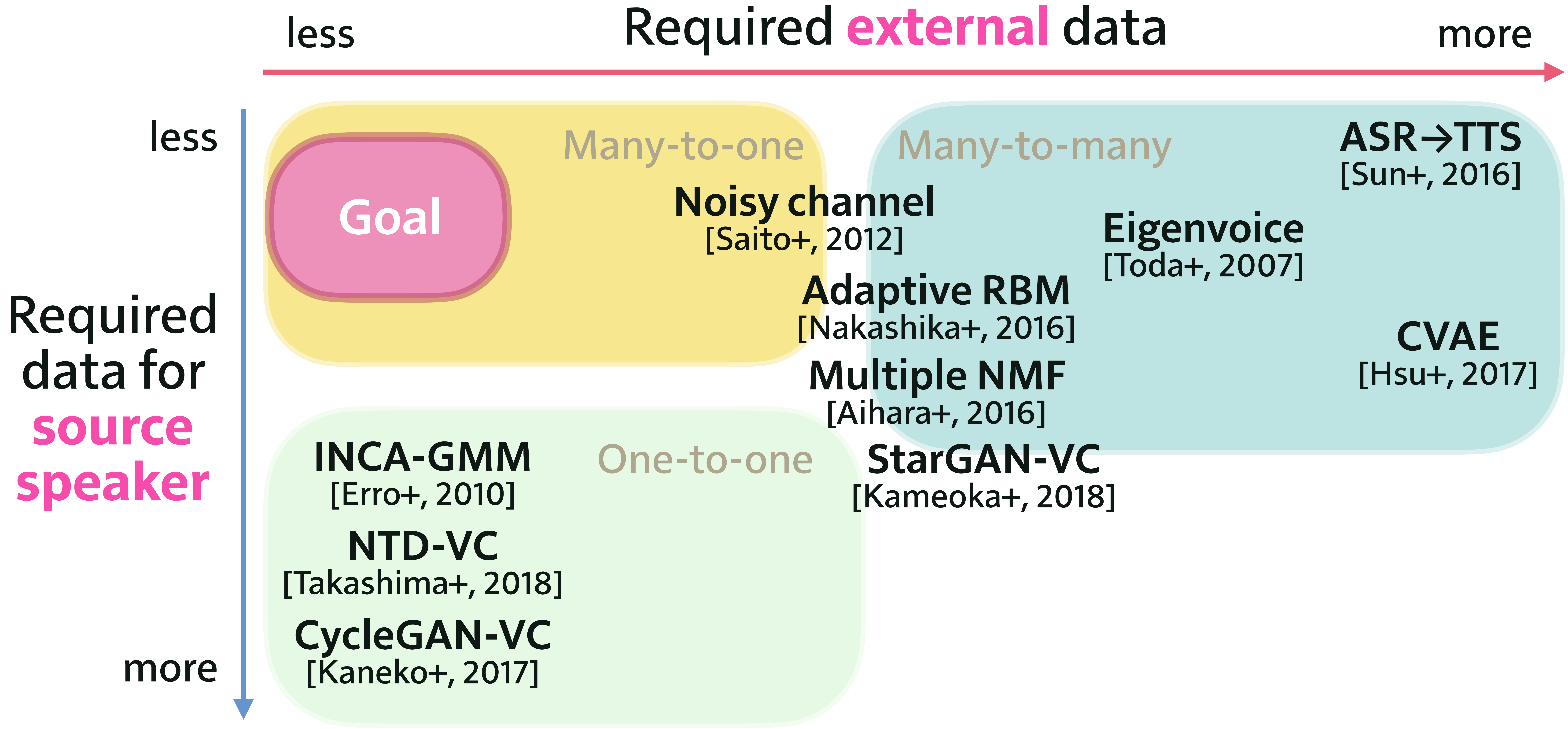
Nonparallel Training of Exemplar-based Voice Conversion System Using INCA-based Alignment Technique

Hitoshi Suda, Gaku Kotani, Daisuke Saito

The University of Tokyo

INTERSPEECH 2020 Virtual Conference

Background: Nonparallel Voice Conversion



less

Required **external** data

more

less

Required data for **source speaker**

more

Many-to-one

Goal

Noisy channel
[Saito+, 2012]

Many-to-many

ASR→TTS
[Sun+, 2016]

Eigenvoice
[Toda+, 2007]

Adaptive RBM
[Nakashika+, 2016]

CVAE
[Hsu+, 2017]

Multiple NMF
[Aihara+, 2016]

StarGAN-VC
[Kameoka+, 2018]

One-to-one

INCA-GMM
[Erro+, 2010]

NTD-VC
[Takashima+, 2018]

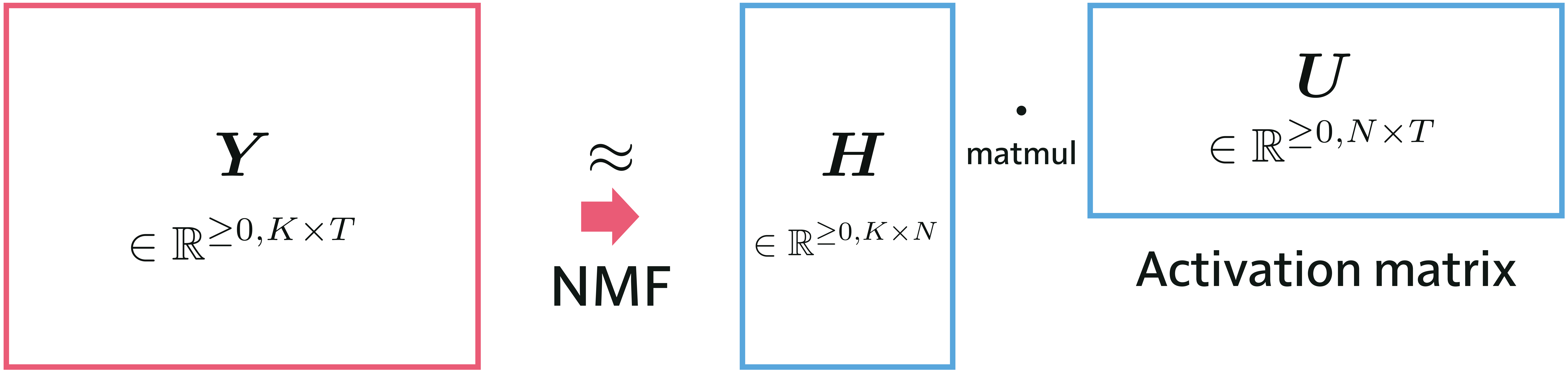
CycleGAN-VC
[Kaneko+, 2017]

Outline

- **Baseline 1: NMF-based voice conversion**
- Baseline 2: INCA algorithm
- Proposed: Nonparallel training of NMF-based voice conversion
- Experiments

Non-negative matrix factorization (NMF)

[D. D. Lee+, 1999]



$\boldsymbol{Y}$

$\in \mathbb{R}^{\geq 0, K \times T}$

Given non-negative matrix

$\approx$

NMF

$\boldsymbol{H}$

$\in \mathbb{R}^{\geq 0, K \times N}$

Base matrix

(Exemplars, Dictionary)

$\cdot$

matmul

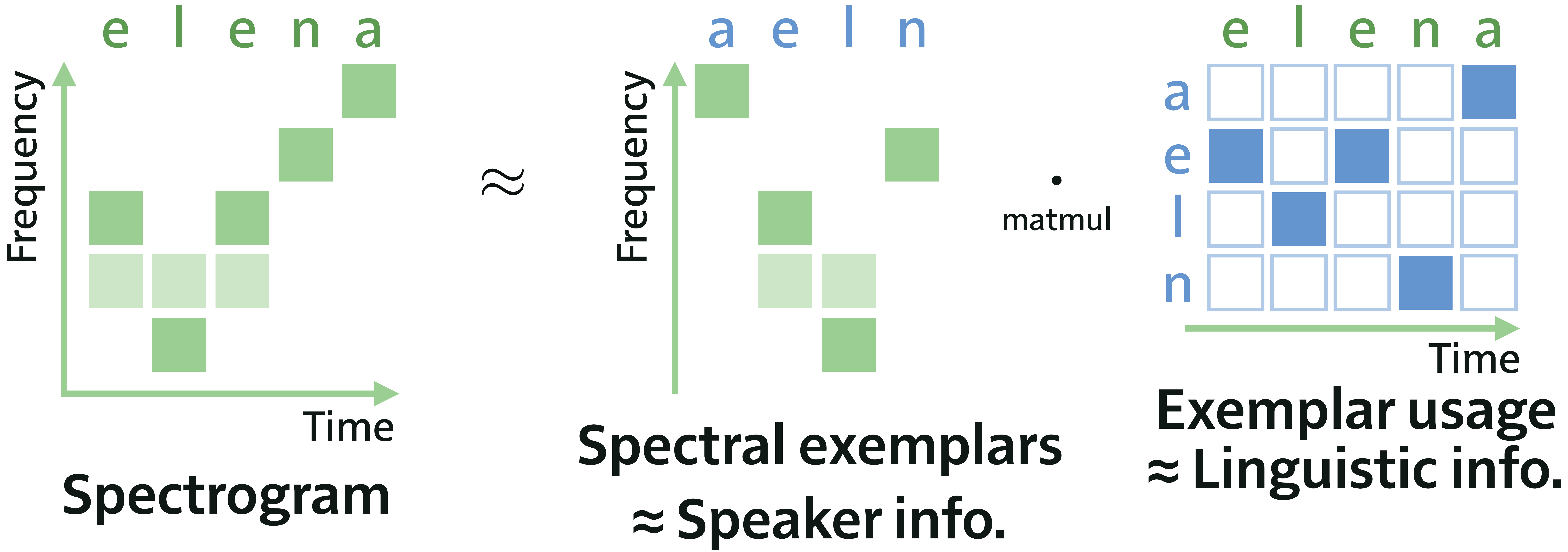
$\boldsymbol{U}$

$\in \mathbb{R}^{\geq 0, N \times T}$

Activation matrix

- NMF acquires $\boldsymbol{H}$ and $\boldsymbol{U}$ by minimizing divergence $\mathcal{D}(\boldsymbol{Y} | \boldsymbol{H}\boldsymbol{U})$

Non-negative matrix factorization (NMF)



	e	l	e	n	a
a	□	□	□	□	■
e	■	□	■	□	□
l	□	■	□	□	□
n	□	□	□	■	□

Spectrogram

Spectral exemplars ≈ Speaker info.

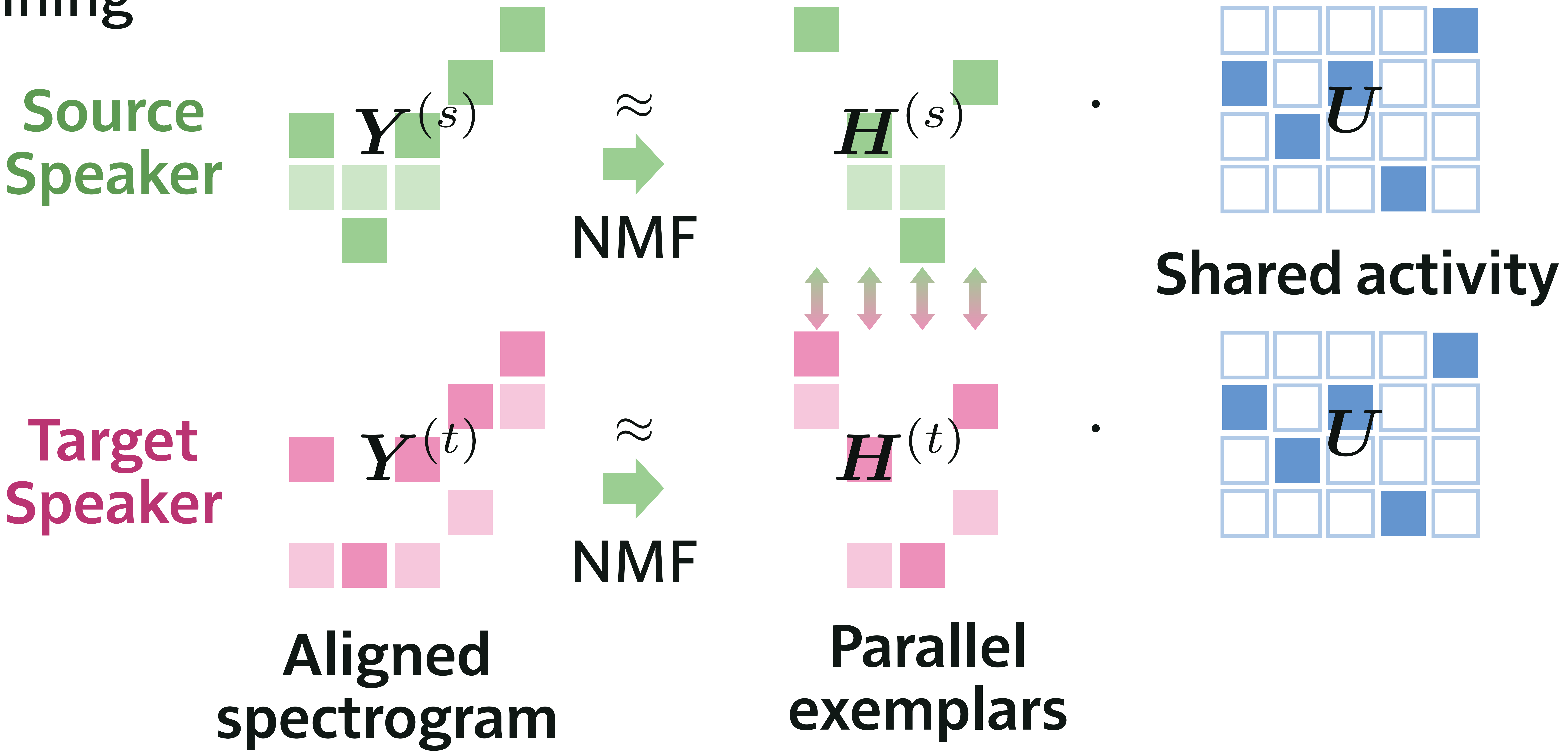
Exemplar usage ≈ Linguistic info.

- Supposing input matrix is spectrogram, **exemplars** contain **individuality** and **activity** contains **linguistic** information

NMF-based Voice Conversion (NMF-VC)

[R. Takashima+, 2012]

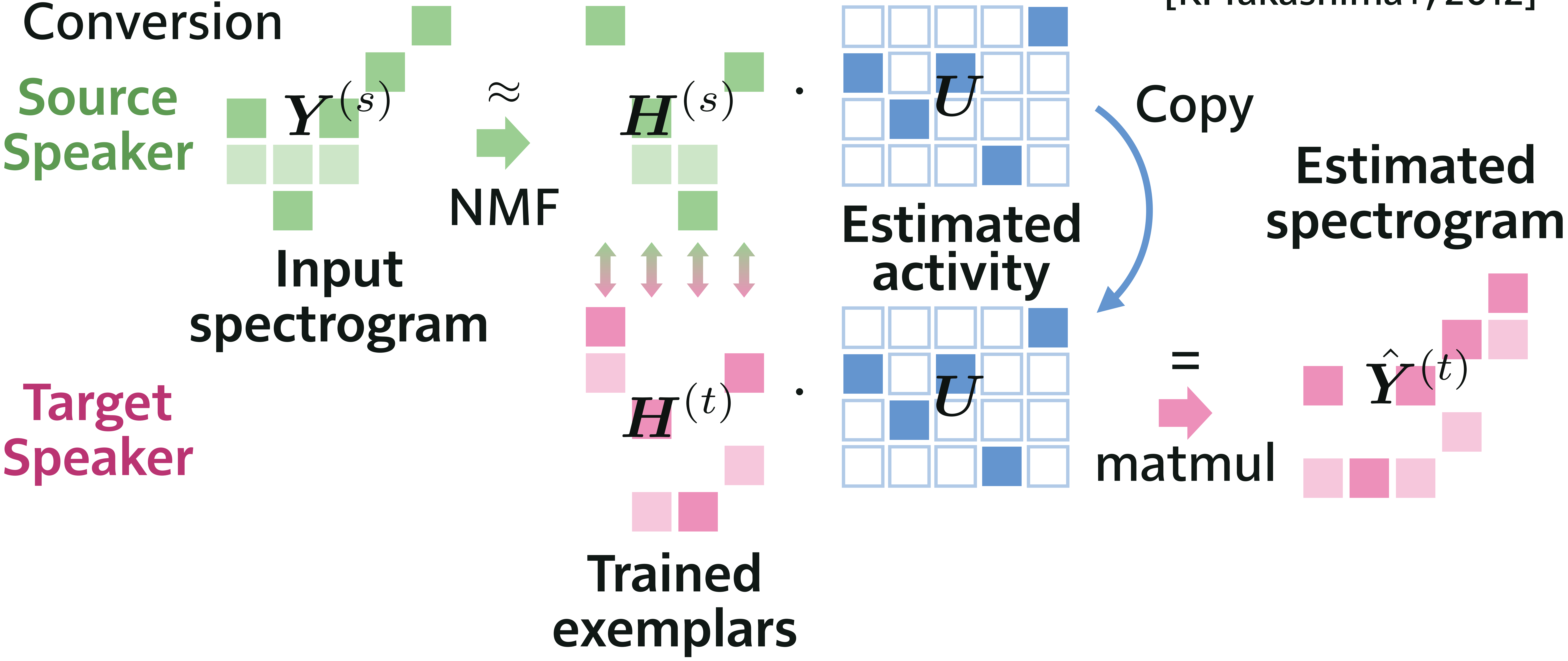
Training



NMF-based Voice Conversion (NMF-VC)

[R. Takashima+, 2012]

Conversion



NMF-based Voice Conversion (NMF-VC)

[R. Takashima+, 2012]

- Uses **parallel exemplars as a conversion model**
  - NMF-VC is called “exemplar-based VC”

✓ **Fast on GPU**, because NMF uses only basic matrix operations

✗ Requires parallel corpora

✗ Degrades because of activation mismatch

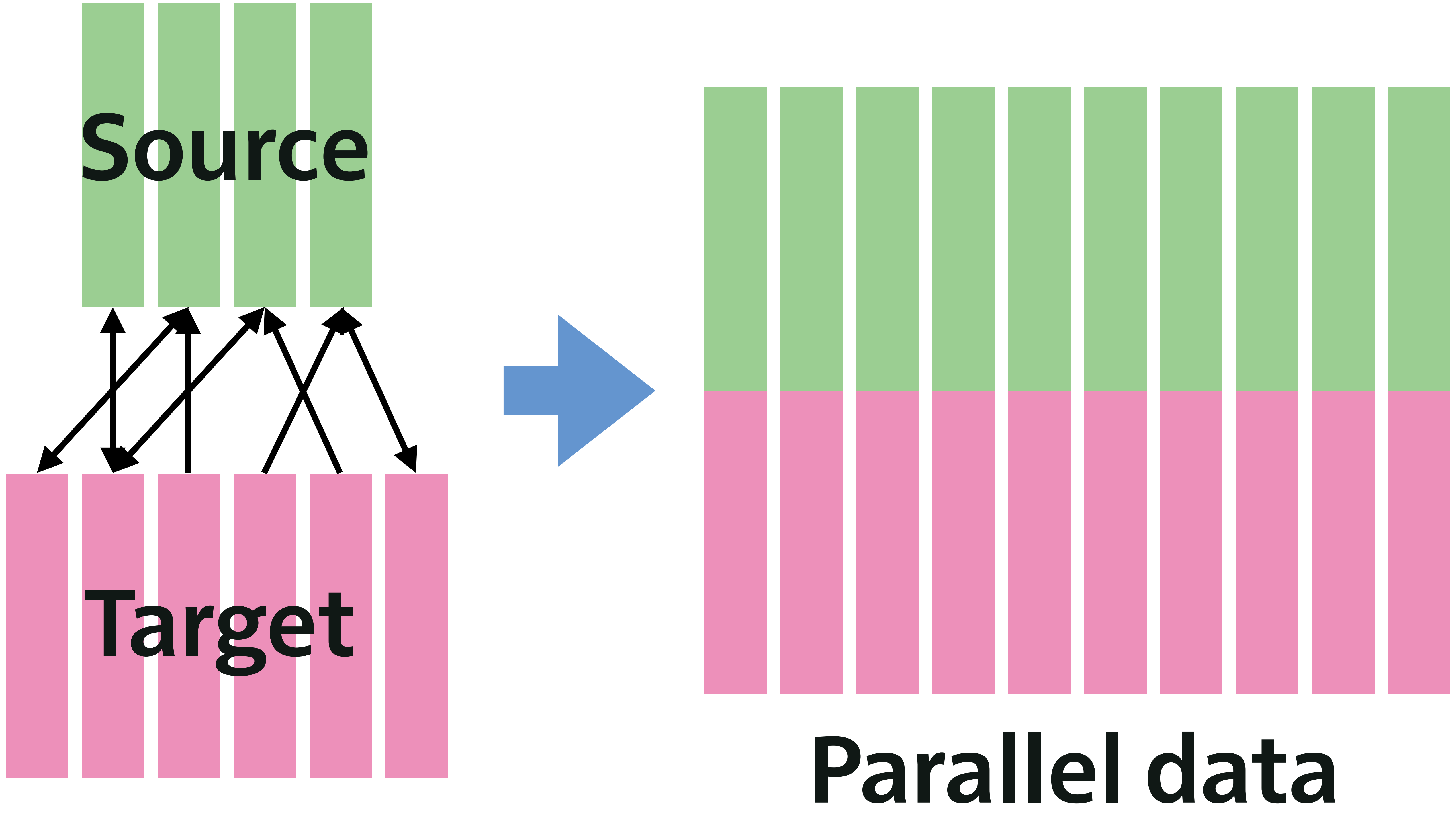
Outline

- Baseline 1: NMF-based voice conversion
- **Baseline 2: INCA algorithm**
- Proposed: Nonparallel training of NMF-based voice conversion
- Experiments

INCA algorithm

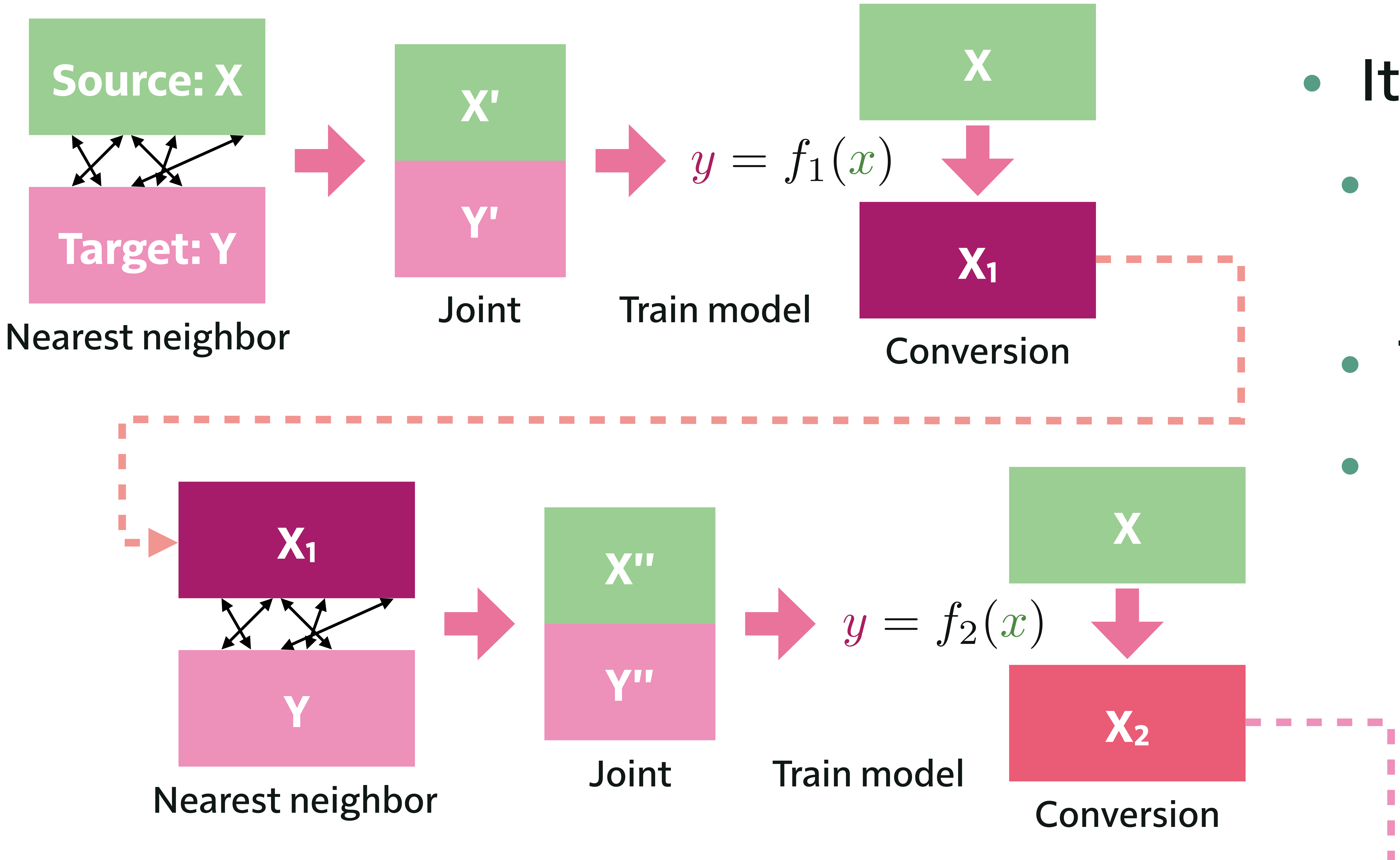
[D. Erro+, 2010]

- **I**terative combination of a **N**earest neighbor search step and a **C**onversion step **A**lignment method
- Algorithm to acquire **alignment of nonparallel corpora**



INCA algorithm

[D. Erro+, 2010]



- Iteration of
  - Nearest neighbor
  - Train model
  - Conversion

INCA algorithm

✓ Applicable to any parallel VC frameworks

✓ Easy to implement

✗ Requires as much data as parallel VC systems

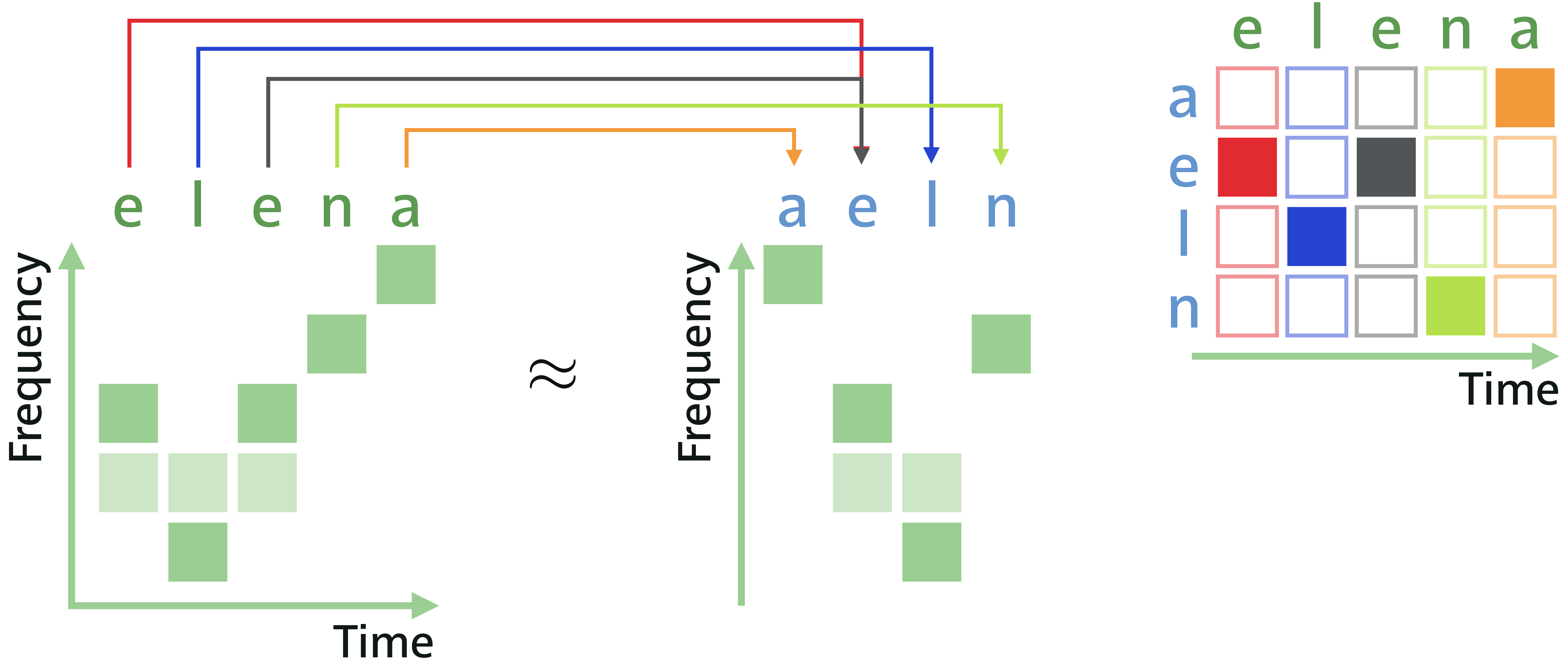
- VC system does not require much data for source speaker
  - **VC system is a generator of target speaker**

Outline

- Baseline 1: NMF-based voice conversion
- Baseline 2: INCA algorithm
- **Proposed: Nonparallel training of NMF-based voice conversion**
- Experiments

Revisit to NMF-VC

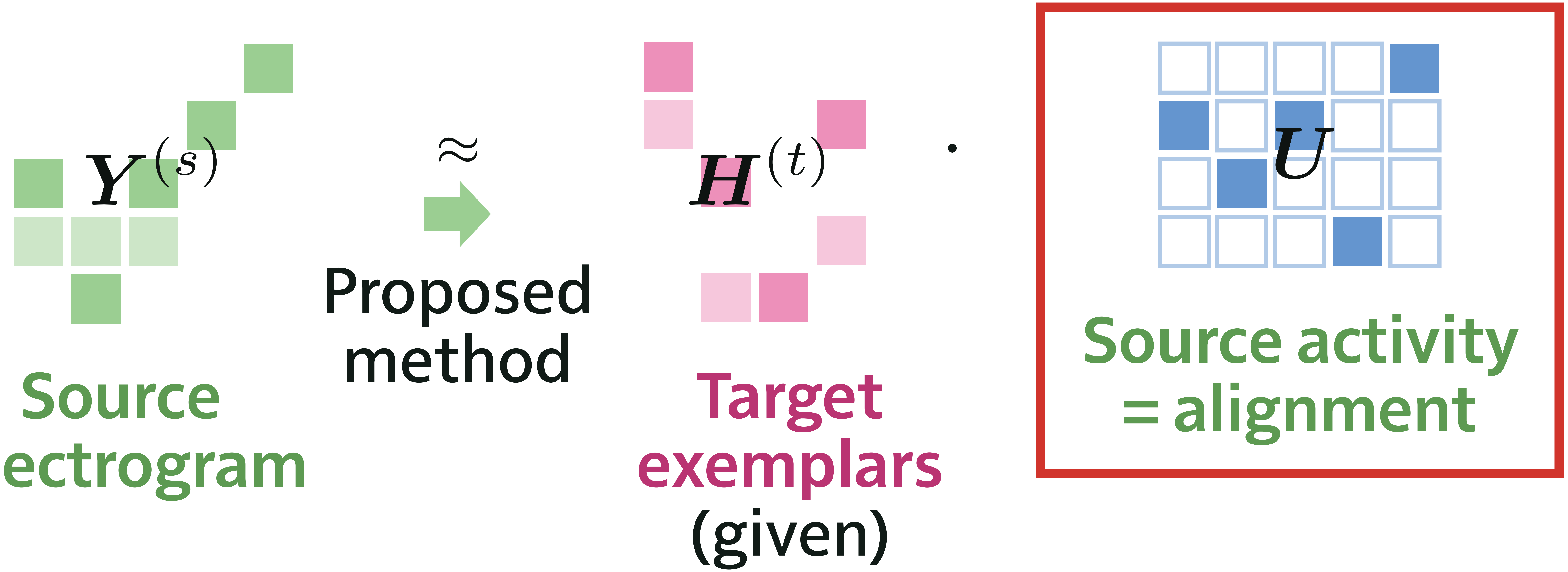
- NMF is equivalent to **alignment from spectrogram to exemplars**



	e	l	e	n	a
a	□	□	□	□	■
e	■	□	■	□	□
l	□	■	□	□	□
n	□	□	□	■	□

Proposed training method of NMF-VC

- Acquires activation of NMF-VC using INCA technique



$\boldsymbol{Y}^{(s)} \approx \boldsymbol{H}^{(t)} \cdot \boldsymbol{U}$

Source spectrogram

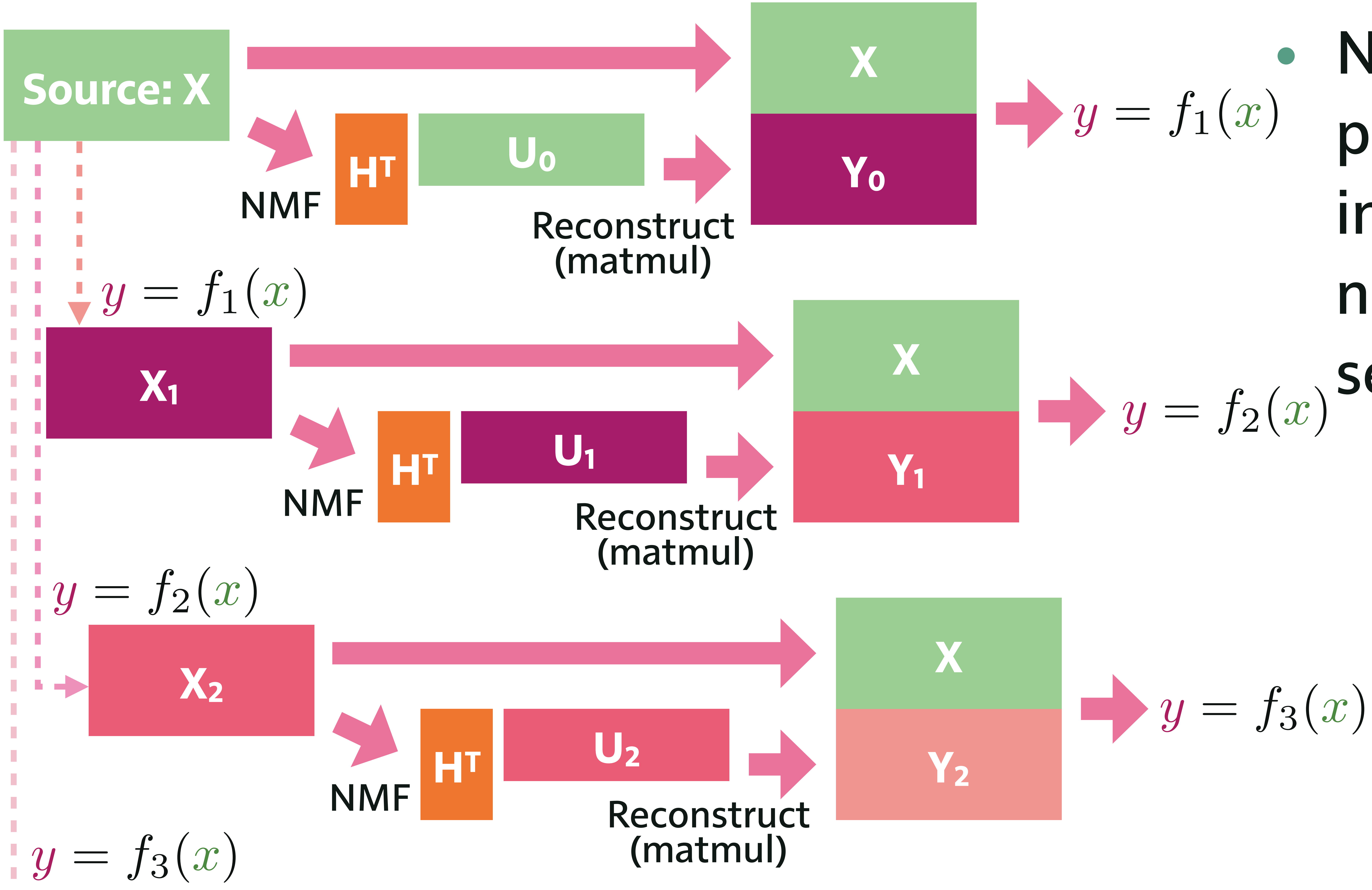
Proposed method

Target exemplars (given)

Source activity = alignment

- Activity is irrelevant to speaker of exemplars

Proposed training method of NMF-VC



- NMF is performed instead of nearest neighbor search

Outline

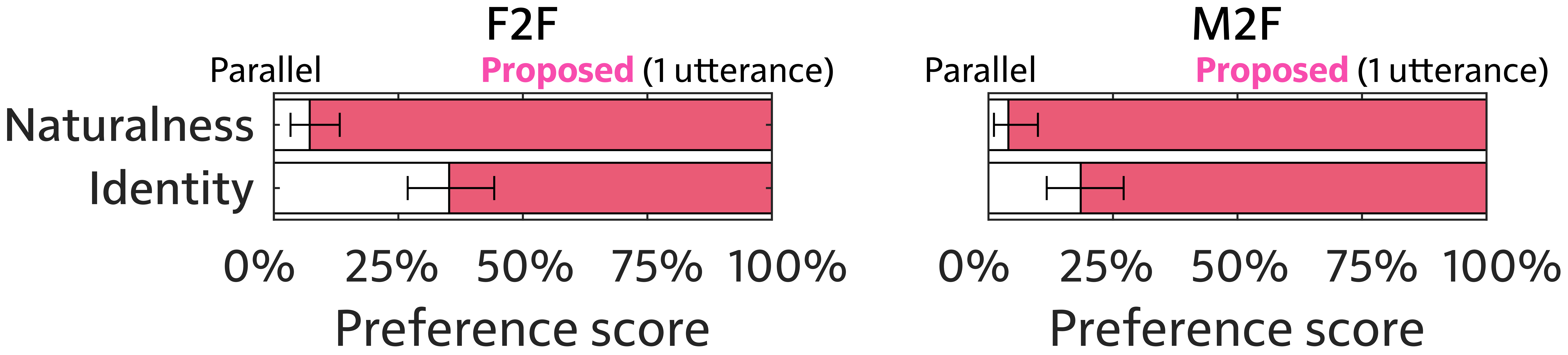
- Baseline 1: NMF-based voice conversion
- Baseline 2: INCA algorithm
- Proposed: Nonparallel training of NMF-based voice conversion
- **Experiments**

Experimental setups

- Methods for comparison
  - CycleGAN-VC [T. Kaneko+, 2019], Parallel NMF-VC [R. Takashima+, 2012]
- Dataset: Japanese versatile speech corpus [S. Takamichi+, 2019]
- Speaker pair: JVS066→JVS010 (F2F), JVS054→JVS010 (M2F)
- Number of sentences: 60 for target speaker, 1 or 10 for source speakers, 20 for test
  - Same 60 sentences are used for source speakers in parallel NMF-VC
- Analysis / Synthesis: WORLD [M. Morise+, 2016]
- Target of decomposition: Amplitude spectrograms of WORLD spectral envelopes
- Dictionary size (number of exemplars): 200
- Conditions of subjective experiments (≥ 25 subjects)
  - A/B preference tests for naturalness
  - ABX tests for speaker similarity

Results of subjective experiments

- Comparison with parallel NMF-VC

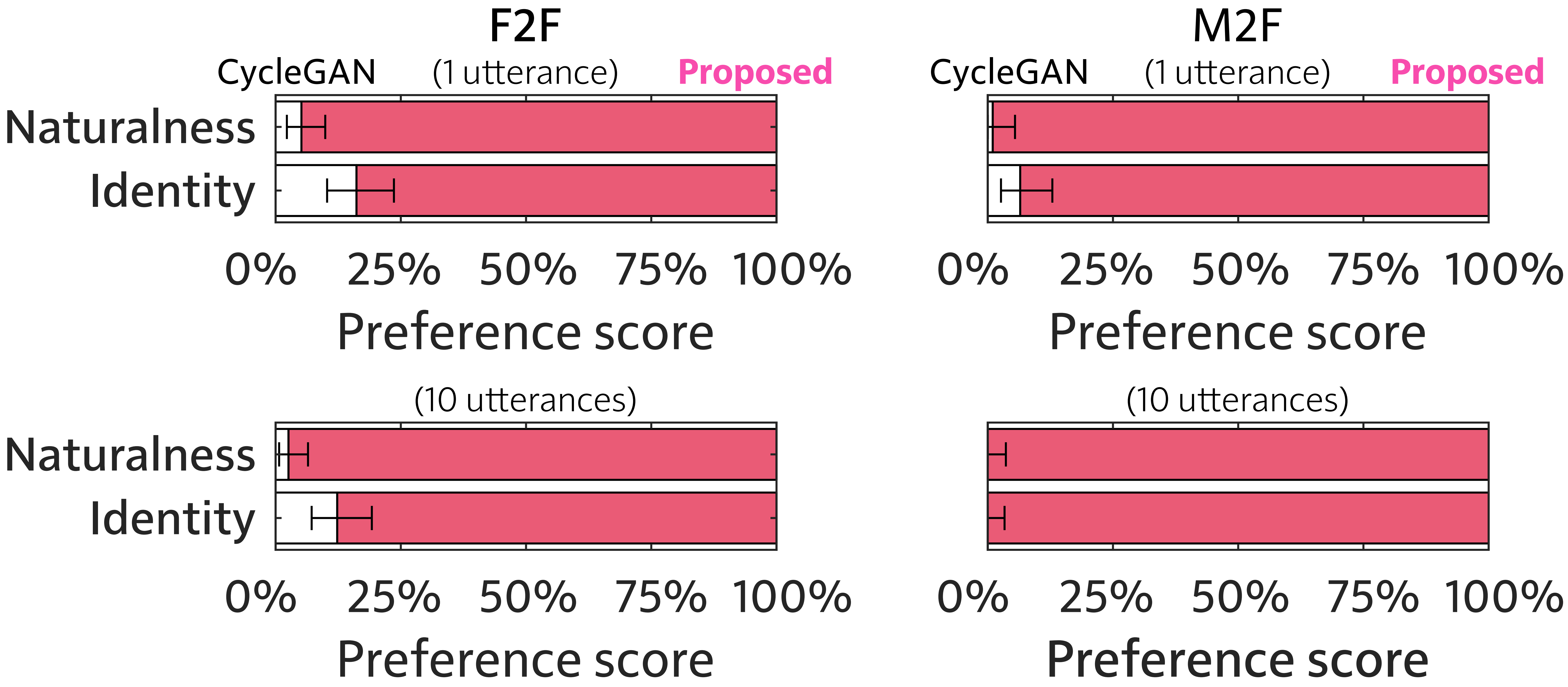


- Proposed framework outperformed conventional parallel NMF-VC framework

⊢⊣ : 95% confidential

Results of subjective experiments

- Comparison with CycleGAN-VC

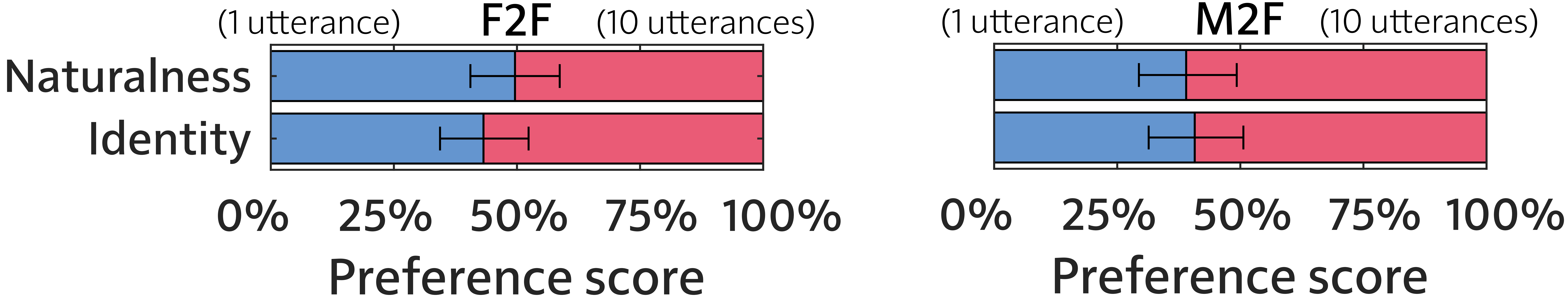


- Proposed framework efficiently used a small source speakers' data

⊢⊣ : 95% confidential

Results of subjective experiments

- Effects of an amount of source speakers' data



- More **source** speakers' data provided
  more speaker similarity about **target** speaker

Audio examples

		Parallel NMF-VC	CycleGAN-VC	Proposed
Target (Female)		🔊		
F2F	Source	🔊		
	Converted	🔊	🔊	🔊
M2F	Source	🔊		
	Converted	🔊	🔊	🔊

- 1 sentence is used for training each source speaker
- Examples are available online (see last slide)

Conclusions

- We introduce **nonparallel training method of NMF-VC** inspired by the INCA algorithm
- The method achieved the goal with the small amount of the training data for source speakers

Future Works

- Higher quality conversion
- Inter-language conversion

https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/nmfvc2020interspeech/