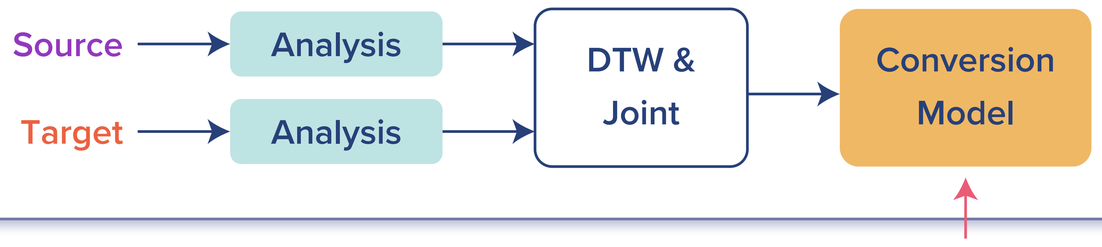
A Revisit to Feature Handling for High-quality Voice Conversion Based on Gaussian Mixture Models

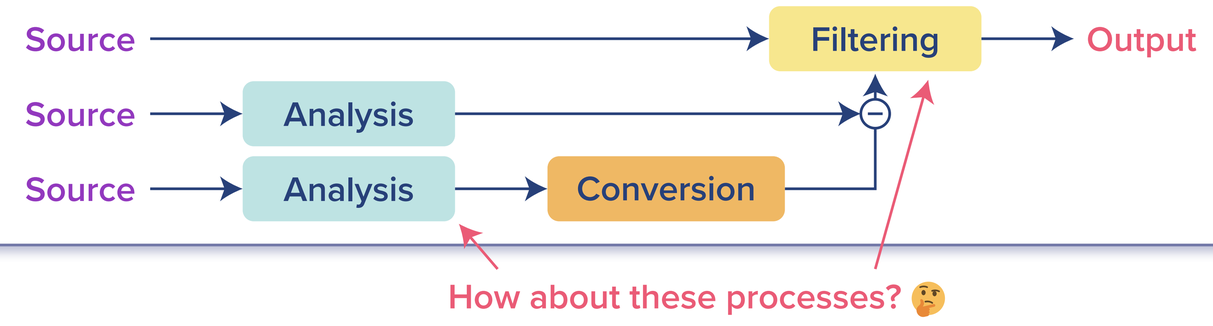
Hitoshi Suda, Gaku Kotani, Shinnosuke Takamichi, and Daisuke Saito (The University of Tokyo)

Backgrounds: VC Frameworks

Training:



Conversion:



Aim of This Study

- To improve quality of existing voice conversion frameworks
- To reveal influences of feature handling via subjective experiments
- To perform best VC with WORLD analysis [M.Morise+, 2016] + GMM-based conversion [A.Kain+, 1998] + diffspec [K.Kobayashi+, 2014]

Experimental Setups

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

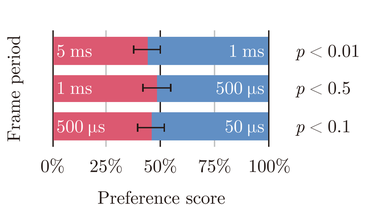
Experiment 1: Analysis Conditions

Source → Analysis → (No conversion) → Synthesis → Output

- Frame periods (frame shifts): How precisely are the waveforms analyzed in time domain?
- Order of mel-cepstral coefficients (mcep): How precisely are the spectral envelopes represented?

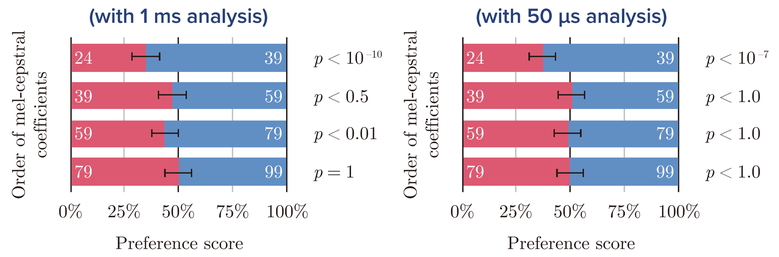
Frame periods:

- 5 ms < 1 ms ≈ 500 μs ≤ 50 μs

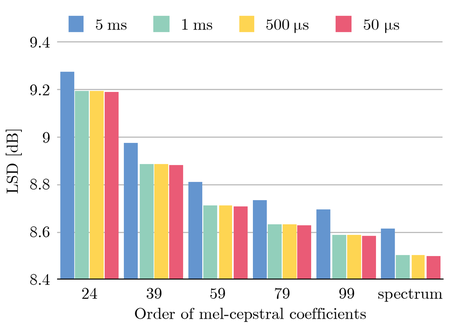


Order of mcep:

- 1 ms: 24 < 39 ≈ 59 < 79 ≈ 99
- 50 μs: 24 < 39 ≈ 59 ≈ 79 ≈ 99



(Objective evaluation)
Log-spectral distances between original and re-synthesized waveforms

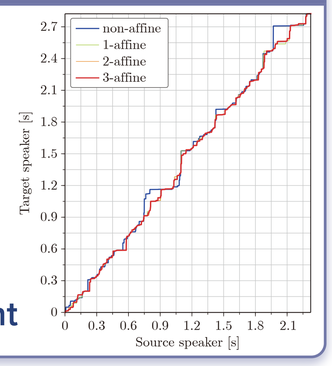


SP-WORLD: New Differential-spectrum Compensation (Diffspec) Implementation

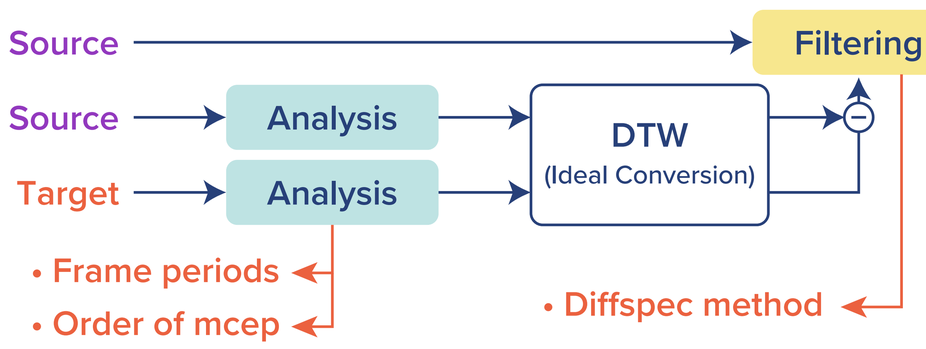
- Famous method (Mel log spectrum approximation (MLSA) filtering [S.Imai+, 1983]) can degrade synthesis quality because of its approximation
- We introduce SP-WORLD inspired by WORLD vocoder
  - Based on minimum phase reconstruction from real cepstra

Affine-DTW: Another DTW Method

- Iteration of general DTW and affine transformation of source features
- The influence on alignment of the difference of speakers can be diminished
- Affince transformation ≈ GMM-based VC with 1 Gaussian component



Experiment 2: Conversion System without Statistical Mapping



Diffspec method:

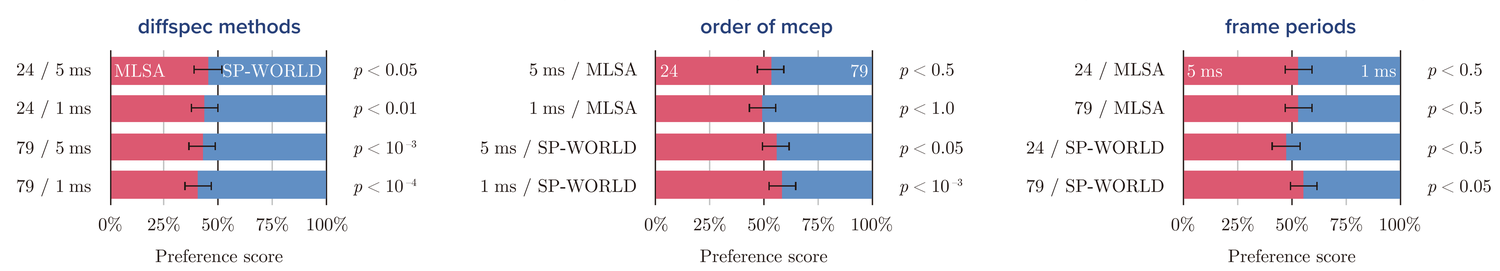
- MLSA < SP-WORLD

Order of mcep:

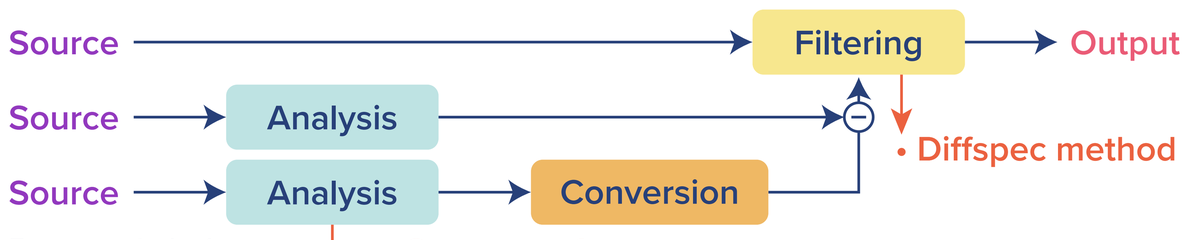
- 24 > 79 (with SP-WORLD)
- 24 ≈ 79 (with MLSA)

Frame periods:

- 5 ms ≥ 1 ms (with SP-WORLD / 79-order)
- 5 ms ≈ 1 ms (with other conditions)



Experiment 3: Total Conversion System



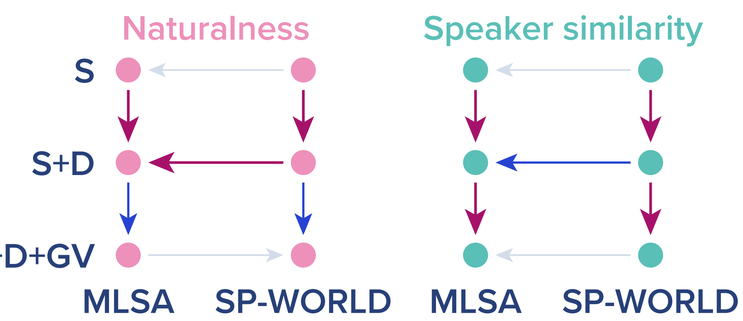
Frame periods: 1 ms
Order of mcep: 24
Gaussian components:
64 without dynamic features / 128 with dynamic features

Diffspec method:

- MLSA > SP-WORLD (S+D)
- MLSA ≈ SP-WORLD (S+D+GV)

Sequence features:

- S < S+D < S+D+GV



S: Static features / D: Dynamic features

→ no significant difference → significantly different ($p < 0.05$) → significantly different ($p < 10^{-3}$)

Conclusion

- SP-WORLD is comparable to MLSA
  - superior in more sophisticated conversion?
- Dynamic features and GV are definitely effective
- Features with higher order are not always effective
  - because of conversion errors in higher order?

Future Works

- F_0 conversion
- Break the 1 ms barrier of WORLD analysis
- Other sophisticated mapping models (e.g. NN)

Acknowledgements This work was partly supported by SECOM Science and Technology Foundation

URL https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/publications/190227-sigslp-poster.pdf