Voice Conversion without Explicit Separation of Source and Filter Components Based on Non-negative Matrix Factorization

Hitoshi Suda, Daisuke Saito, and Nobuaki Minematsu (The University of Tokyo)

Synthesis methods for VC

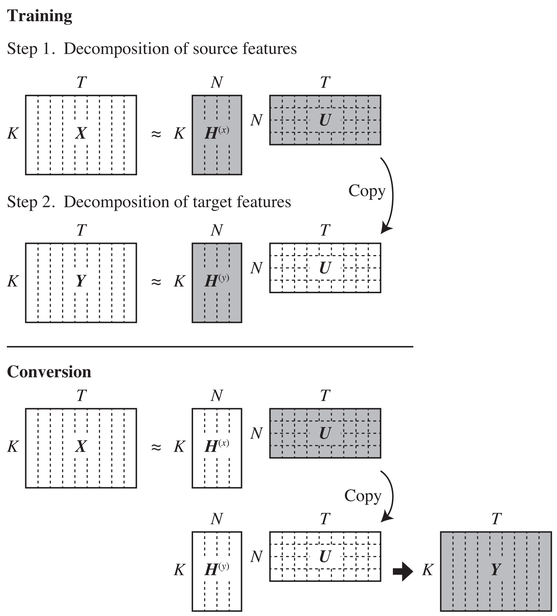
- Vocoders (e.g. WORLD [Morise+, 2016], STRAIGHT [Kawahara+, 2006])
  - ✓ very lightweight
  - ✗ limited quality due to strict parameterization
- Spectral differential compensation [Kobayashi+, 2014]
  - ✓ high-quality and lightweight
  - ✗ difficult to modify source component
- **Phase reconstruction** / Waveform generation from spectra (e.g. Griffin-Lim [Griffin+, 1984], von-Mises-distribution NNs [Takamichi+, 2018], WaveNet [van den Oord+, 2016], WaveRNN [Kalchbrenner+, 2018])
  - ✓ flexible and capable of high-quality synthesis
  - ✗ estimation of spectra themselves is difficult

VC based on non-negative matrix factorization [Takashima+, 2012]

- NMF: decomposition of a non-negative matrix into 2 matrices

$$\boldsymbol{x}_t \approx \sum_n u_{n,t} \boldsymbol{h}_n^{(x)}$$

  - $u_{n,t}$: how active the n-th template $\boldsymbol{h}_n$ is at the time t
- Concept of NMF-VC: both speakers' utterances can be represented by weighted summation of corresponding spectral templates



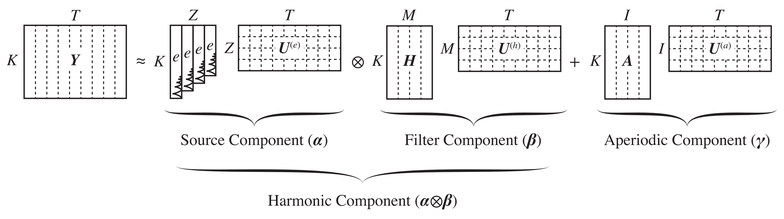
VC based on simplified source-filter NMF

- Source-filter NMF [Virtanen+, 2006]: extended model of NMF for spectrograms of polyphonic sounds with source structures

$$y_t(k) \approx \sum_{n,m,z,t} e_n(k-z) h_m(k)$$

  - ✗ So complicated for single speaker's utterances
- Proposition: simplified model of SF-NMF

$$y_t(k) \approx \left(\sum_z u_{z,t}^{(e)} e(k-z)\right)\left(\sum_m u_{m,t}^{(h)} h_m(k)\right) + \sum_i u_{i,t}^{(a)} a_i(k)$$



- SF-NMF-VC can be achieved in the same way as NMF-VC

Scalogram and its phase reconstruction [Irino+, 1993]

- SF-NMF models source structures as summation of **shifted** templates
  - Decomposed matrices must be log-frequency spectrograms
- We use **continuous wavelet transform (CWT)** to obtain spectrograms (scalograms) in an arbitrary scale
  - Basis function of CWT: wavelets (c.f. windowed sinusoidal in STFT)
- CWT: convolution of waveforms and wavelets

$$\boldsymbol{s}_l = \boldsymbol{\psi}_l \otimes \boldsymbol{x}$$

  - l: frequency index, $\boldsymbol{\psi}_l$: wavelet with l-th center frequency
  - Equivalent to multiplication of waveforms $\boldsymbol{x}$ and conv. matrix $\boldsymbol{W}$

$$\boldsymbol{s} = \left[\boldsymbol{s}_0^\top, \ldots, \boldsymbol{s}_{L-1}^\top\right]^\top = \boldsymbol{W}\boldsymbol{x}$$

- Phase reconstruction: estimation of most consistent phase

$$\tilde{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \left\| \boldsymbol{a}e^{i\boldsymbol{\phi}} - \boldsymbol{W}\boldsymbol{W}^{+}\boldsymbol{a}e^{i\boldsymbol{\phi}} \right\|$$

  - $\boldsymbol{W}^{+}$: pseudo inverse of $\boldsymbol{W}$
  - c.f. Griffin-Lim [Griffin+, 1984]
- CWT and phase reconstruction can be accelerated [Nakamura+, 2014]
  - $\boldsymbol{\psi}_l$ is band-limited ($\boldsymbol{W}$ is sparse)

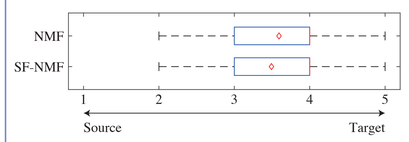
Experiment

Experimental Setups

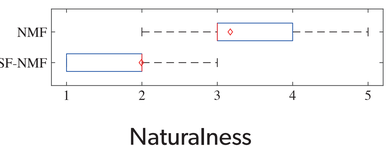
- Dataset: ATR Japanese phonetically balanced sentence sets
  - Subset A (50 sentences) for training, subset J (53 sentences) for test
  - Source: male / Target: female / Sampling frequency: 16 kHz
  - Speaker similarity and naturalness is evaluated by MOS (28 subjects for similarity, 29 for naturalness)
- SF-NMF-based system:
  - N = 1 (source), M = 200 (filter), I = 5 (aperiodicity)
  - Z = 96 (48 bins per oct.)
  - $\boldsymbol{U}^{(e)}$ is initialized based on WORLD analysis
  - Mother wavelet: log-normal wavelet [Kameoka, 2007]
  - Frequency bins of scalograms: 50 Hz – $2^{7.25}$ (≈ 7611) Hz (349 bins)
- NMF-based system (for comparison):
  - Number of dictionaries: 200
  - Analysis, synthesis: WORLD [Morise+, 2016] (D4C edition [Morise, 2016])

Experimental Results

- Proposed framework achieved speaker conversion, but the quality of generated utterances sounded lower
- Audio samples:
  `https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/publications/190920-ssw/`



Speaker identity



Naturalness

Conclusion

- Proposition: **spectrogram-to-spectrogram VC based on SF-NMF**
- The quality did not reach frameworks using vocoders

Future works

- Investigation of the effectiveness of wavelets
- Conversion from/to abnormal utterances (e.g. gravelly or creaky voices)

Acknowledgements
This research and development work was supported by the MIC/SCOPE #182103104.

PDF data of this poster
`https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/publications/190920-ssw-poster.pdf`