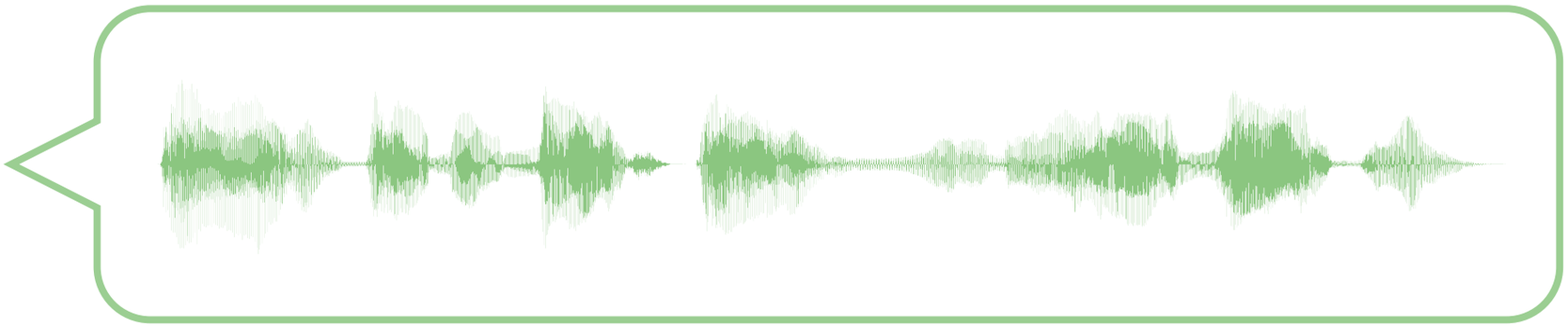
A Revisit to Feature Handling for High-quality Voice Conversion Based on Gaussian Mixture Model

Hitoshi Suda, Gaku Kotani,
Shinnosuke Takamichi, and Daisuke Saito
(The University of Tokyo)

APSIPA ASC 2018 / Nov. 14, 2018 No. 1370 / WE-A2-8.3

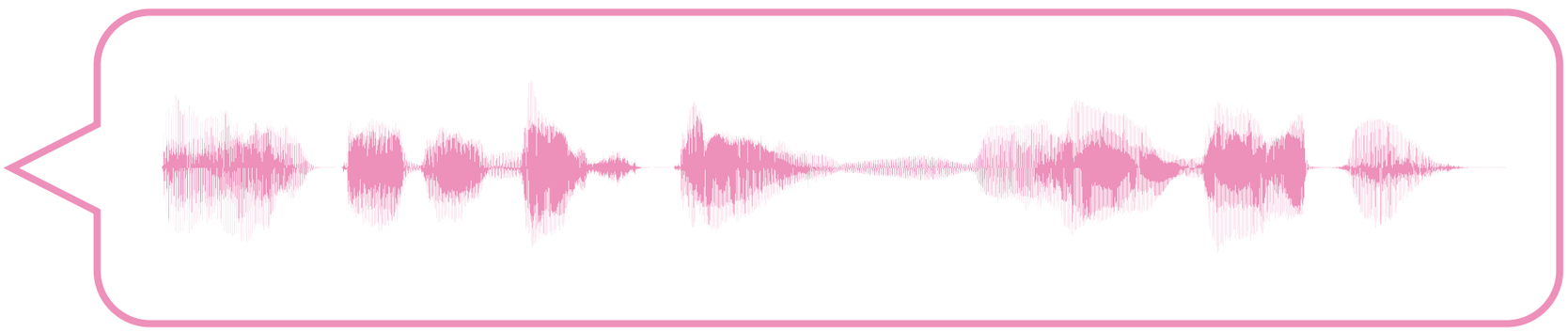
Voice Conversion

Source Speaker



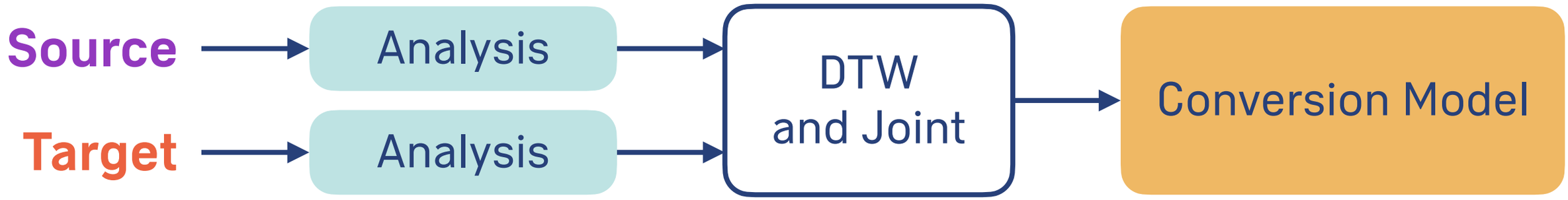
Modifying personalities in the utterance

Target Speaker

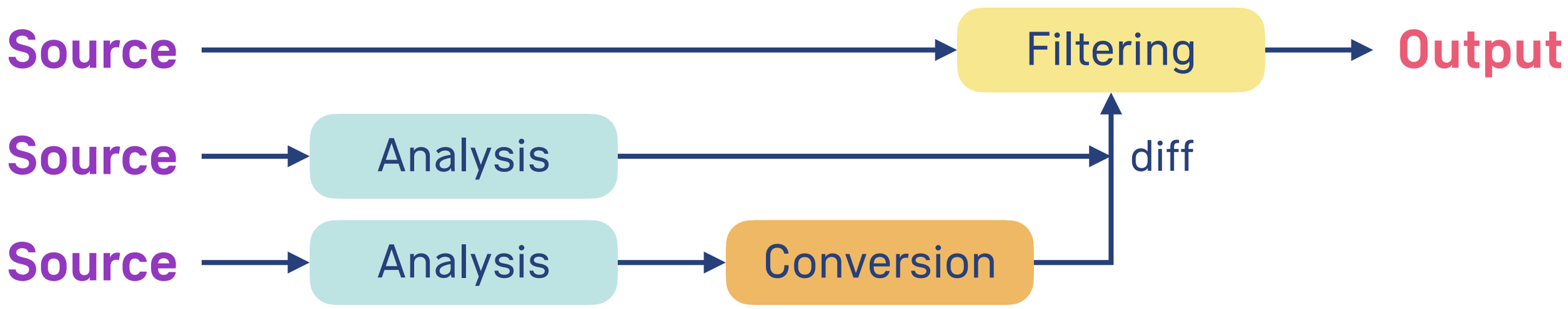


Voice Conversion Framework

- Training

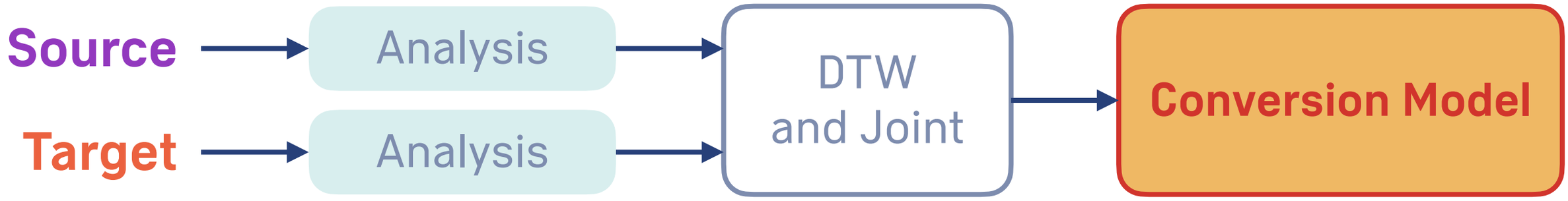


- Conversion

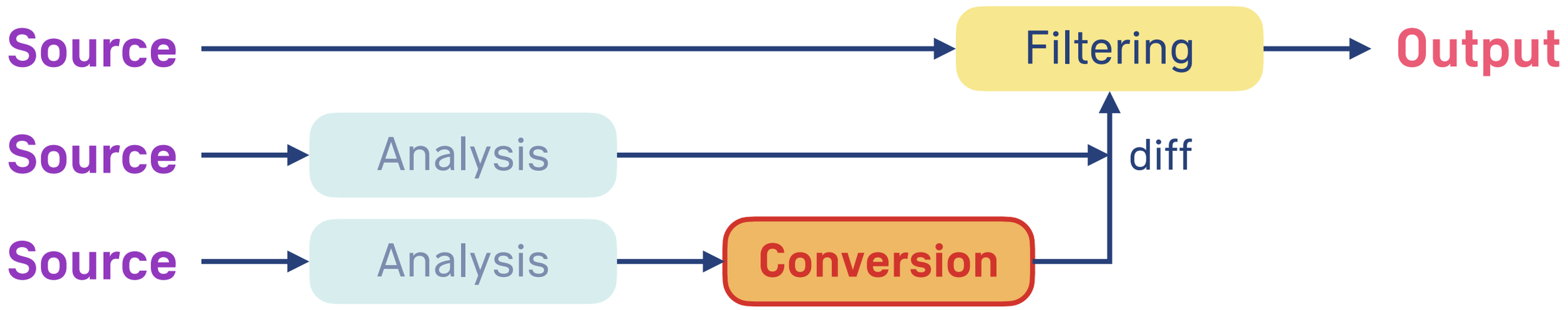


Voice Conversion Framework

- Training

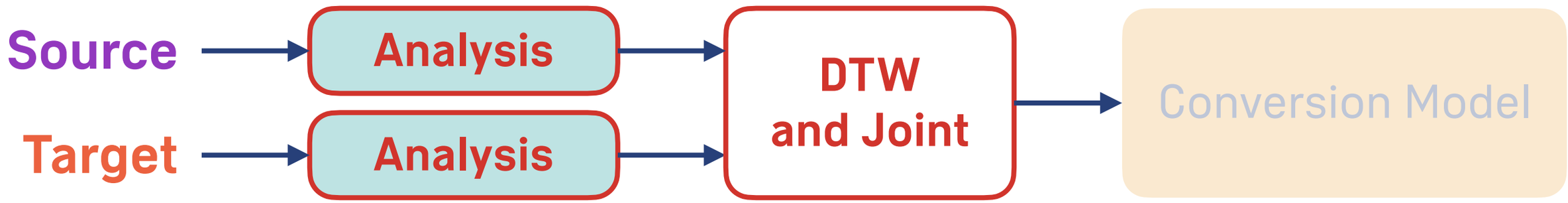


- Conversion

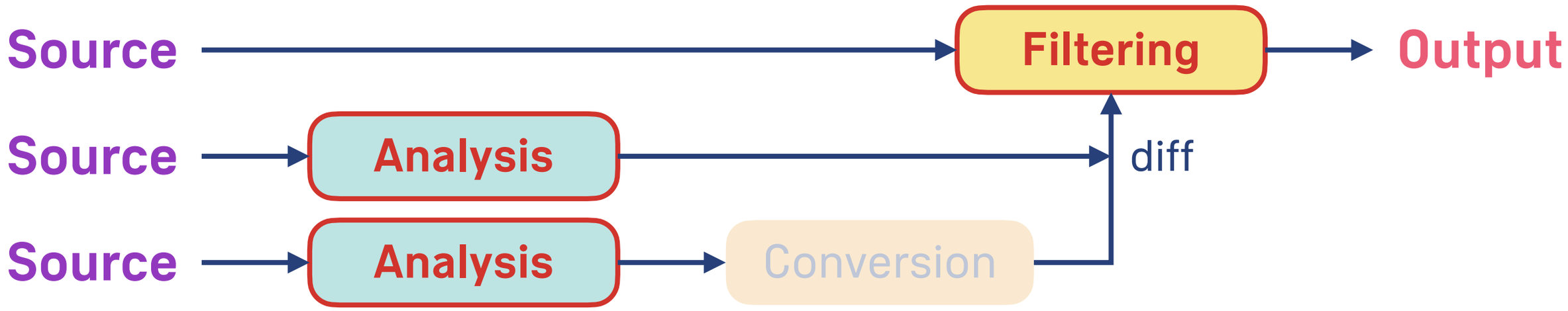


Voice Conversion Framework

- Training



- Conversion



Aim of This Study

A Revisit to Feature Handling for High-quality Voice Conversion Based on Gaussian Mixture Model

- To **improve conversion quality** of existing voice conversion frameworks
- To experimentally reveal influences of feature handling
  - via subjective experiments

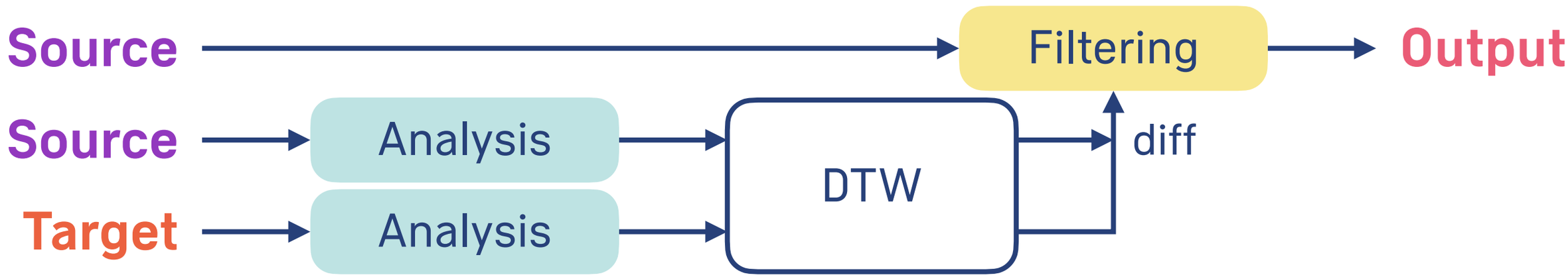
Experimental Setups

- 50 sentences from ATR Japanese phonetically balanced sentence sets [Kurematsu+, 1990]
  - 40 for training, 10 for evaluation
- Sampling frequency: 22050 Hz
- Analysis and synthesis: WORLD [Morise+, 2016]
- Speakers: 2 males and 2 females
- Only intra-gender conversion / No F_0 conversion
- 23 listeners answered questions in each preference test via our crowdsourcing system

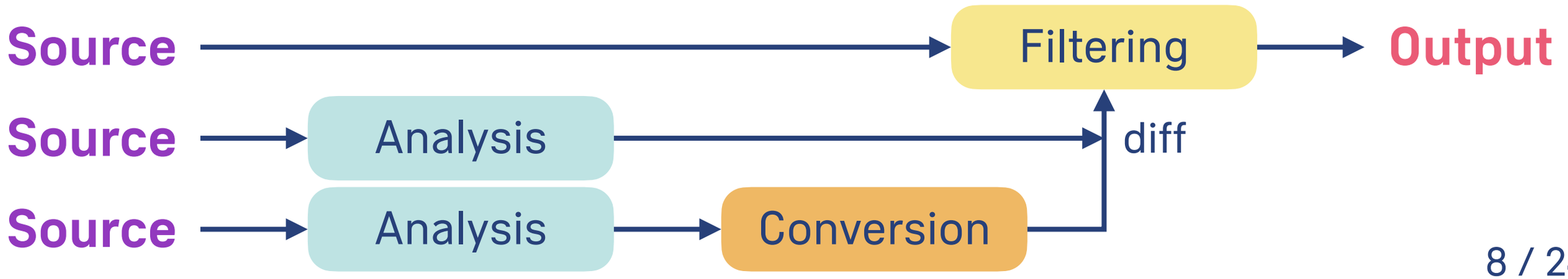
3 Experiments

1. Analysis conditions

2. Conversion system without statistical mapping



3. Total conversion system

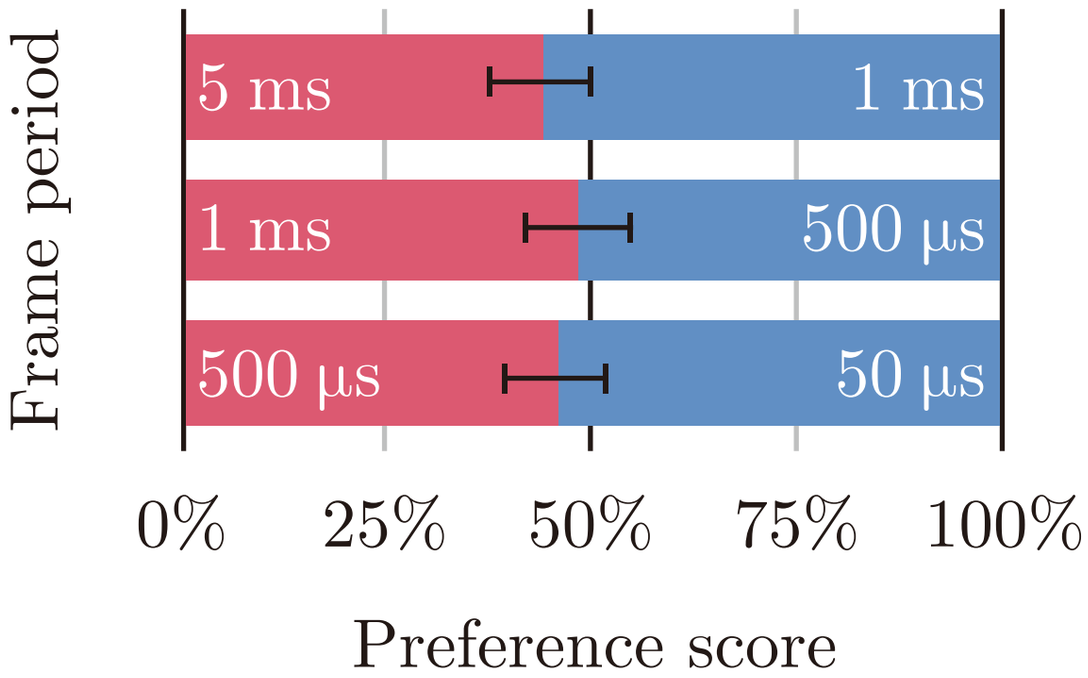


Experiment 1: Analysis Conditions

- To reveal the effects of conditions of analysis
  - Frame periods (or frame shifts)
    - How precisely the waveforms are analyzed in time domain
  - Order of mel-cepstral coefficients (mcep)
    - How precisely the spectral envelopes are represented

Experiment 1: Analysis Conditions

- Frame periods:



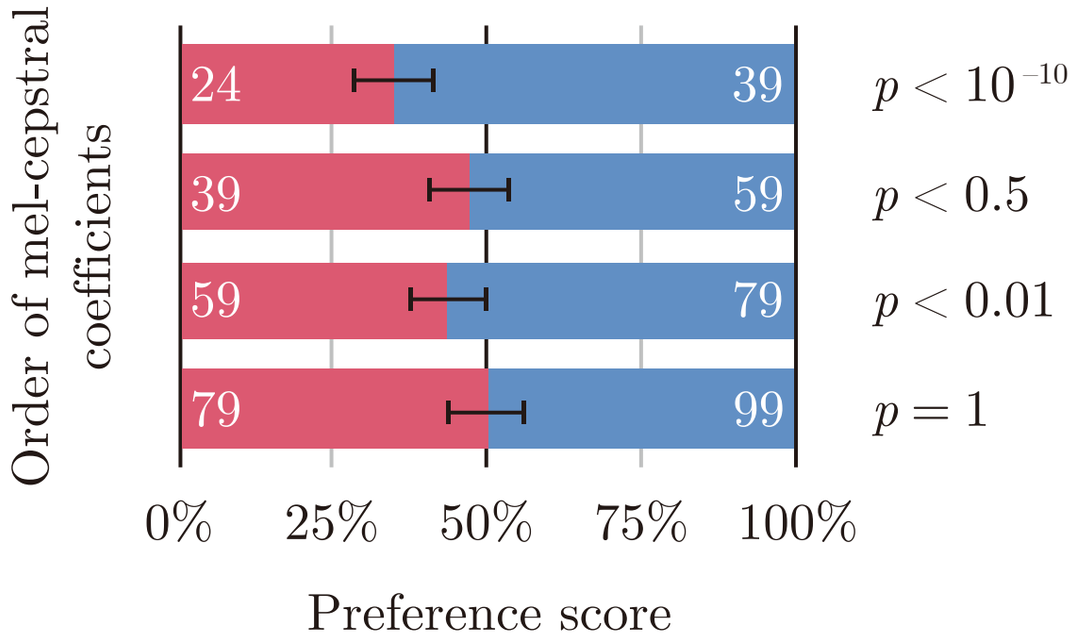
$p < 0.01$ 5 ms < 1 ms

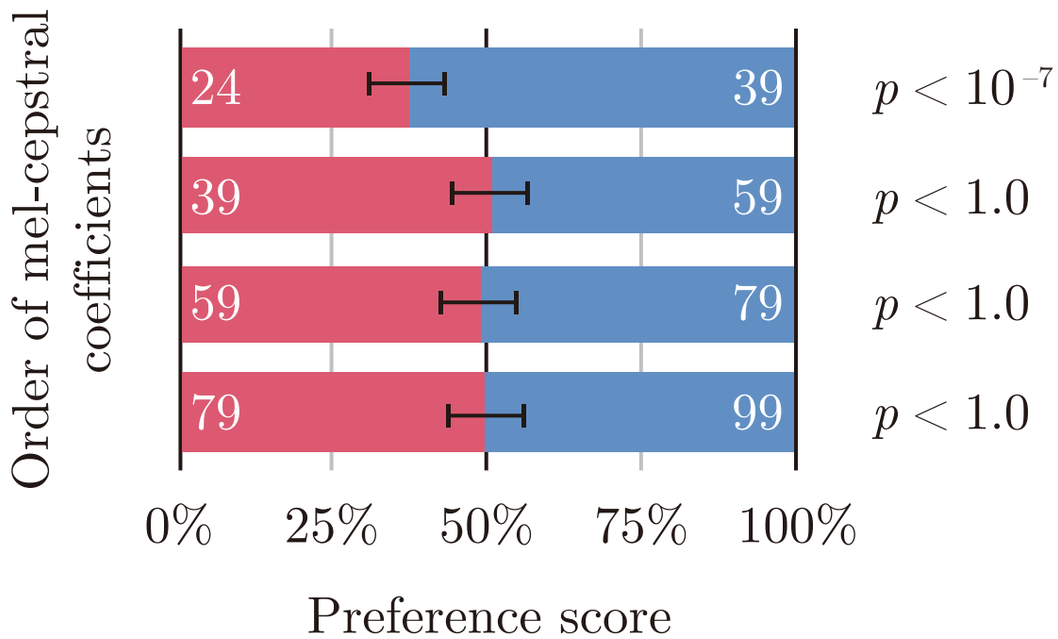
$p < 0.5$ 1 ms ≈ 500 µs

$p < 0.1$ 500 µs ≤ 50 µs

Experiment 1: Analysis Conditions

- Order of mcep:





with 1 ms analysis

24 < 39

39 ≈ 59

59 < 79

79 ≈ 99

with 50 μs analysis

24 < 39

39 ≈ 59

59 ≈ 79

79 ≈ 99

Experiment 1: Analysis Conditions

- Frame periods:

  5 ms < **1 ms** ≈ 500 μs ≤ 50 μs

- Order of mcep:

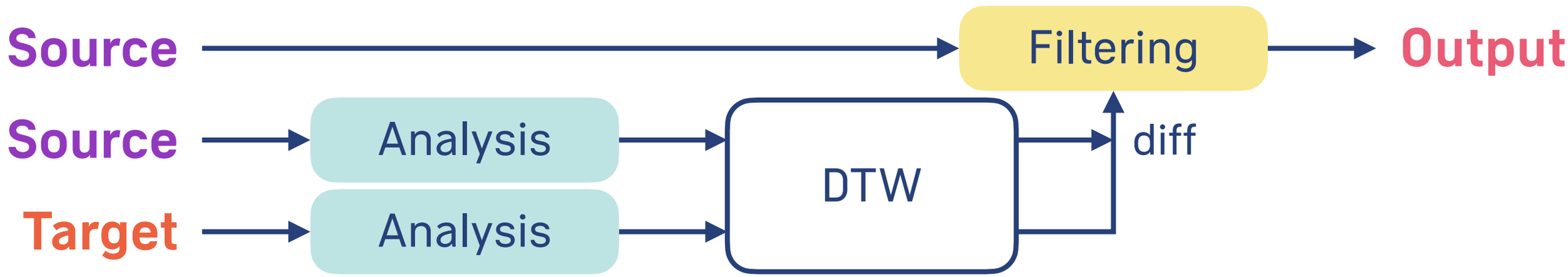
  24 < 39 ≈ 59 < **79** ≈ 99　(with 1 ms frames)

  24 < **39** ≈ 59 ≈ 79 ≈ 99　(with 50 μs frames)

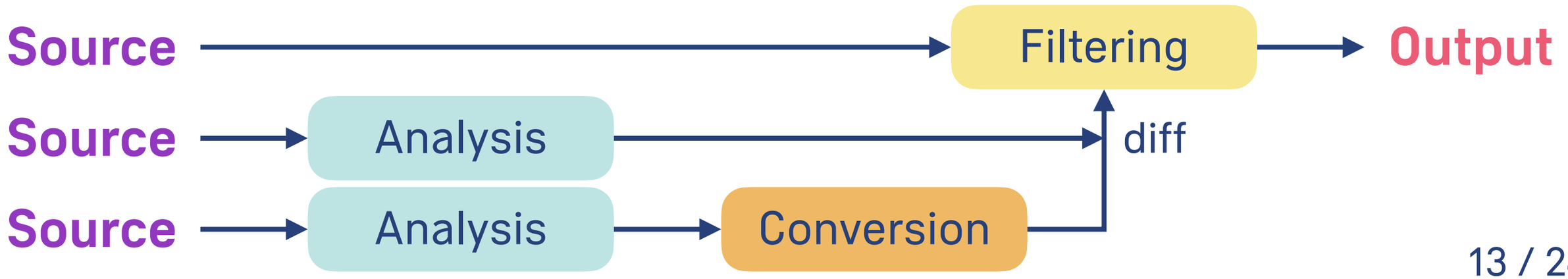
3 Experiments

1. Analysis conditions

2. **Conversion system without statistical mapping**

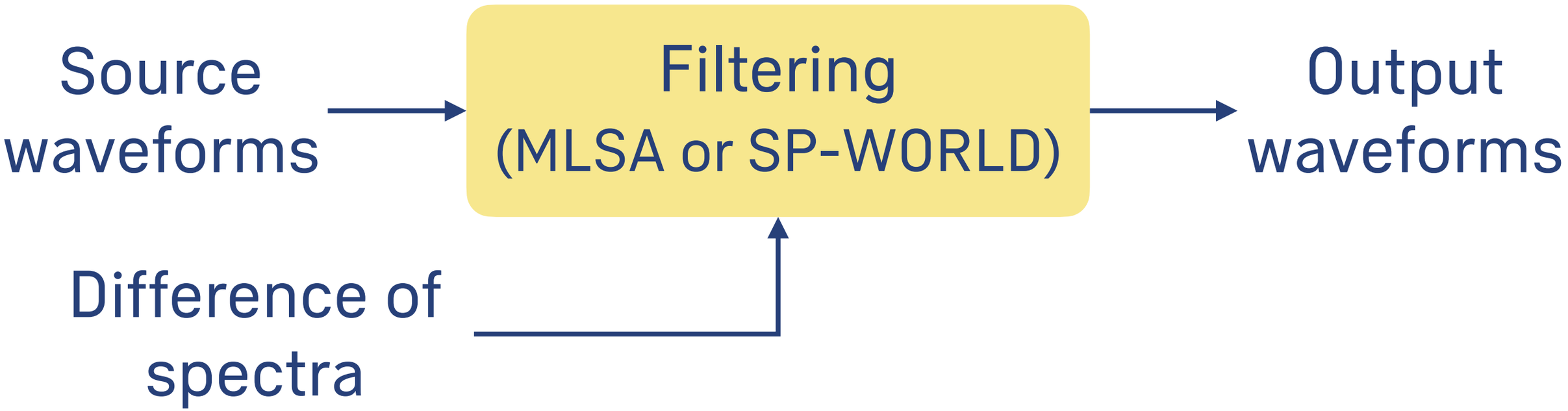


3. Total conversion system



Differential-spectrum Compensation

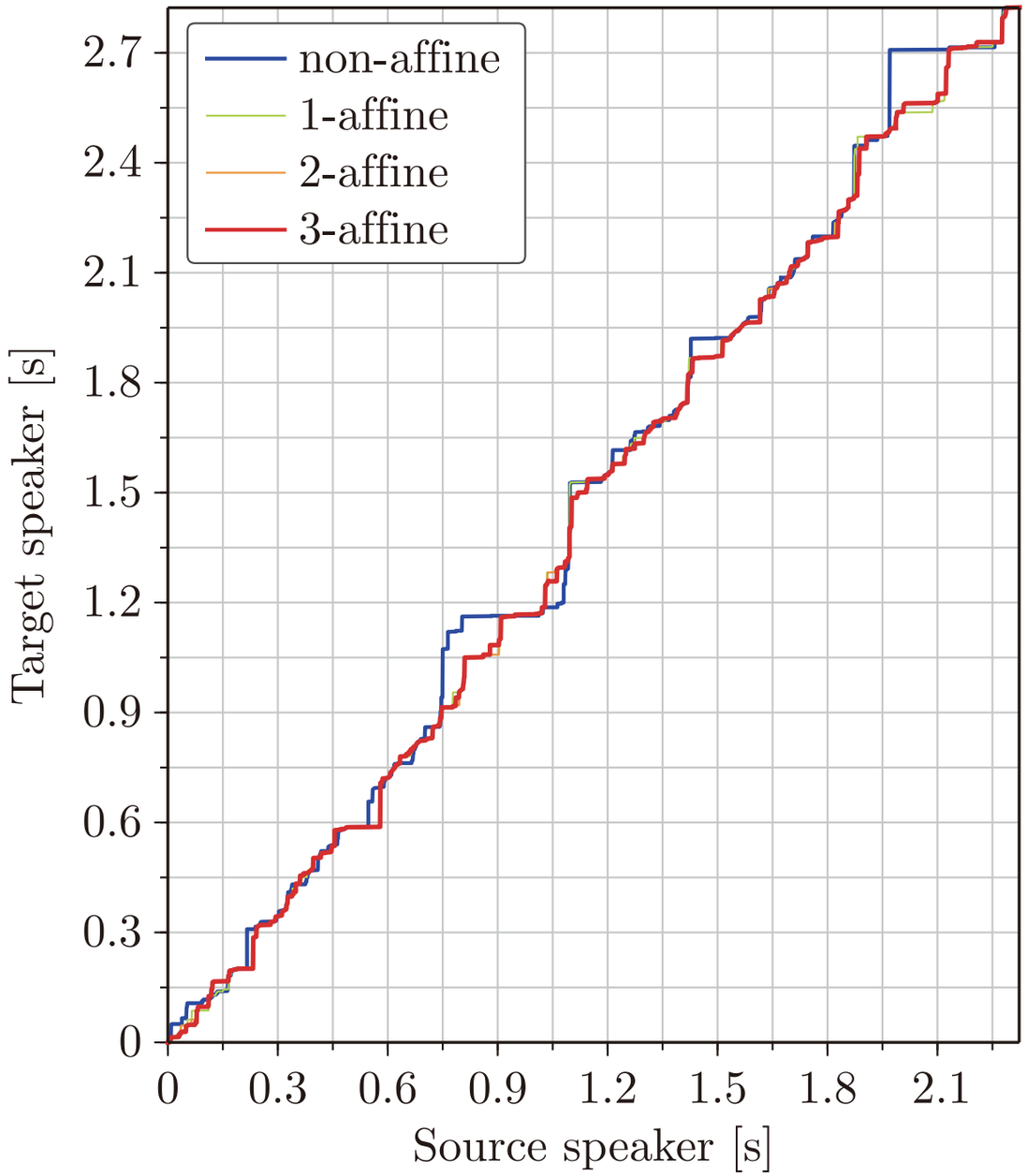
(diffspec) [Kobayashi+, 2014]



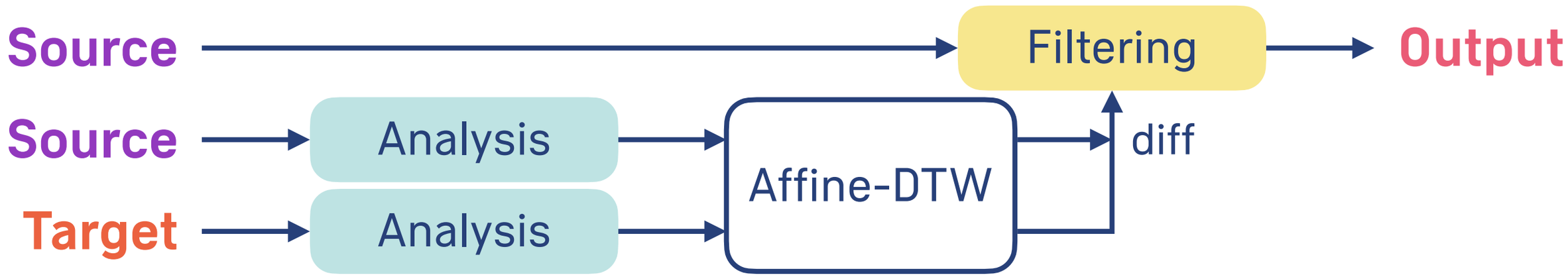
- Famous implementation: Mel log spectrum approximation (MLSA) Filtering [Imai+, 1983]
- We introduce a diffspec method "**SP-WORLD**" inspired by WORLD vocoder

Dynamic Time Warping (DTW)

- Alignment of features
- Sensitive to difference of individuality
- We introduce “Affine-DTW”
  - Iteration of alignment and coarse conversion



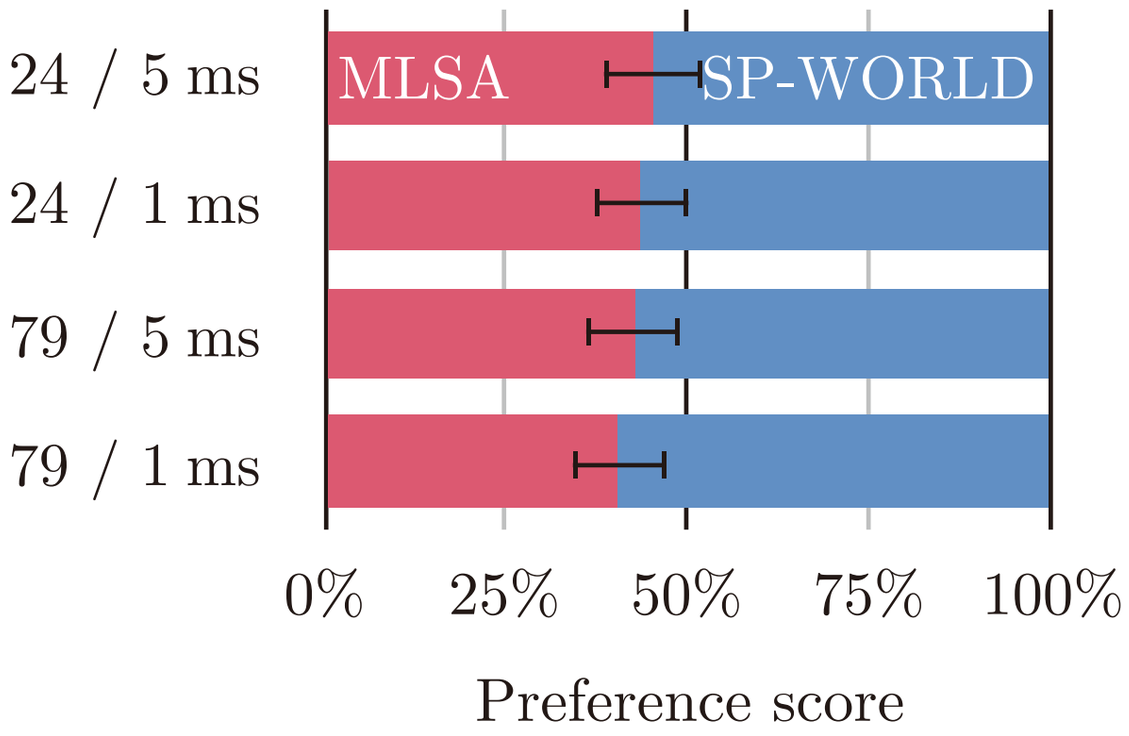
Experiment 2: Ideal Conversion



- To reveal the effects on quality of conversion of the conditions except mapping models
- Diffspec method: MLSA or SP-WORLD
- Analysis conditions
  - Frame period and order of mcep

Experiment 2: Ideal Conversion

- Diffspec method:



$p < 0.05$ MLSA ≦ SP-WORLD

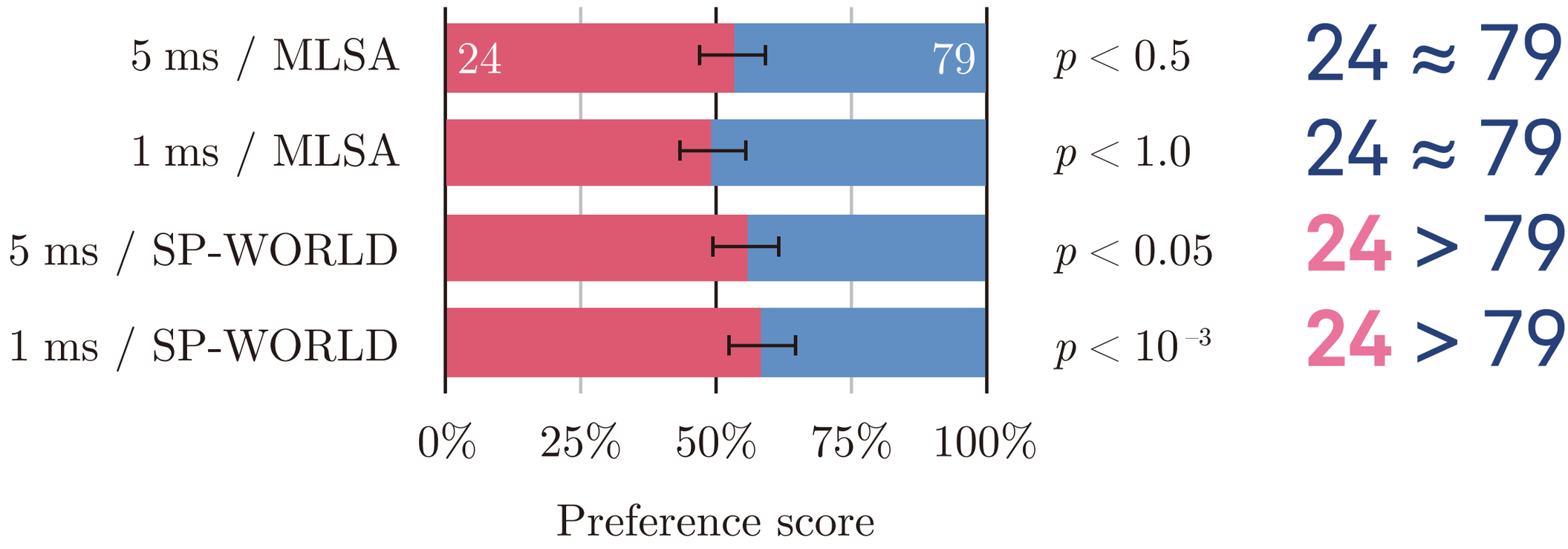
$p < 0.01$ MLSA < SP-WORLD

$p < 10^{-3}$ MLSA < SP-WORLD

$p < 10^{-4}$ MLSA < SP-WORLD

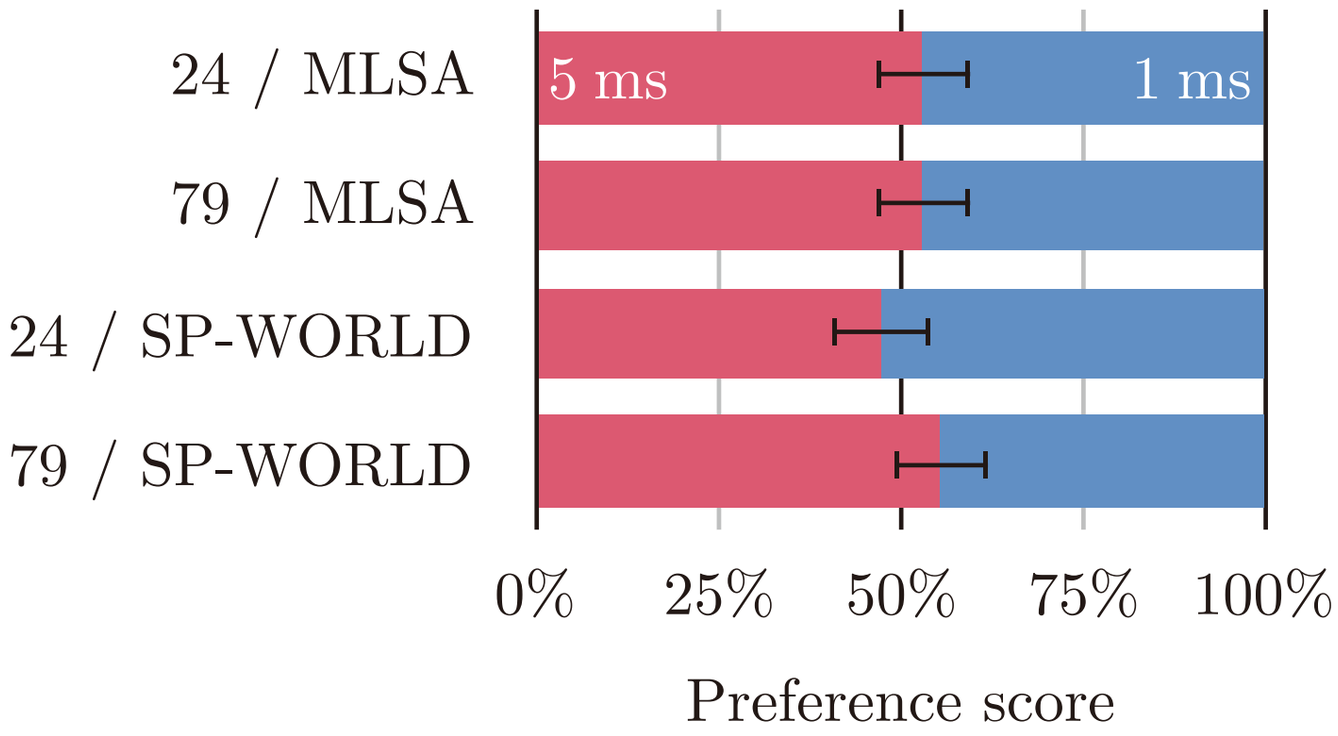
Experiment 2: Ideal Conversion

- Order of mcep:



Experiment 2: Ideal Conversion

- Frame periods:



$p < 0.5$

$p < 0.5$

$p < 0.5$

$p < 0.05$

5 ms ≈ 1 ms

5 ms ≈ 1 ms

5 ms ≈ 1 ms

5 ms ≧ 1 ms

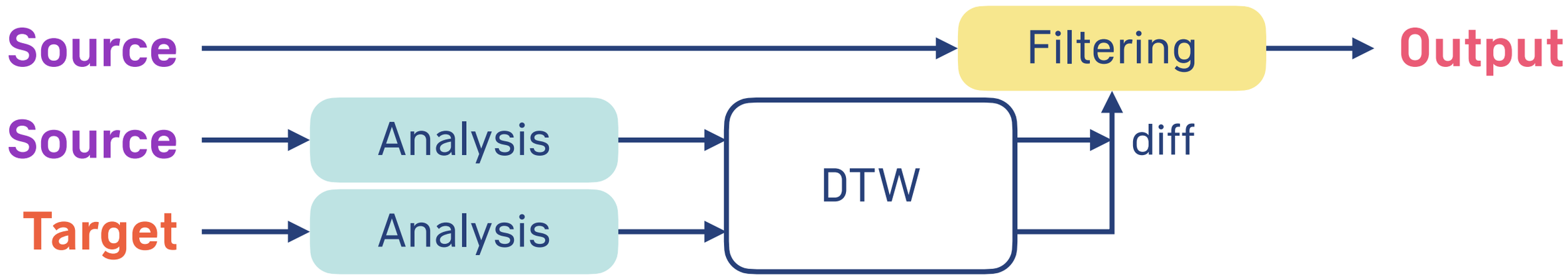
Experiment 2: Ideal Conversion

- Diffspec method: MLSA < **SP-WORLD**
- Order of mcep: **24** > 79 (with SP-WORLD)
  24 ≈ 79 (with MLSA)
- Frame periods:
  5 ms ≧ 1 ms (with SP-WORLD / 79-order)
  5 ms ≈ 1 ms (with other conditions)

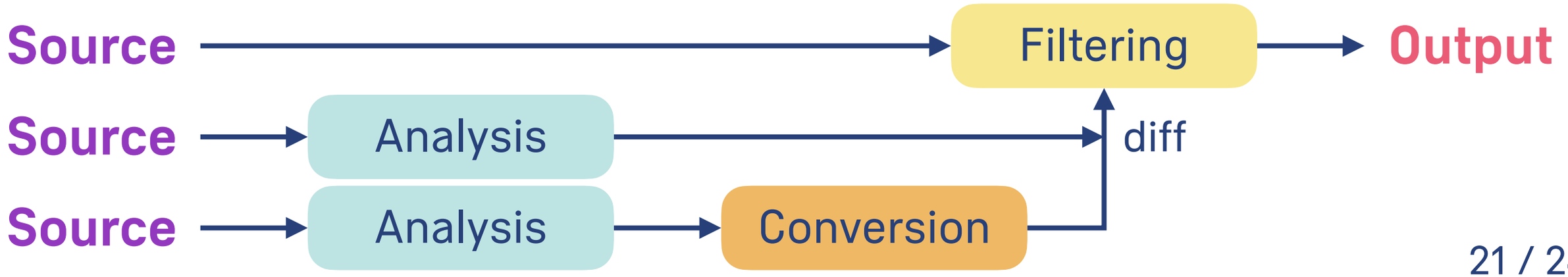
3 Experiments

1. Analysis conditions

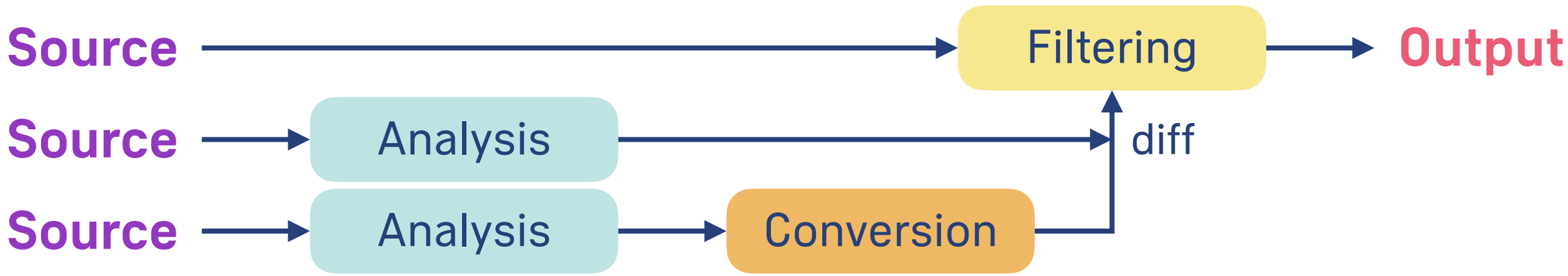
2. Conversion system without statistical mapping



3. Total conversion system



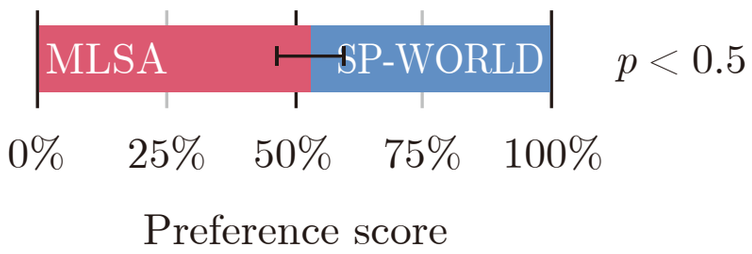
Experiment 3: Statistical Conversion



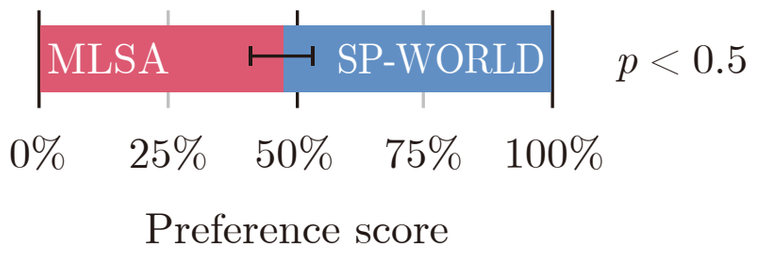
- To reveal the influences of below components
  - Diffspec method: MLSA or SP-WORLD
  - Sequence features [Toda+, 2007]
    - Dynamic features and global variances
- 1 ms period / 24-order of mcep

Experiment 3: Statistical Conversion

- Diffspec method: MLSA ≈ SP-WORLD



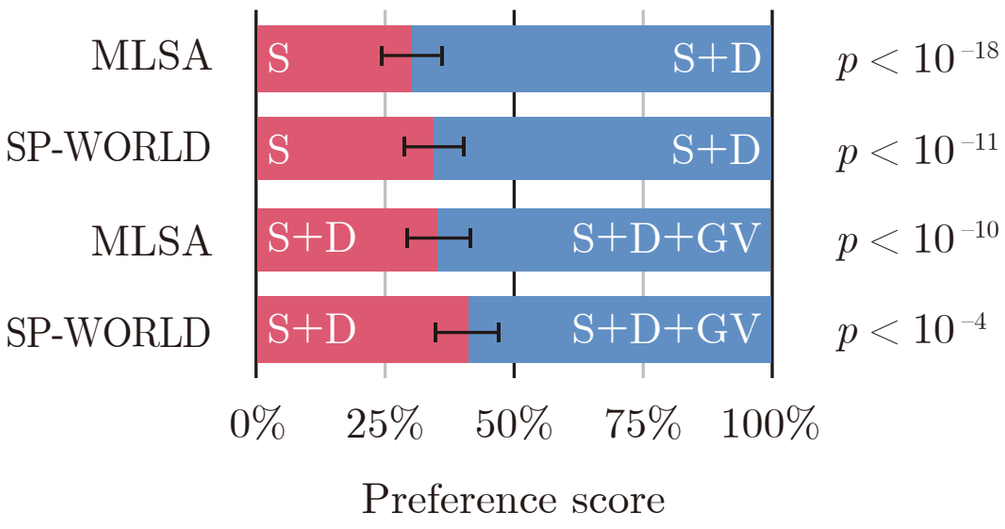
speaker similarity



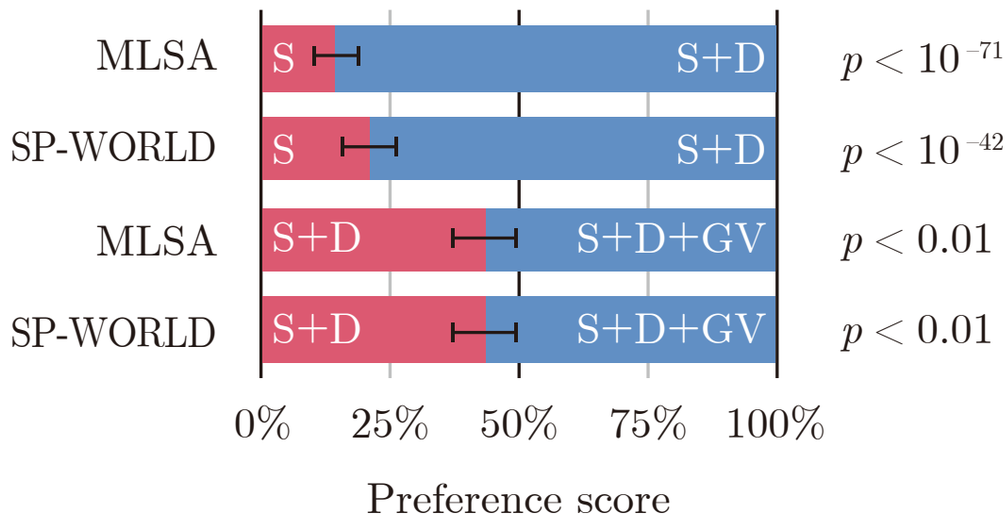
naturalness

- Sequence features:

  static < static+dynamic < **static+dynamic+GV**



speaker similarity



naturalness

Conclusion

- In GMM-based statistical voice conversion,
  - Dynamic features and GV: definitely effective
  - SP-WORLD: comparable to MLSA
    - Superior in ideal conversion
  - Higher order of features: **not** always effective
- High time-resolution analysis is effective in analysis-synthesis
  - Potential of effectiveness also in conversion

Future Works

- F_0 conversion
- Break the 1 ms barrier in WORLD analysis
- Other mapping models such as neural networks