

EPSRC

Engineering and Physical Sciences
Research Council



A study of speaker adaptation for DNN-based speech synthesis

Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals, Simon King

The Centre for Speech Technology Research (CSTR)
University of Edinburgh, United Kingdom



Background

- A speaker-dependent TTS system requires several hours recordings in studio
 - It is expensive to collect

- Adaptation for speech synthesis
 - Create a new voice using minimal data, for example 1 minute speech

Related work

- Speaker adaptation for statistical parametric speech synthesis
 - MLLR, CMLLR, MAP, MAPLR, CSMAPLR, etc

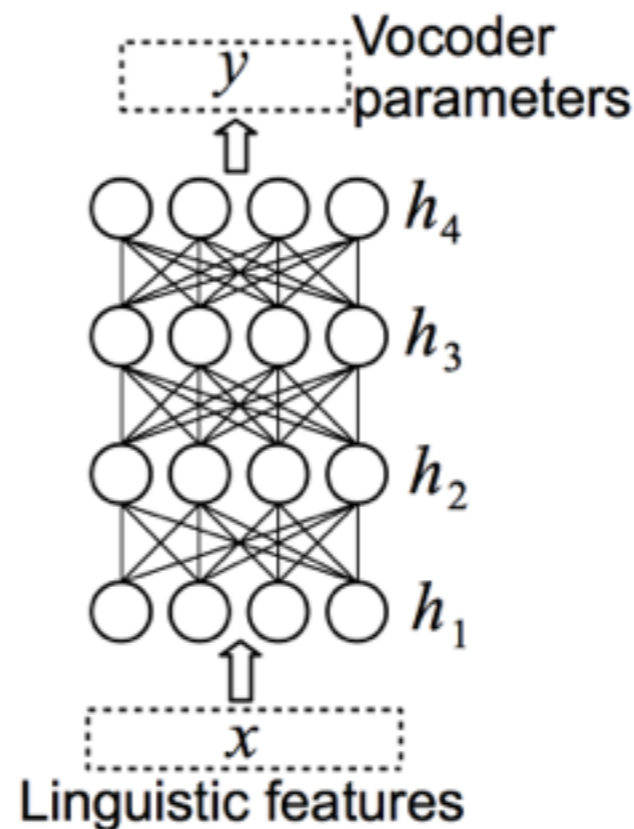
- Voice conversion for unit-selection concatenation speech synthesis

Yamagishi, Junichi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai. "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm." *IEEE Transactions on Audio, Speech, and Language Processing*, 17, no. 1 (2009): 66-83.

Kain, Alexander, and Michael W. Macon. "Spectral voice conversion for text-to-speech synthesis." In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998. vol. 1, pp. 285-288.

DNN-based speech synthesis

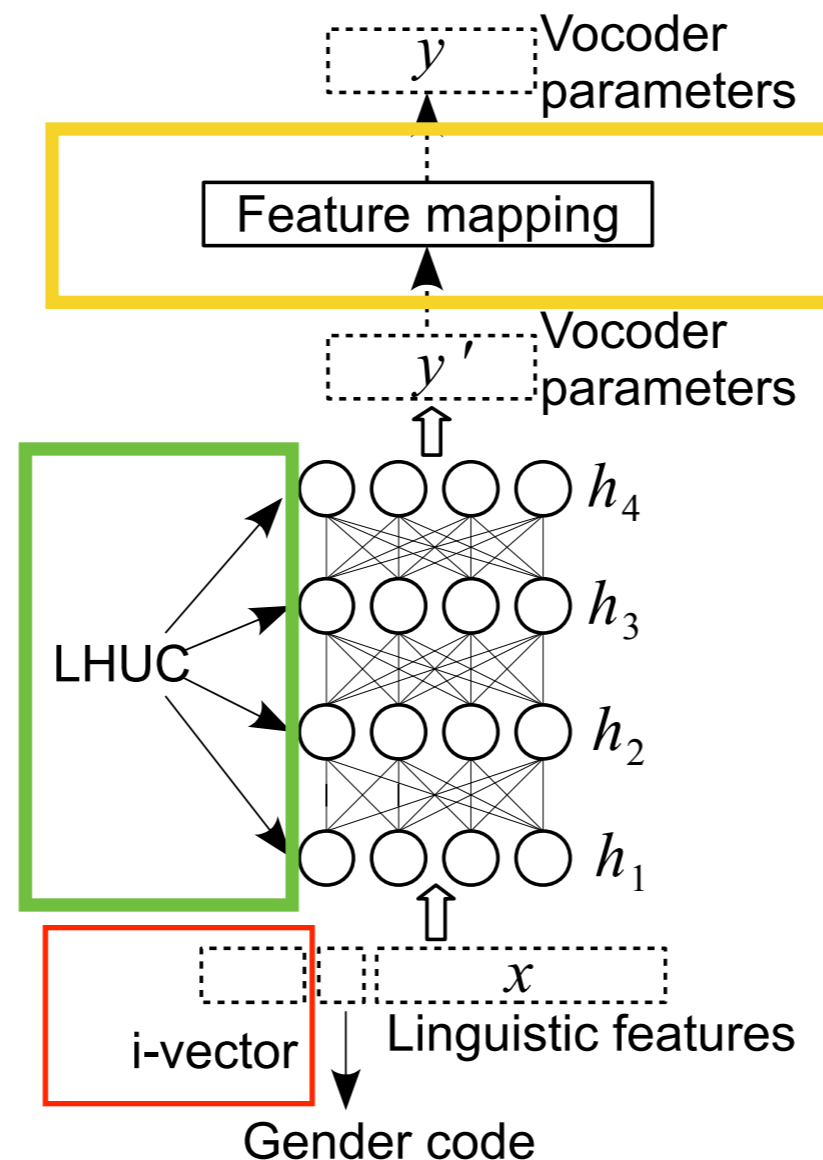
- Mapping linguistic features to vocoder parameters using a deep neural network
 - Outperform HMM-based speech synthesis in terms of naturalness



Heiga Zen, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks." ICASSP 2013
Yao Qian, Yuchen Fan, Wenping Hu, and Frank K. Soong. "On the training aspects of deep neural network (DNN) for parametric TTS synthesis." ICASSP 2014

Proposed adaptation framework for DNN-based speech synthesis

- Performing speaker adaptation at three different levels



LHUC: Learning hidden unit contributions

Adaptation framework: i-vector

- I-vector extraction

$$\mathbf{s} \approx \mathbf{m} + \mathbf{T}\mathbf{i}, \quad \mathbf{i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- \mathbf{m} is the mean supervector of a speaker-independent universal background model (UBM)
- \mathbf{s} is the mean supervector of the speaker-dependent GMM model (adapted from the UBM)
- \mathbf{T} is the total variability matrix estimated on the background data
- \mathbf{i} is the speaker identity vector, also called the i-vector

Dehak, Najim, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. "Front-end factor analysis for speaker verification." IEEE Transactions on Audio, Speech, and Language Processing, 19, no. 4 (2011): 788-798.

Adaptation framework: LHUC

- Learning hidden unit contribution

$$\mathbf{h}_m^l = \alpha(\mathbf{r}_m^l) \odot (\mathbf{W}^{l\top} \mathbf{h}_m^{l-1})$$

- \mathbf{h}_m^l is the activations of the l^{th} hidden layer
- $\alpha(\mathbf{r}_m^l)$ is an element-wise function to constrain the range of \mathbf{r}_m^l
- \mathbf{W}^l is the weight matrix of the l^{th} hidden layer

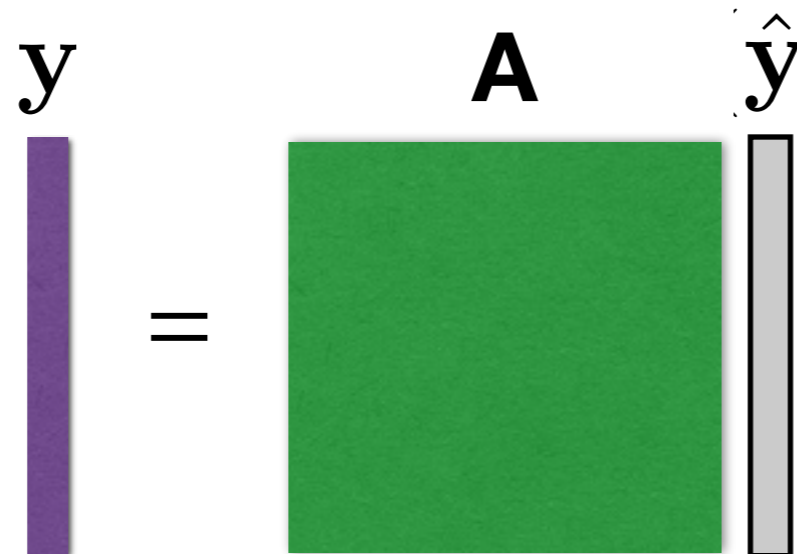
$$\mathbf{h}_m^l = \mathbf{W}^{l\top} \mathbf{h}_m^{l-1}$$

- setting $\alpha(\mathbf{r}_m^l) = 1$, the hidden activation will become the normal one

Swietojanski, Pawel, and Steve Renals. "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models." In IEEE Spoken Language Technology Workshop (SLT), 2014

Adaptation framework: feature space adaptation

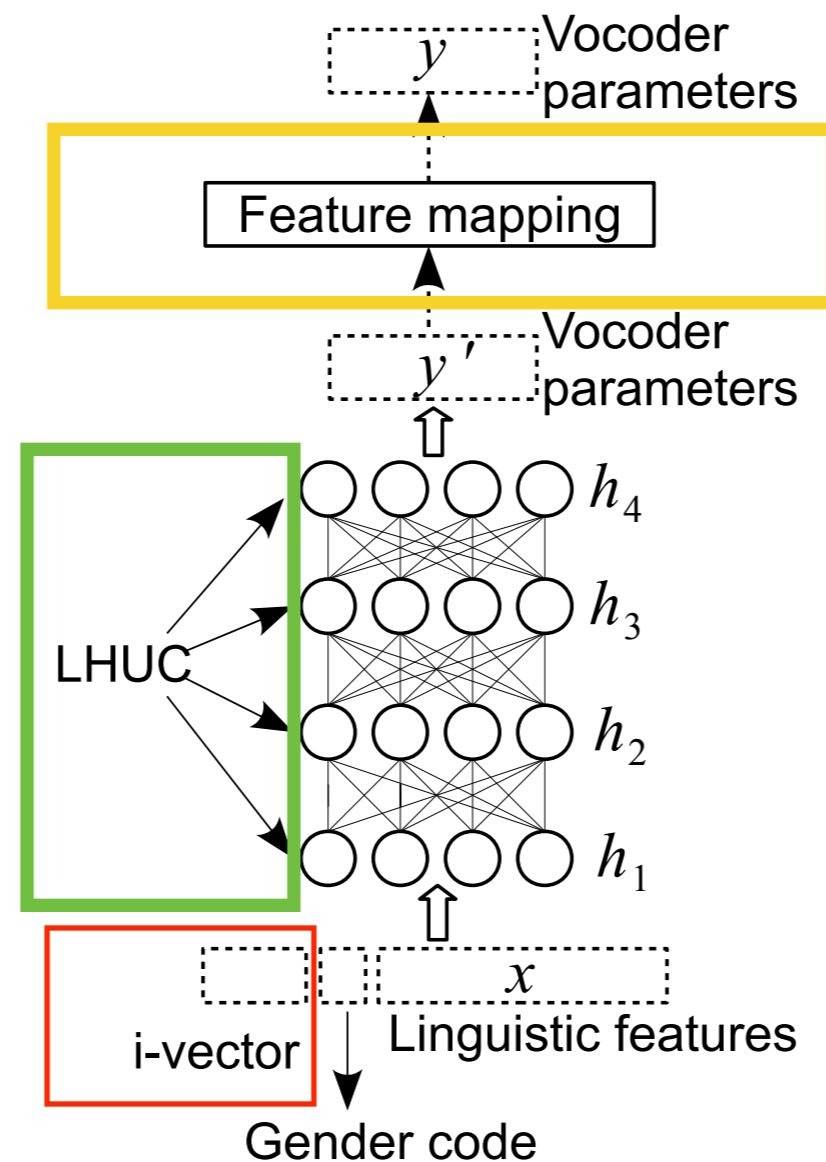
- Feature transformation: Transform the output of DNN using a linear transformation

$$\mathbf{y} = \mathbf{A} \hat{\mathbf{y}}$$


- **A** is a linear transformation matrix

Adaptation framework: combination of individual techniques

- As each adaptation method is applied at different level, they can easily combined



Experimental setups

- Corpus
 - Voice bank database: 96 speakers (41 male, 55 female)
 - To build speaker-independent average DNN model
 - Sampling rate: 48 kHz
 - Each speaker has around 300 utterances
 - Two target speakers (one male, one female)
 - 10 utterances for adaptation, 70 development, 72 testing
- Vocoder parameters (extracted by STRAIGHT)
 - 60-D Mel-Cepstral Coefficients with delta, delta-delta
 - 25-D Band Aperiodicities (BAP) with delta, delta-delta
 - 1-D fundamental frequency (F0) (linearly interpolated) with delta, delta-delta
 - 1-D voiced/unvoiced binary feature
 - In total 259 dimension

Experimental setups

- Neural network architecture
 - 6 hidden layers, each layer has 1536 hidden units
 - Tangent activation function for hidden layers, linear activation function for output layer
- Data normalisation
 - Vocoder parameters: speaker-dependent normalisation to zero mean and unit variance
 - Linguistic features: normalised to [0.01 0.99] on the whole database

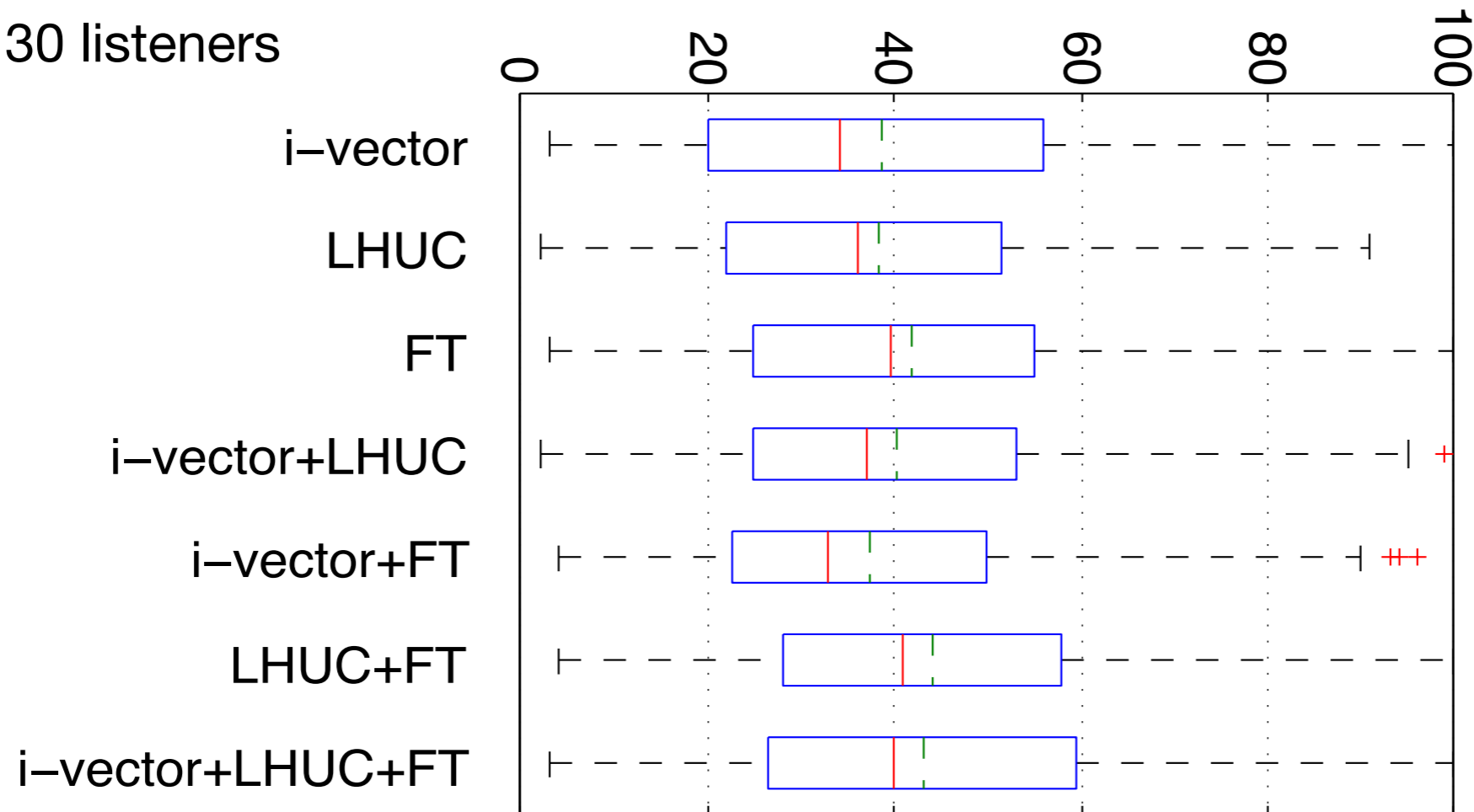
Experimental setups (cont'd)

- Baseline HMM system
 - The open-source HTS toolkit, and the best the setting on our dataset
 - CSMAPLR adaptation algorithm
- Adaptation
 - i-vector
 - background model: voice bank database
 - i-vector dimension: 32
 - Toolkit: ALIZE
 - LHUC
 - applied to all the hidden layers
 - Feature transformation
 - Joint density Gaussian mixture model based voice conversion

Subjective results – DNN adaptation methods

- Naturalness

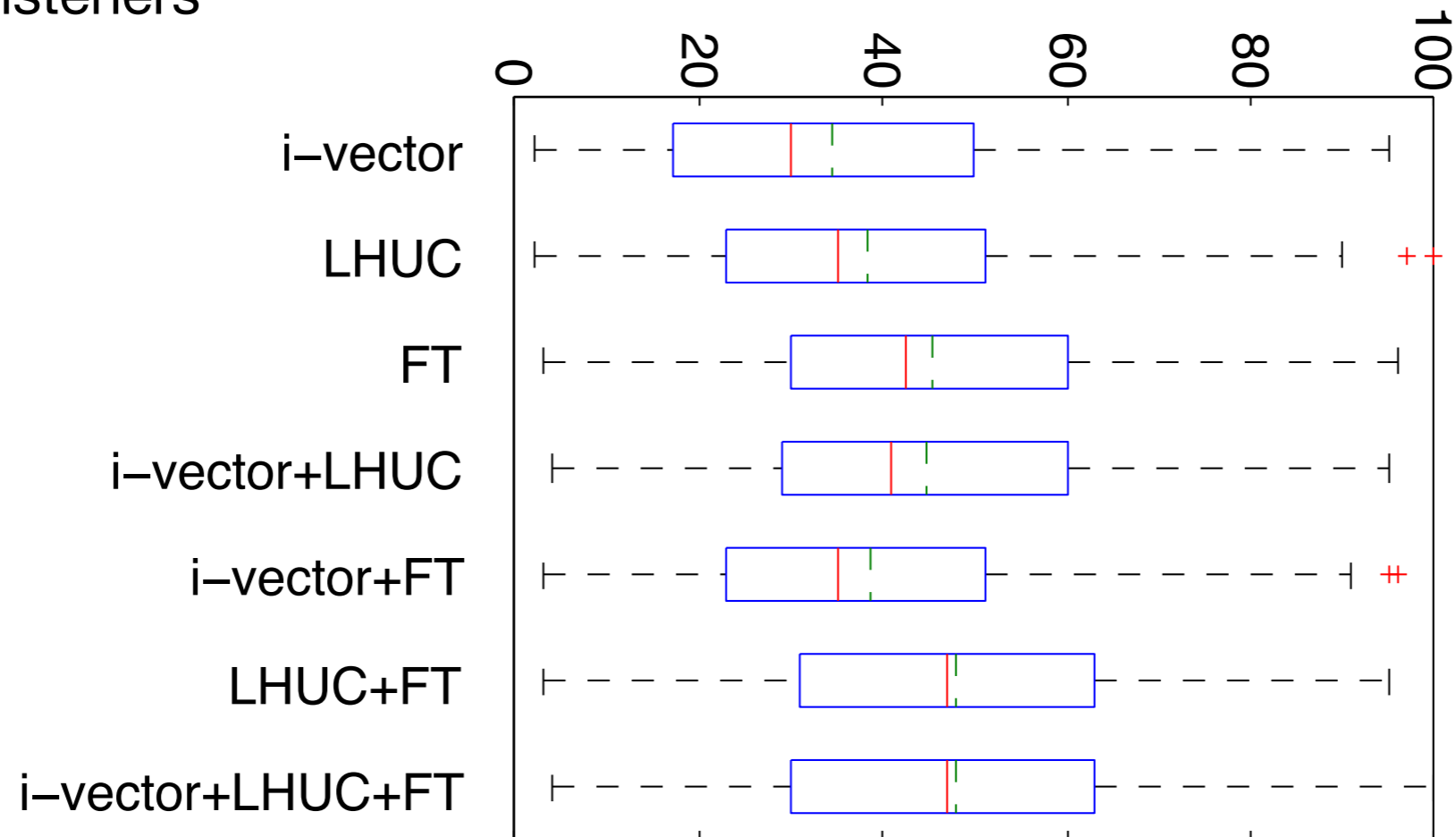
- MSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test
- 30 listeners



only i-vector+LHUC+FT vs LHUC+FT, and LHUC vs i-vector+LHUC are not significantly different

Subjective results – DNN adaptation methods

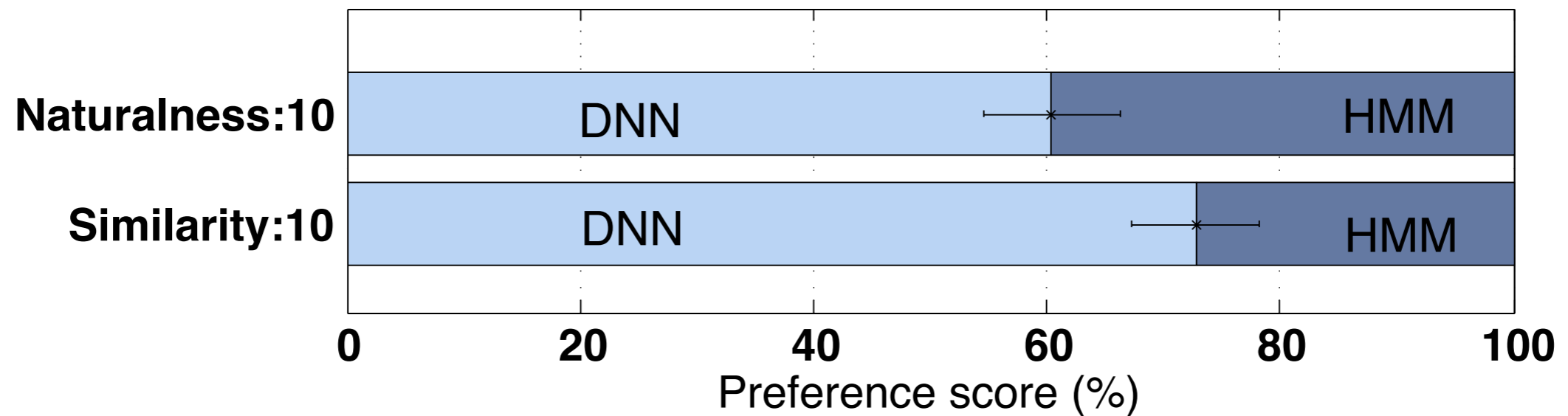
- Similarity
 - 30 listeners



only i-vector+LHUC+FT vs LHUC+FT, FT vs i-vector+LHUC and LHUC vs i-vector+FT are not significantly different

Subjective results — DNN vs HMM

- Preference test
 - 30 native English speakers



Conclusions

- Adaptation for DNN-based synthesis can be applied at three different levels
- The performance of DNN adaptation is significantly better than HMM adaptation
- Future work
 - Speaker adaptive training for the average DNN model
 - Joint optimisation of adaptation at three different levels

All the samples used in the listening tests are available at:
<http://dx.doi.org/10.7488/ds/259>