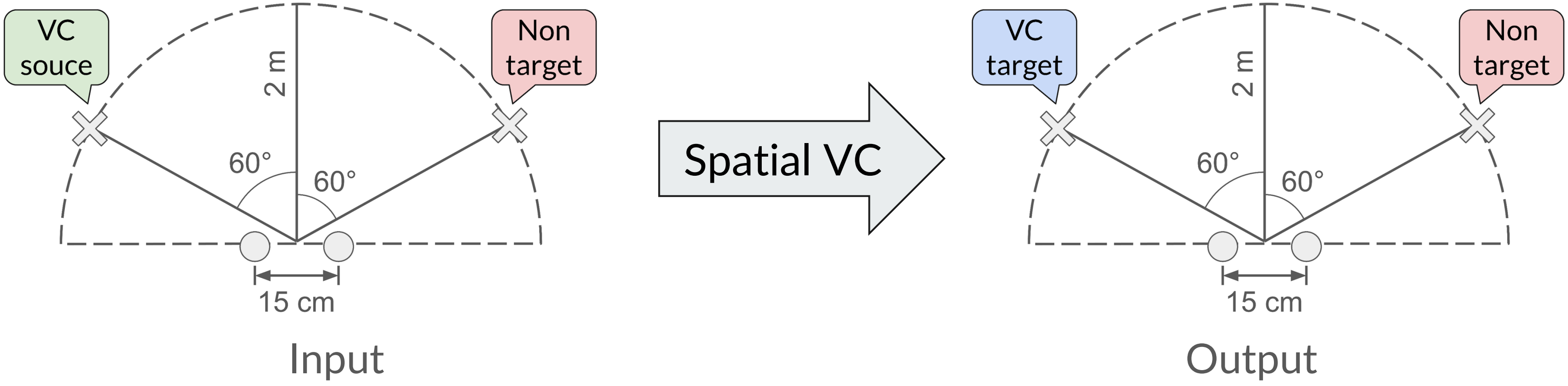
- Talk with virtual avatar in **Mixed Reality** (MR)



Real world | Mixed Reality

Task Description

- **Input**: Recording of mixed speech (VC source + non target)
  - Multi-channel signal (containing spatial information)
- **Output**: Recording of mixed speech (VC target + non target)



Problem Formulation

- **Input:** Sum of VC source speech and non-target signals

$$y_j(t) = h_{1j}(t) * x_1(t) + \sum_{i=2}^{N} h_{ij}(t) * x_i(t) \quad (j = 1, ..., N)$$

  - $x_1(t)$: VC source speech, $x_2(t), ..., x_N(t)$: Non-target signals
  - $h_{ij}(t)$: Transfer function from i-th source to j-th microphone
- **Output:** Applying VC exclusively to VC source speech

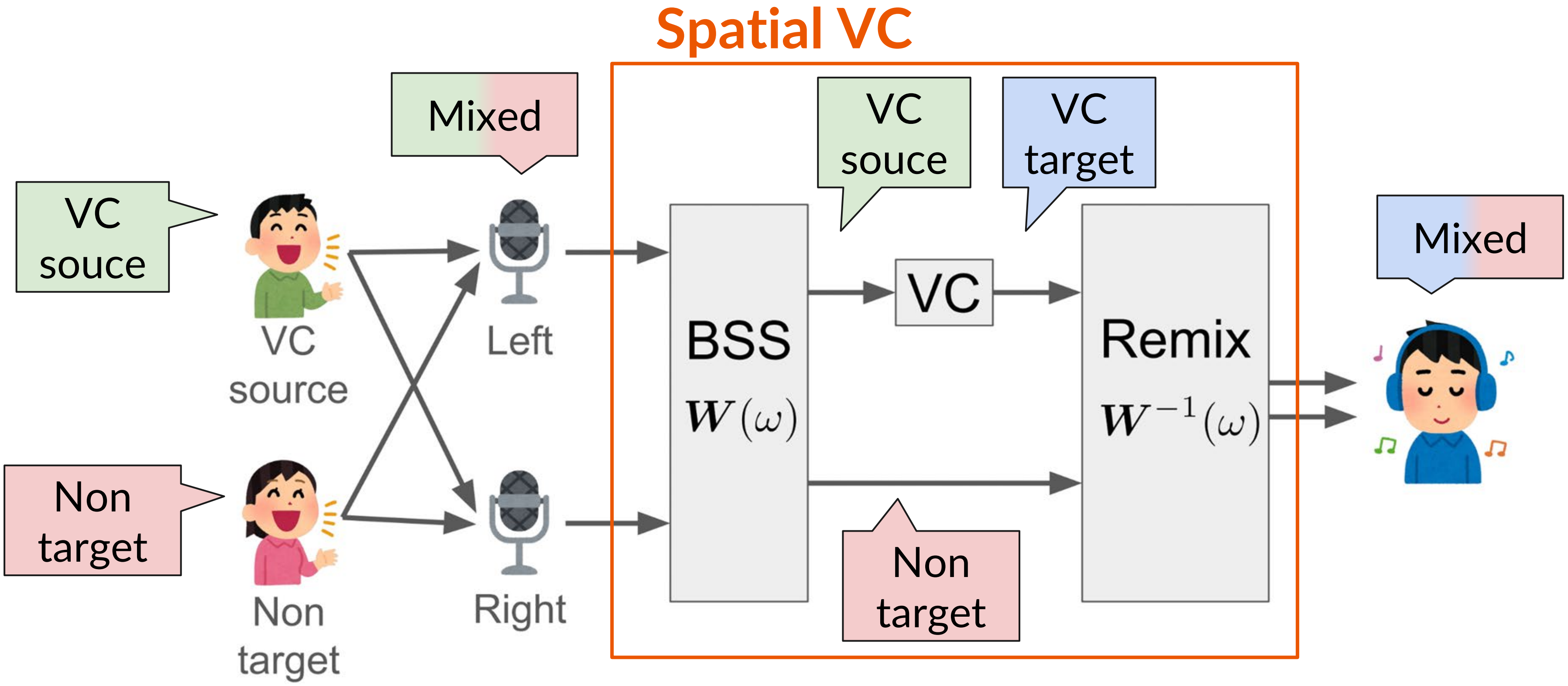
$$z_j(t) = h_{1j}(t) * \mathrm{VC}[x_1(t)] + \sum_{i=2}^{N} h_{ij}(t) * x_i(t) \quad (j = 1, ..., N)$$

  - $\mathrm{VC}[\cdot]$: Voice Conversion

Outline

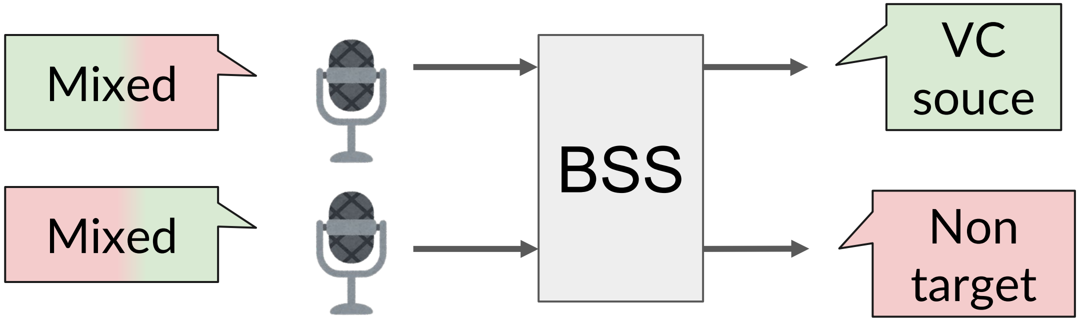
1. Background
2. Problem description
3. **Method**
4. Experimental results and Analysis

Method: Overview



Method: Blind Source Separation (BSS)

- Purpose of this step
  - Enhance VC-source speaker's voice
  - Extract non-target signals (to preserve non-target information)
- Apply BSS to obtain each signal separately



Method: Projection Back (in BSS)

- BSS is based on statistical independence, but cannot determine scale
- BSS output introduces a constant scaling factor (resulting in distortion)
- Projection Back: Determine scaling factor to reproduce microphone signal

New separation matrix

STFT of observed signals

STFT of 1st microphone

By PB, we update separation matrix

$$\underset{\boldsymbol{W}(\omega)}{\text{minimize}} \quad \mathbb{E}\left[\left\| \boxed{\mathbf{1}_N^\top \boxed{\boldsymbol{W}(\omega)} \boxed{\boldsymbol{Y}(\omega, t)}} - \boxed{Y_1(\omega, t)} \right\|^2\right],$$

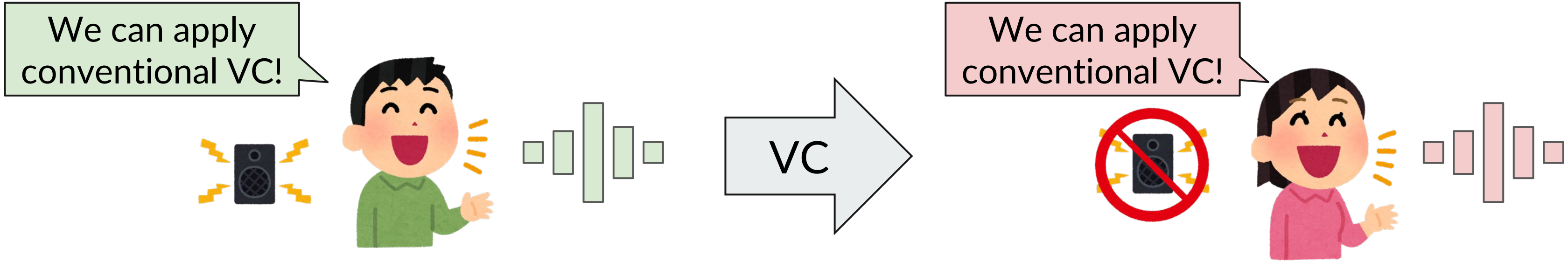
Summation of separated signals

Separation matrix determined by BSS

$$\text{subject to} \quad \exists k_1(\omega), \ldots, k_N(\omega),$$
$$\boldsymbol{W}(\omega) = \mathrm{diag}\{k_1(\omega), \ldots, k_N(\omega)\} \boxed{\boldsymbol{V}(\omega)}$$

Method: Voice Conversion (VC)

- Apply VC to extracted VC source speaker's speech
- This step: noisy-to-clean setting
  - Non-target signals are extratec in another channel
  - This channel can remove non-target signal



Method: Remix 1

- **Simple approach**: Apply inverse matrix of separation matrix ($\hat{\boldsymbol{A}}(\omega) := \boldsymbol{W}(\omega)^{-1}$)
  - "Mixing" is inverse process of "Separation"
  - Each component of $\hat{\boldsymbol{A}}(\omega)$ is expected to be transfer function

$$\hat{\boldsymbol{Z}}(\omega, t) = \hat{\boldsymbol{A}}(\omega) \begin{bmatrix} \mathrm{VC}[\hat{X}_1(\omega, t)] \\ \hat{\boldsymbol{X}}_{2:N}(\omega, t) \end{bmatrix}$$

- In reality, this may degrade audio quality of VC output!
  - After PB procedure, $\hat{\boldsymbol{A}}(\omega)$ correspond to relative transfer function
    - From PB-target microphone to another microphone)
  - This implicitly include unstable inverse filter

Method: Remix 2

- **Another approach**: Direct estimation of transfer function
  - We use steering vector as estimation
    - This model is based on the direction of arrival (DoA)
  - This model does not account for other spatial factors, (e.g., reverberation)
- For non-target signals, we applyinverse of separation matrix

Steering vector

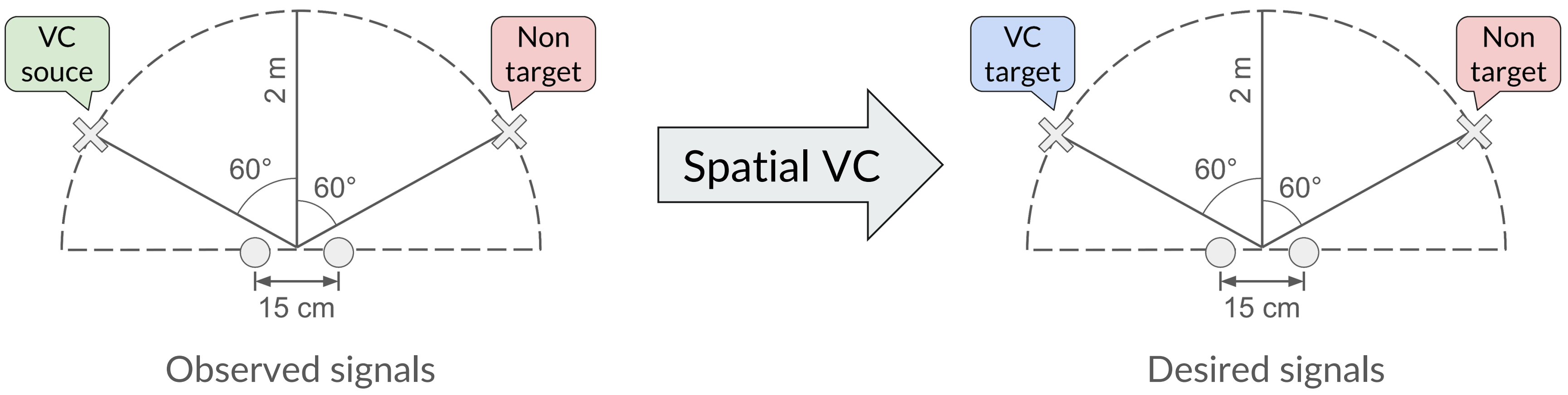
$$\hat{\boldsymbol{Z}}(\omega, t) = \left[\boxed{\boldsymbol{d}(\omega)} \quad \hat{\boldsymbol{A}}_{2:N}(\omega)\right] \begin{bmatrix} \mathrm{VC}[\hat{X}_1(\omega, t)] \\ \hat{\boldsymbol{X}}_{2:N}(\omega, t) \end{bmatrix}$$

Outline

1. Background
2. Problem description
3. Method
4. **Experimental results and Analysis**

Settings: Test Data

- Simulated with pyroomacoustics [4]
  - Reverberation time (RT_{60}): Approx. 200ms
  - Speaker: Randomly sampled from 10 speakers in JVS corpus [5]



Settings: Proposed Method

- **BSS:** Geometrically Constrained Independent Vector Analysis (GC-IVA) [6]
  - Features of GC-IVA:
    - Based on statistical independence of source signals
    - Utilize spatial regularization to specify VC-source speaker
  - Challenges in “Remixing” with inverse matrix:
    - Issues also arise with other BSS methods (e.g., ILRMA [7])
    - This problem is due to PB (common in linear BSS methods)
- **VC:** DDSP-SVC [8]
  - Open-source many-to-many voice conversion model

Compared Methods

- **Ideal**: Simulated desired signal (for comparison)
- **Inverse**: Spatial VC using inverse matrix for remixing
  - Preserve spatial information, but degrade quality
- **Steering**: Spatial VC using steering vector for remixing
  - Maintain audio quality, but discard spatial information except DoA.

Inverse of separation matrix

$$\hat{\boldsymbol{Z}}(\omega, t) = \boxed{\hat{\boldsymbol{A}}(\omega)} \begin{bmatrix} \mathrm{VC}[\hat{X}_1(\omega, t)] \\ \hat{\boldsymbol{X}}_{2:N}(\omega, t) \end{bmatrix}$$

Inverse

Steering vector

$$\hat{\boldsymbol{Z}}(\omega, t) = [\boxed{\boldsymbol{d}(\omega)} \quad \hat{\boldsymbol{A}}_{2:N}(\omega)] \begin{bmatrix} \mathrm{VC}[\hat{X}_1(\omega, t)] \\ \hat{\boldsymbol{X}}_{2:N}(\omega, t) \end{bmatrix}$$

Steering

Results: Acoustic Quality

- Calculated Mean Opinion Score (MOS) on naturalness
- "Inverse" exhibited a significantly lower score
  - Attributed to right channel (=other than PB target)
- **Indicating: Inverse matrix degrade audio quality**

Methods	Stereo	Monaural-Left (PB target)	Monaural-Right
Ideal	4.254 ± 0.060	4.062 ± 0.060	4.102 ± 0.058
Inverse	1.992 ± 0.064	4.142 ± 0.060	1.388 ± 0.042
Steering	3.405 ± 0.062	3.975 ± 0.066	3.132 ± 0.070

MOS (↑)

Results: Spatial Information

- Calculated Log-determinant divergence (LDD) [9] of spatial covariance matrix from “Ideal” to “Inverse” and “Steering”
  - Indicator for accuracy of spatial information reproduction
- “Inverse” is better than “Steering”
- **Indicating: Steering discards spatial information** (e.g., reverberation)

Methods	$RT_{60} \simeq 70$ ms	$RT_{60} \simeq 200$ ms	$RT_{60} \simeq 400$ ms
Inverse	$\mathbf{0.468 \pm 0.074}$	$\mathbf{0.979 \pm 0.147}$	$\mathbf{2.058 \pm 0.270}$
Steering	1.990 ± 0.367	3.097 ± 0.552	4.666 ± 0.535

LDD (↓)

Conclusion

- We proposed a new task: **Spatial Voice Conversion (Spatial VC)**
  - VC preserving spatial Information and non-target signals
- We proposed a baseline method for spatia VC
  - **Combining BSS and VC** (BSS: blind source separation)
- We identified key challenges inherent in Spatial VC
  - Preserving spatial information may degrade audio quality

Reference

[1] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noiserobust text-to-speech." in SSW, 2016, pp. 146–152.

[2] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy voice conversion framework with denoising model," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2021, pp. 814–820.

[3] M. Togami, J.-M. Valin, K. Helwani, R. Giri, U. Isik, and M. M. Goodwin, "Real-time stereo speech enhancement with spatialcue preservation based on dual-path structure," arXiv preprint arXiv:2402.00337, 2024

[4] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A ´ python package for audio room simulation and array processing algorithms," in Proc. ICASSP. IEEE, 2018, pp. 351–355.

Reference

[5] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, “Jvs corpus: free japanese multi-speaker voice corpus,” arXiv preprint arXiv:1908.06248, 2019.

[6] L. Li and K. Koishida, “Geometrically constrained independent vector analysis for directional speech enhancement,” in Proc. ICASSP. IEEE, 2020, pp. 846–850.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE/ACM TASLP, 24(9), 1626-1641.

[8] “DDSP-SVC,” https://github.com/yxlllc/DDSP-SVC.

[9] B. Kulis, M. A. Sustik, and I. S. Dhillon, “Low-rank kernel learning with bregman matrix divergences.” Journal of Machine Learning Research, vol. 10, no. 2, 2009.