SP-P6.5
Wednesday April 18

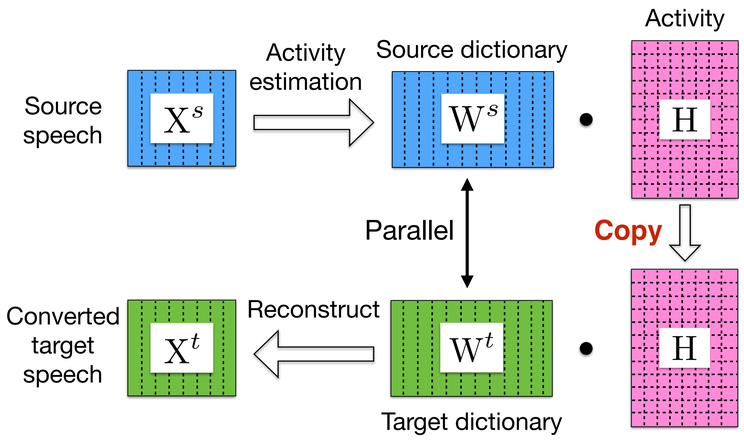
PARALLEL-DATA-FREE DICTIONARY LEARNING FOR VOICE CONVERSION USING NON-NEGATIVE TUCKER DECOMPOSITION

Yuki Takashima[1], Hajime Yano[1], Toru Nakashika[2], Tetsuya Takiguchi[1], Yasuo Ariki[1]
1. Graduate School of System Informatics, Kobe University, Japan,
2. Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan

Voice Conversion

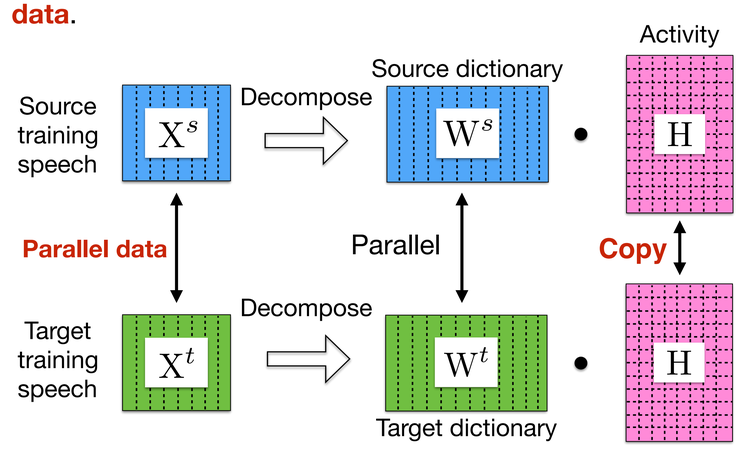
Voice conversion (VC) is a technique for changing speaker information in input speech signal while maintaining linguistic and emotion information.

Non-negative matrix factorization (NMF)-based voice conversion
An input source signal is decomposed into a linear combination of basis from the source dictionary.
A target signal is constructed from the replaced bases of the target dictionary and the weights of source bases.



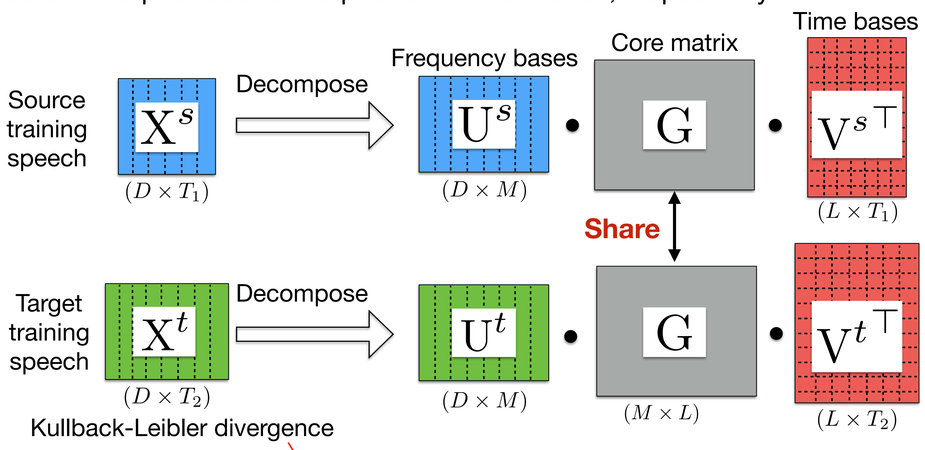
NMF-based Dictionary Learning

In NMF-based dictionary learning, source and target dictionaries are learned while sharing a same activity matrix.
This method needs a large number of **parallel training data**.



Proposed Method

We propose a non-negative Tucker decomposition (NTD)-based dictionary learning method. The NTD is a non-negative extension of Tucker decomposition that decomposes the input observation into a set of matrices and one core tensor.
In the spectral domain, the NTD decompose the input spectrogram into three matrices; a frequency basis matrix, a time basis matrix, and a core matrix.
A core matrix is shared between speakers, and the time-varying matrices are dependent on each speaker.
We assume that the frequency basis matrices, the core matrix and the time basis matrices represent the speaker information, the codebook between the frequency bases and the phones and the phonemic information, respectively.

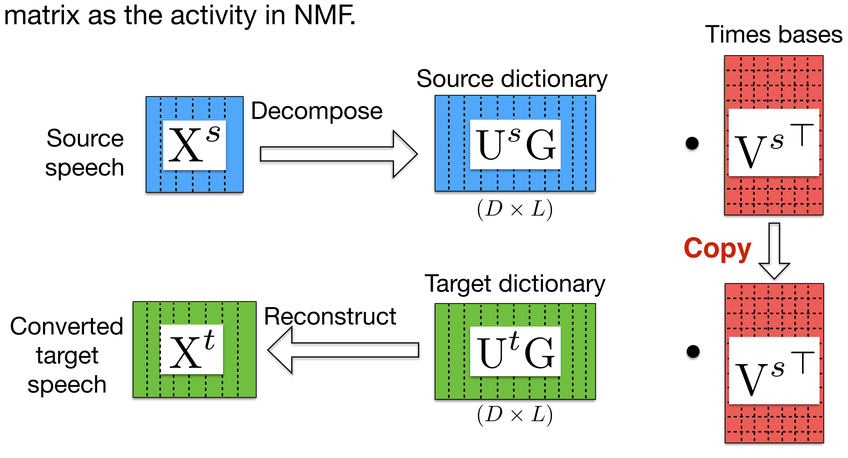


Kullback-Leibler divergence

$$L_{NMF} = d_{KL}(\mathbf{X}^s, \mathbf{W}^s\mathbf{H}) + d_{KL}(\mathbf{X}^t, \mathbf{W}^t\mathbf{H})$$

$$L_{NTD} = d_{KL}(\mathbf{X}^s, \mathbf{U}^s\mathbf{G}\mathbf{V}^{s\top}) + d_{KL}(\mathbf{X}^t, \mathbf{U}^t\mathbf{G}\mathbf{V}^{t\top})$$

After each matrix in the model is estimated, the source and target parallel dictionaries are calculated multiplying the frequency matrix by the core matrix.
When converting, for the given source spectrogram, we estimate only the time bases matrix as the activity in NMF.



Experiments

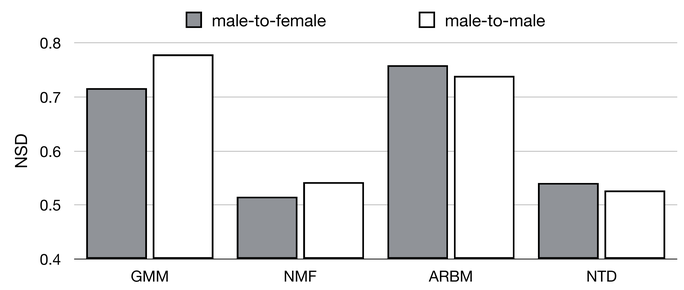
Condition

Features	Spectrum (NMF, NTD) Cepstrum (GMM, ARBM)
Sampling rate	8 kHz
Speaker	1 male, 1 female
#training data	50 utterances

$$NSD = \sqrt{\frac{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2}{||\mathbf{X}^t - \mathbf{X}^s||^2}}$$

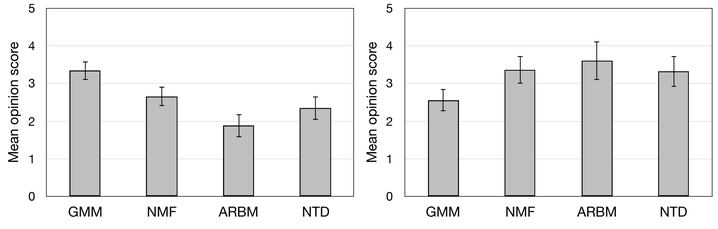
$\hat{\mathbf{X}}^t$: Converted
$\mathbf{X}^t$: Target
$\mathbf{X}^s$: Source

Objective Evaluation (50 utterances)



Subjective Evaluation (10 utterances)

10-native Japanese listened, 5-scale MOS test



Conclusion
We proposed a dictionary learning of NMF-based VC which allows NMF-VC for non-parallel training based on NTD.
We obtained equivalent performance compared with a conventional method using parallel data.