Exemplar-based voice conversion using non-negative spectrogram deconvolution

Zhizheng Wu[1], Tuomas Virtanen[2], Tomi Kinnunen[3], Eng Siong Chng[1], Haizhou Li[1,4]

[1]Nanyang Technological University, Singapore

[2]Tampere University of Technology, Finland

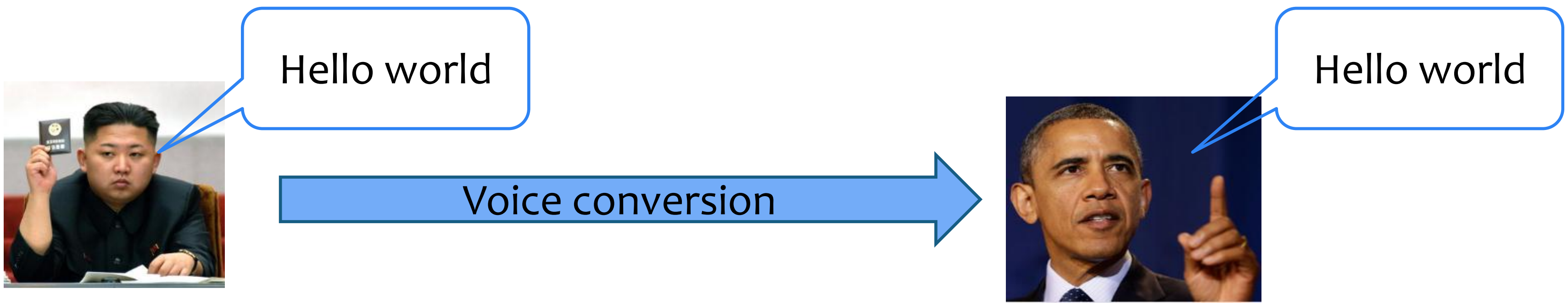
[3]University of Eastern Finland, Finland

[4]Institute for Infocomm Research, Singapore

Email: wuzz@ntu.edu.sg

Introduction of voice conversion

- Techniques for modifying the para-linguistic information (*speaker identity, speaking styles, and so on*) while keeping linguistic information (*language content*) unchanged.

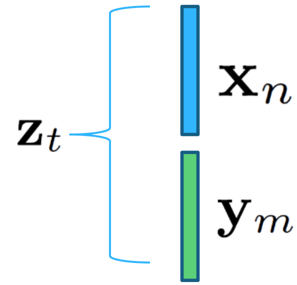


Source speaker's voice

Target speaker's voice

Baseline method

- JD-GMM: joint density Gaussian mixture model
  - Joint probability density



$$P(\mathbf{X},\mathbf{Y}) = P(\mathbf{Z}) = \sum_{k=1}^{K} w_k^{(z)} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k^{(z)}, \boldsymbol{\Sigma}_k^{(z)})$$

$$\boldsymbol{\mu}_k^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_k^{(x)} \\ \boldsymbol{\mu}_k^{(y)} \end{bmatrix} \quad \boldsymbol{\Sigma}_k^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(xx)} & \boldsymbol{\Sigma}_k^{(xy)} \\ \boldsymbol{\Sigma}_k^{(yx)} & \boldsymbol{\Sigma}_k^{(yy)} \end{bmatrix}$$

  - Conversion function:

$$\hat{\mathbf{y}} = F(\mathbf{x}) = \sum_{k=1}^{K} p_k(\mathbf{x})(\boldsymbol{\mu}_k^{(y)} + \boldsymbol{\Sigma}_k^{(yx)}(\boldsymbol{\Sigma}_k^{(xx)})^{-1}(\mathbf{x} - \boldsymbol{\mu}_k^{(x)}))$$

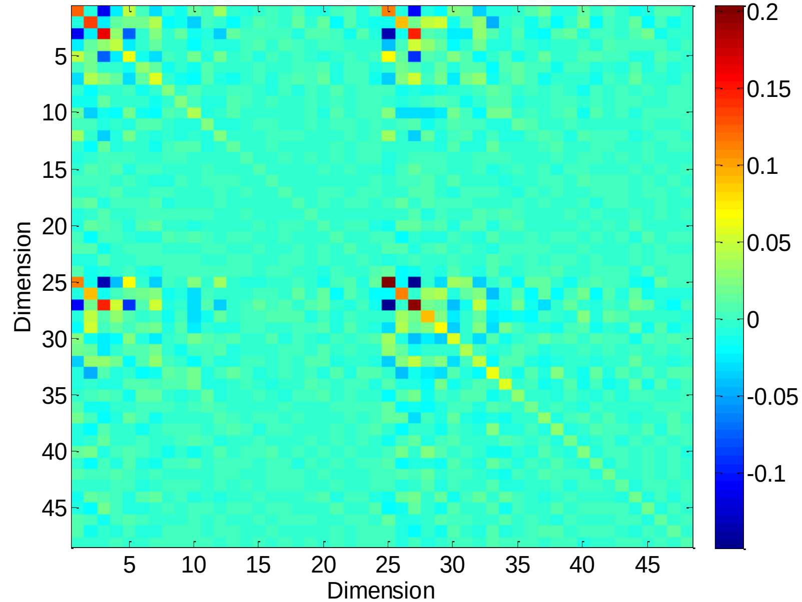
    - $p_k(\mathbf{x}) = \dfrac{w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^{x}, \boldsymbol{\Sigma}_k^{xx})}{\sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^{x}, \boldsymbol{\Sigma}_k^{xx})}$ is the posteriori probability of x belong to k^{th} Gaussian component

Problems in JD-GMM

- Statistical average
  - Estimation of mean and covariance

$$\boldsymbol{\mu}_k^{(z)} = \frac{\sum_{t=1}^{T} \mathbf{z}_t p_k(\mathbf{z}_t, \lambda^{(z)})}{\sum_{t=1}^{T} p_k(\mathbf{z}_t, \lambda^{(z)})} \qquad \boldsymbol{\Sigma}_k^{(z)} = \frac{\sum_{t=1}^{T} p_k(\mathbf{z}_t, \lambda^{(z)})(\mathbf{z}_t - \boldsymbol{\mu}_k^{(z)})(\mathbf{z}_t - \boldsymbol{\mu}_k^{(z)})^\top}{\sum_{t=1}^{T} p_k(\mathbf{z}_t, \lambda^{(z)})}$$

Average over all the training samples



Motivation

- Avoid estimating covariance matrix which usually 'bad' estimated
- To transform relative high-dimensional spectral envelopes directly
- Include temporal constraint in generation of spectrogram

Non-negative spectrogram factorization (NMF)

- Basic idea: to represent magnitude spectra as a linear combination of a set of basis spectra (speech atoms)

$$\mathbf{x} = \sum_{t=1}^{T} \mathbf{a}_t^{(\mathrm{X})} \cdot h_t = \mathbf{A}^{(\mathrm{X})} \cdot \mathbf{h}$$

- NMF for voice conversion

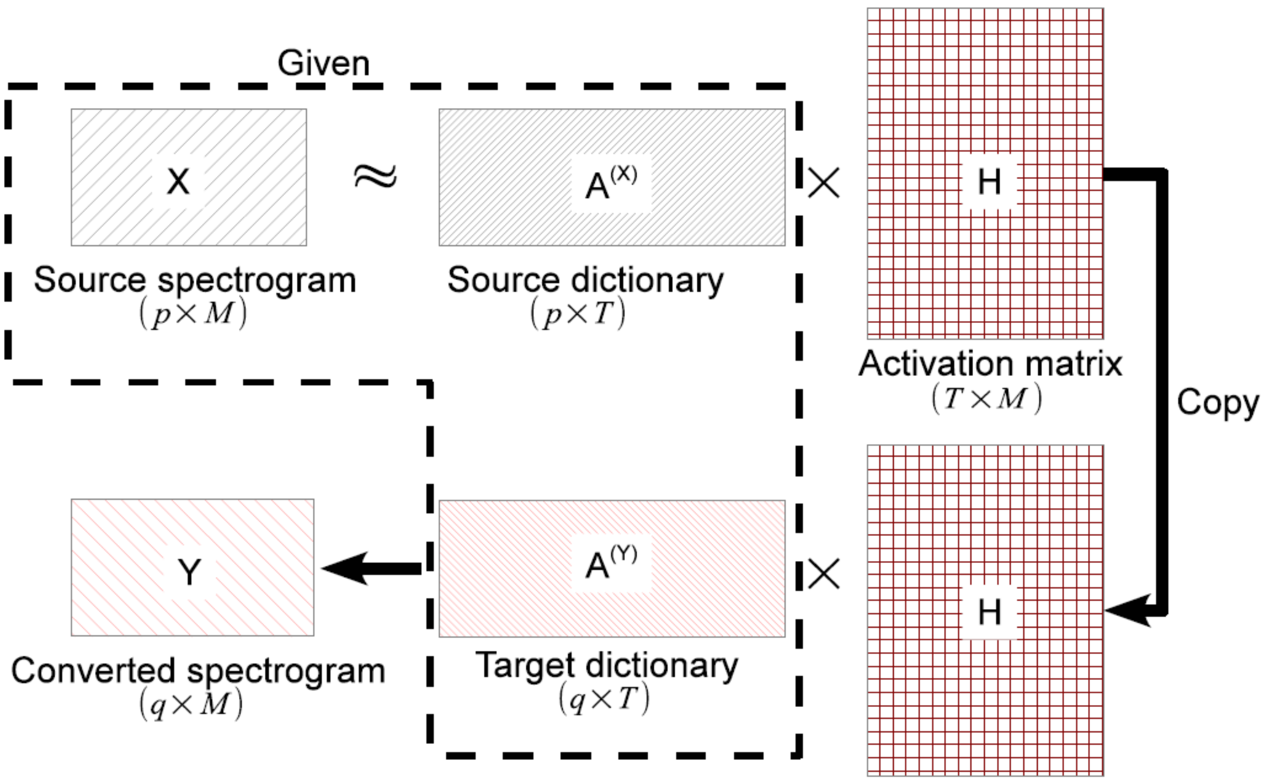
$$\mathbf{X} = \mathbf{A}^{(\mathrm{X})} \cdot \mathbf{H}$$

$$\mathbf{Y} = \mathbf{A}^{(\mathrm{Y})} \cdot \mathbf{H}$$

  - $\mathbf{X}$ and $\mathbf{Y}$ are source and converted spectrograms, respectively
  - $\mathbf{A}^{(\mathbf{X})}$ and $\mathbf{A}^{(\mathbf{Y})}$ are source and target exemplar dictionaries, respectively
  - $\mathbf{H}$ is the activation matrix, column vector, $\mathbf{h}$, of $\mathbf{H}$ consists of non-negative weights

Non-negative spectrogram factorization (NMF)

- Illustration of NMF



Non-negative spectrogram deconvolution (NMD)

- The idea: to include temporal constraint in the estimation of activation matrix and also the generation of spectrogram
- Formulation:

$$\mathbf{X} = \sum_{l=1}^{L} \mathbf{A}_l^{(\mathrm{X})} \cdot \overset{\rightarrow (l-1)}{\mathbf{H}}$$

$$\mathbf{Y} = \sum_{l=1}^{L} \mathbf{A}_l^{(\mathrm{Y})} \cdot \overset{\rightarrow (l-1)}{\mathbf{H}}$$

  - $\mathbf{A}_l^{(\mathrm{X})} \in \mathcal{R}^{p \times T}$ and $\mathbf{A}_l^{(\mathrm{Y})} \in \mathcal{R}^{q \times T}$ are the matrices consisting of the l^{th} frame of the source and target atoms, respectively
  - L is the number of adjacent frames within an exemplar
  - $\overset{\rightarrow (l-1)}{(\cdot)}$ operator shifts the matrix entries (columns) to the right by $(l-1)$ unit

Features

- Magnitude spectrum (MSP): use 513-dimenaional spectral envelope extracted by STRAIGHT. We use MSP to reconstruct speech signal.
- Mel-scale magnitude spectrum (MMSP): pass MSP to a 23-channel Mel-scale filter-bank. The minimum frequency is set to be 133.33 Hz, and the maximum frequency is set to be 6,855.5 Hz.
- Mel-cepstral coefficient (MCC): MCC is obtained by employing mel-cepstral analysis on magnitude spectrum and keeping 24 coefficients as the feature

Dictionary construction

- Processes to build source and target dictionaries
  - Extract magnitude spectrograms (MSP) using STRAIGHT;
  - Apply Mel-cepstral analysis on MSP to obtain Mel-cepstral coefficients (MCCs);
  - Apply 23-channel Mel-scale filter-bank on the spectrograms to obtain 23-dimensional Mel-scale magnitude spectra (MMSP);
  - Perform dynamic time warping (DTW) to the source and target MCC sequence to align source and target speech to obtain source-target frame pairs;
  - Apply the alignment information to the source MMSP (or MSP) and target MSP. The resulting spectrum pairs are stored in the source and target dictionaries (column vectors), respectively.

Experimental setups

- Corpus
  - VOICES database: parallel corpus
    - Male-to-female and female-to-male conversions are conducted
    - 10 utterances from each speaker are used as training set
    - 20 utterances from each speaker as testing set

- Fundamental frequency (F0) is converted by equalizing the means and variances of source and target speaker in log-scale.

Objective evaluation measure

- Mel-cepstral distortion: calculation is done frame-by-frame

$$\mathrm{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}(c_{m,d} - c_{m,d}^{\mathrm{conv}})}$$

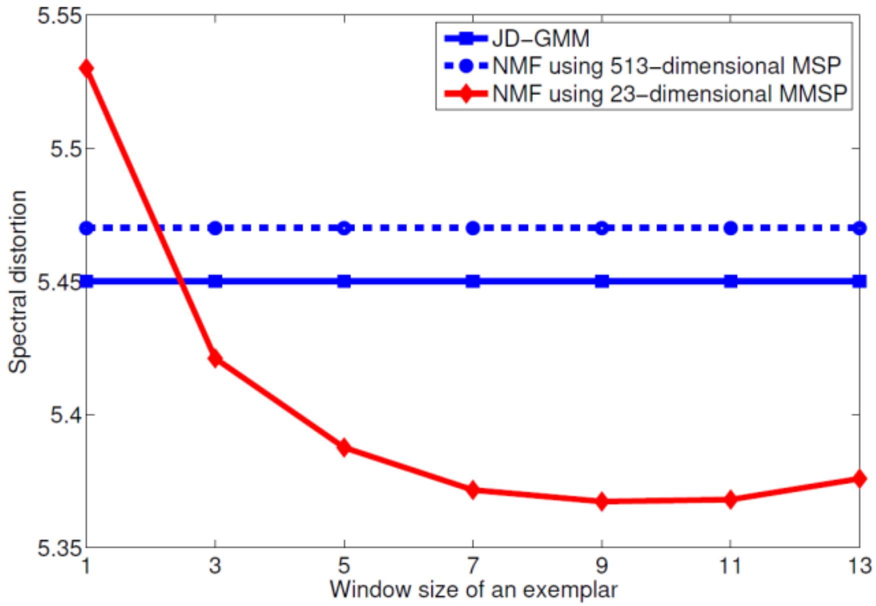
- Correlation coefficient: calculation is done dimension-by-dimension

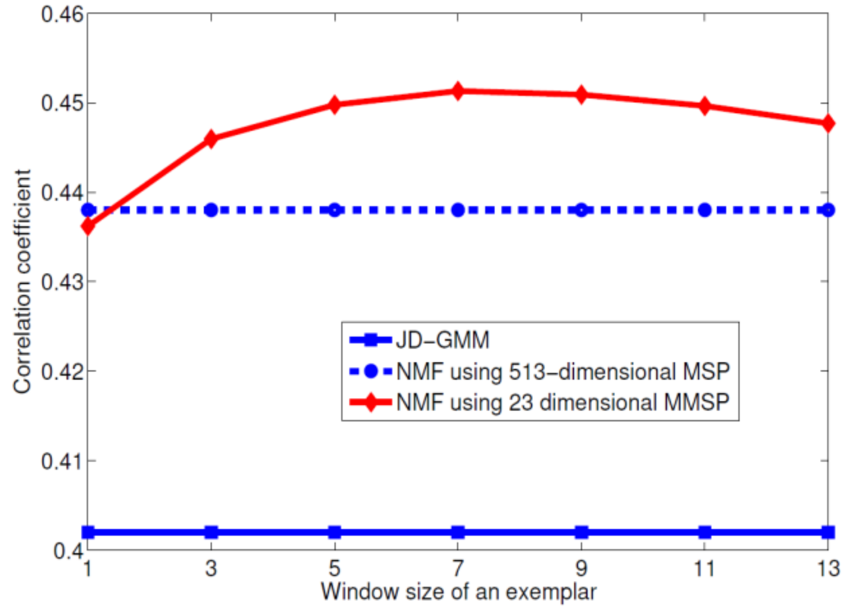
$$\gamma_d = \frac{\sum_{m=1}^{M}(c_{m,d} - \overline{c_d})(c_{m,d}^{\mathrm{conv}} - \overline{c_d^{\mathrm{conv}}})}{\sqrt{\sum_{m=1}^{M}(c_{m,d} - \overline{c_d})^2}\sqrt{\sum_{m=1}^{M}(c_{m,d}^{\mathrm{conv}} - \overline{c_d^{\mathrm{conv}}})^2}}$$

  - $c_{m,d}$ and $c_{m,d}^{\mathrm{conv}}$ are the d^{th} dimension feature of the m^{th} frame original target and converted MCC vector, respectively.
  - $\overline{c_d}$ and $\overline{c_d^{\mathrm{conv}}}$ are the mean values of the d^{th} dimension original target and converted MCC trajectories, respectively.

Experimental results

- Comparison of NMF using 513-dimension MSP and 23-dimensional MMSP in the source dictionary
  - Spectral distortion and correlation results as a function of the window size of an exemplar

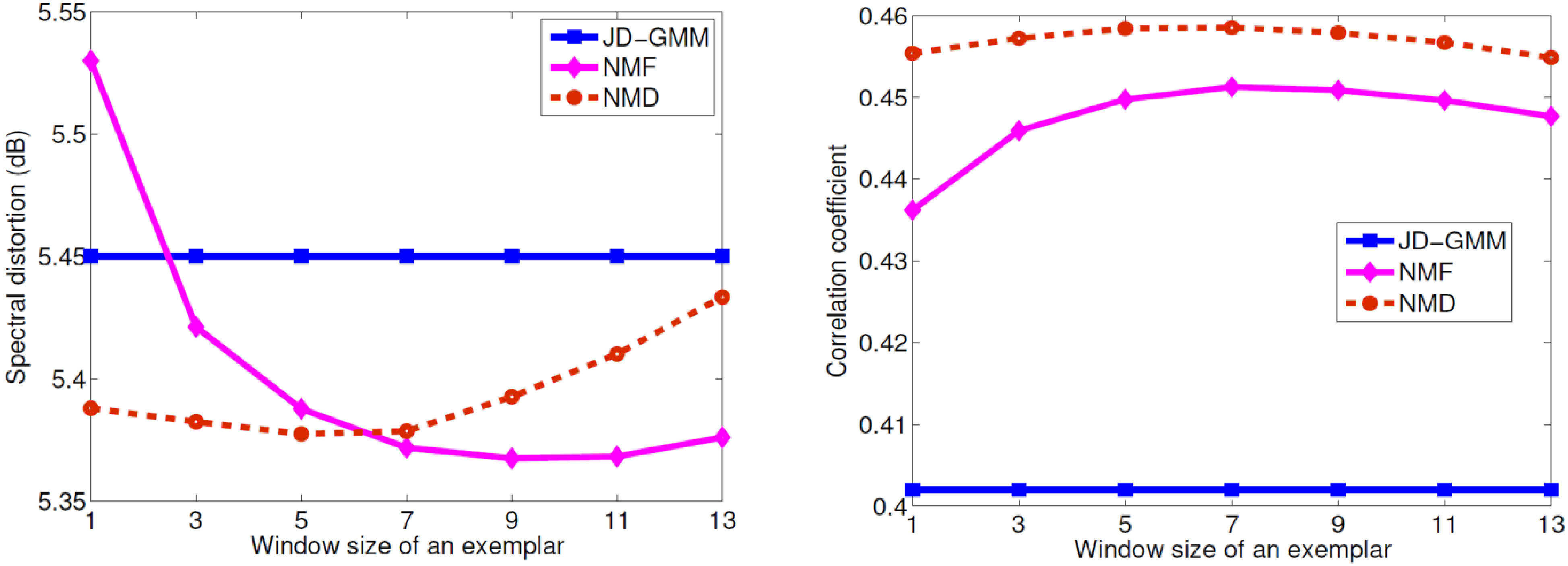




23-dimensional MMSP yields lower MCD and higher correlation coefficient than 513-dimensional MSP

Experimental results

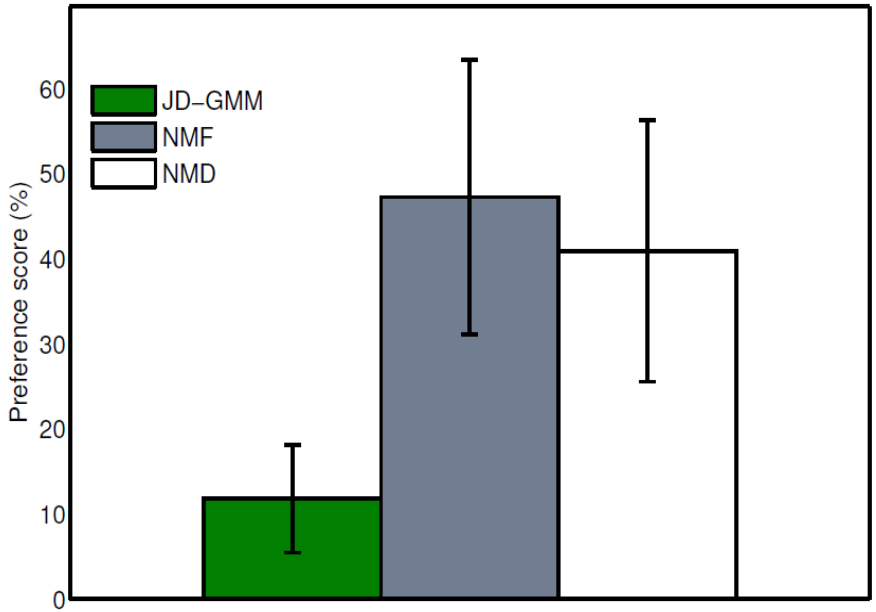
- Spectral distortion and correlation results comparison of JD-GMM, NMF and NMD methods as a function of the window size of an exemplar.



1, Both NMF and NMD obtain lower distortion and higher correlation than JD-GMM.
2, NMD method obtains higher correlation than NMF method.

Subjective evaluation results

- Preference score with 95% confidence interval for speaker similarity



Both NMF and NMD **outperform** JD-GMM method!

Converted speech quality? Listen to our demo!

Conclusions

- We proposed an exemplar-based voice conversion method utilizing the matrix/spectrogram factorization techniques.

- Both non-negative spectrogram factorization and non-negative spectrogram deconvolution are implemented to use original target spectrogram directly without any dimension reduction to synthesize the converted speech.

- NMF and NMD both outperforms the conventional JD-GMM method.