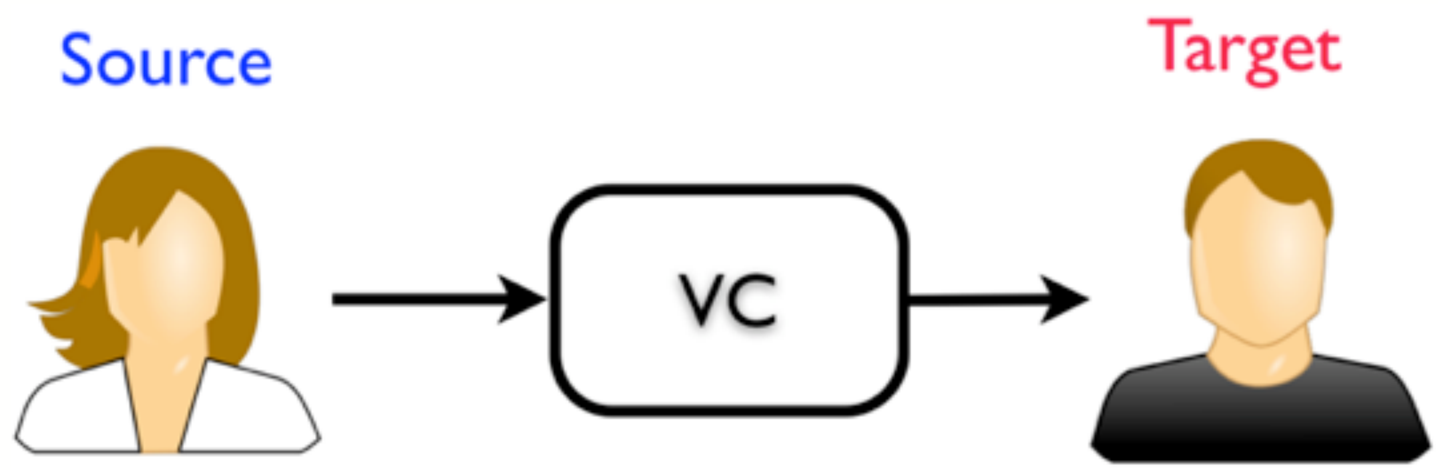
Semi-supervised Training of a Voice Conversion Mapping Function using a Joint-Autoencoder

Hamid Mohammadi
10/01/2015
CSLU OHSU

Voice Conversion

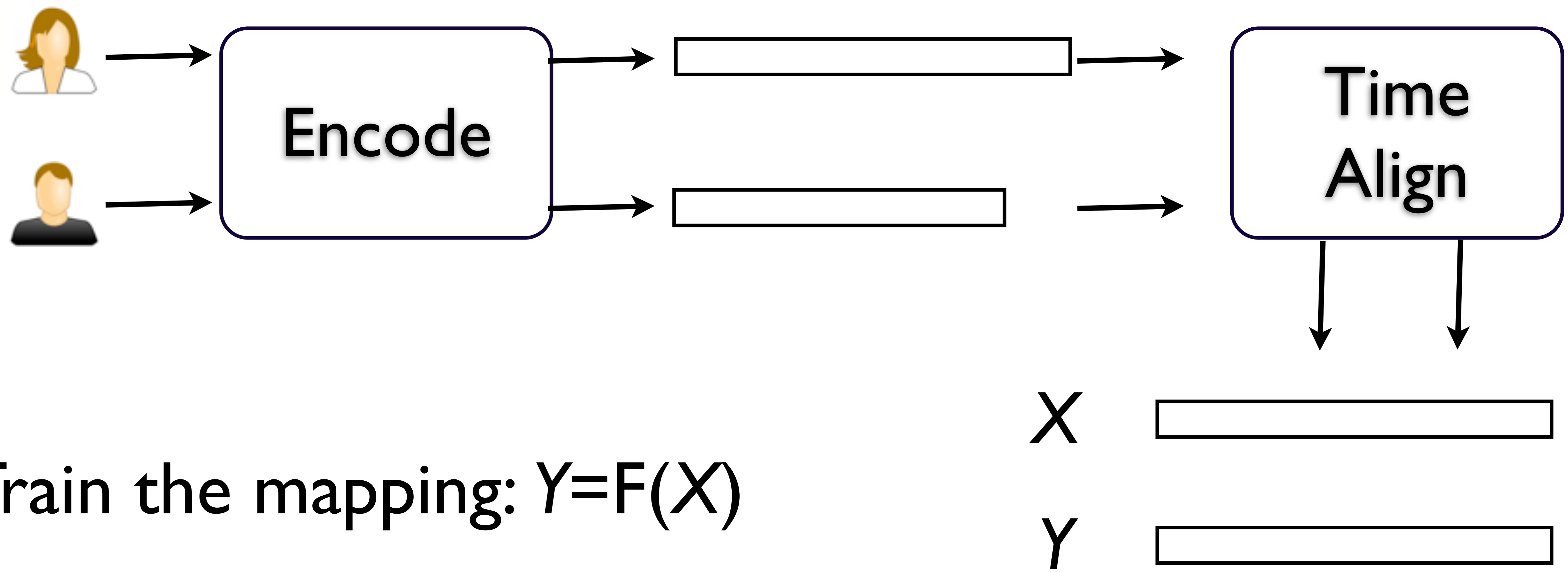
- Voice Conversion (VC): Processing a source speaker's speech to sound like a target speaker



Voice Conversion

- A typical VC system:
  - Given source and target speakers' training waveforms, extract MCEP features
  - Align source features, *X*, and target features, *Y*
  - Train a mapping that predicts *Y* from *X*
  -

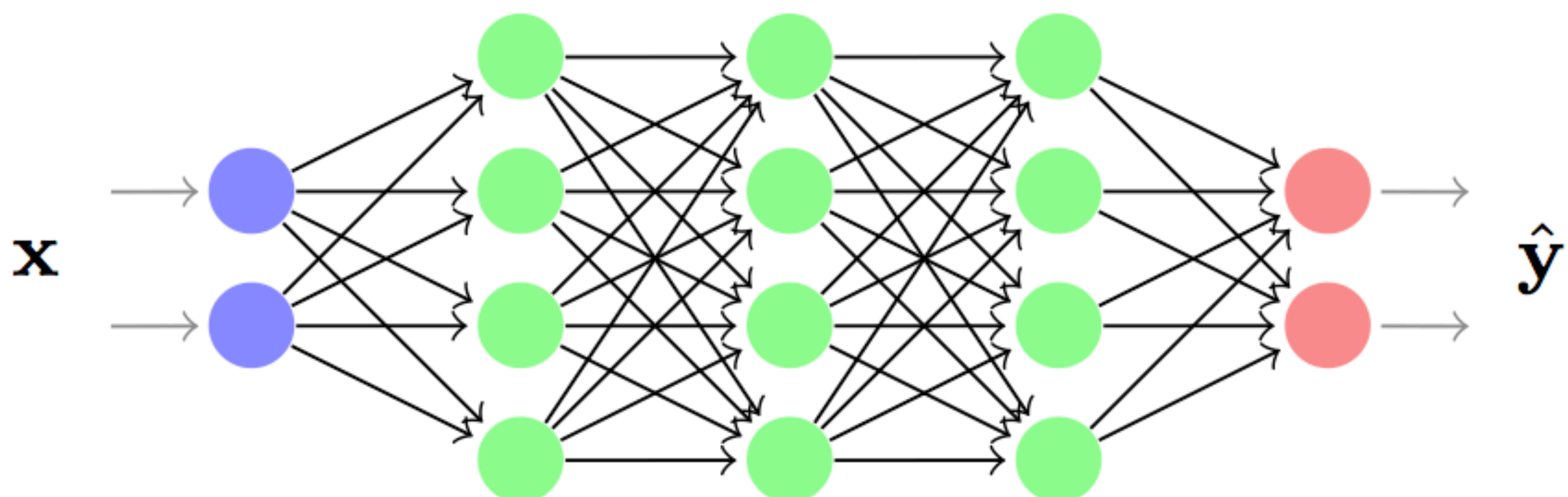
Voice Conversion



Train the mapping: *Y*=F(*X*)

VC approaches

- The mapping can be various approaches:
  - GMM
  - DNN



DNN-based VC

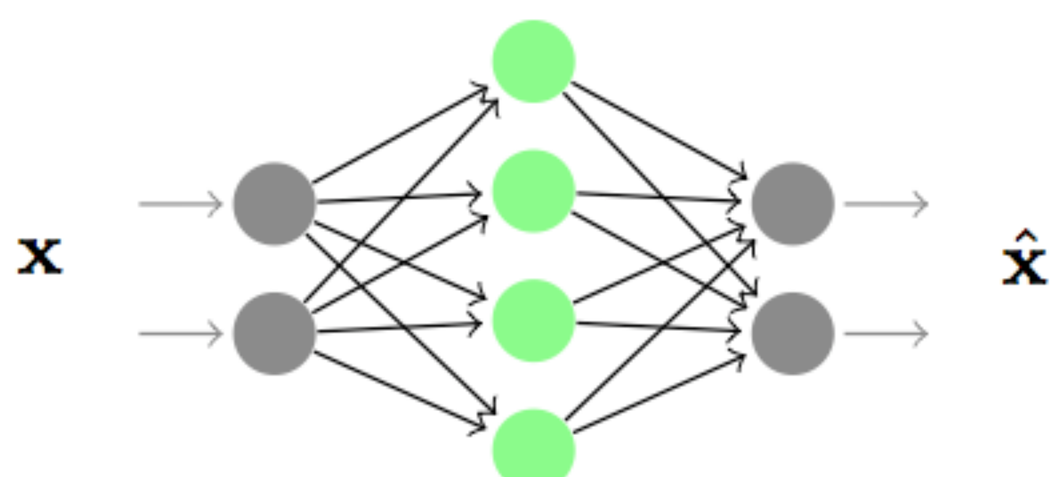
- Challenges:
  - Challenging to train multiple layers
  - Gradient-fading
  - Random Initialization
  - Local minima

DNN-based VC

- Proposed Solution, Part 1:
  - Use a lot of unseen data to “pre-train” the DNN
  - The pre-training would help to capture general spectral patterns
  - We used all TIMIT speakers to pre-train the DNN

Autoencoder

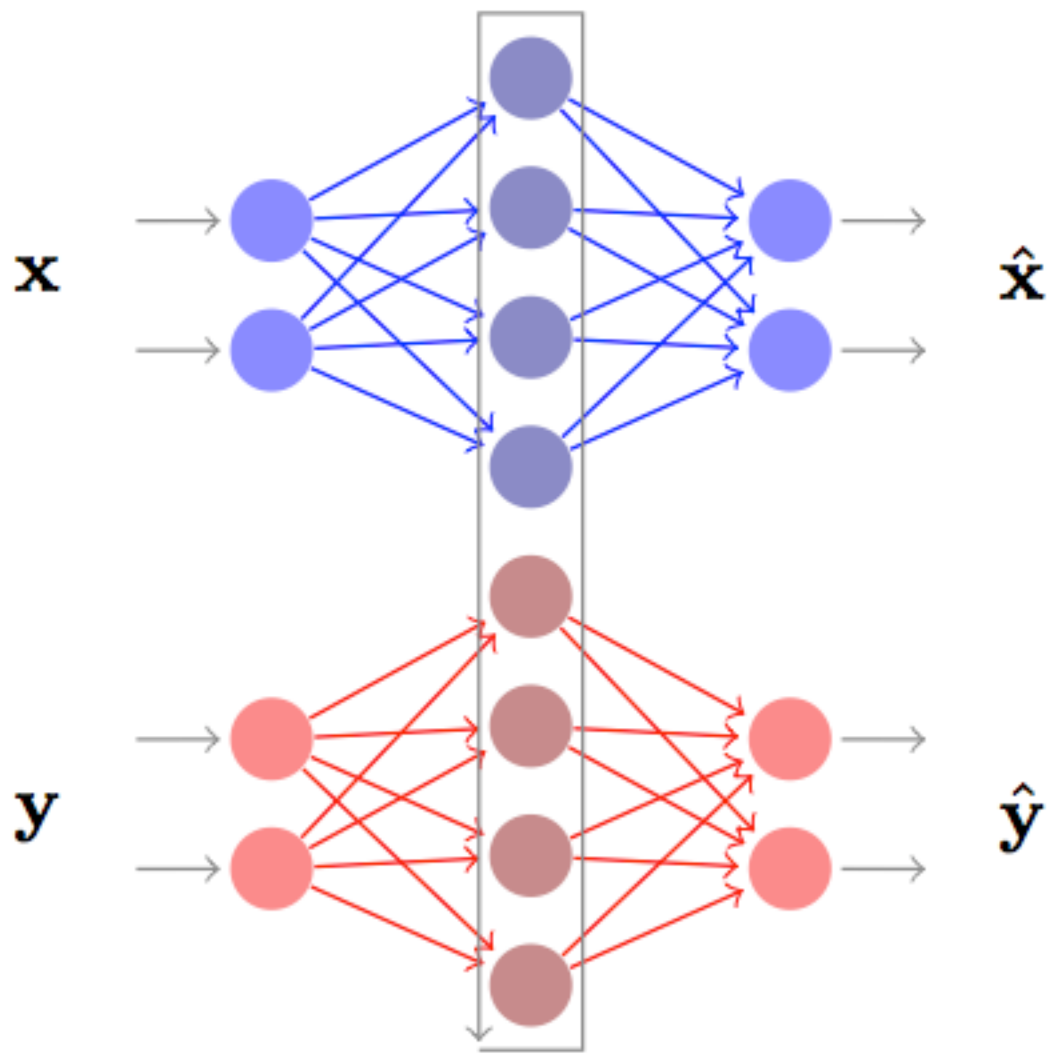
- The pre-training is done using a stacked autoencoder
- Each layer is trained using an autoencoder and then they are stacked together



Joint Autoencoder

- Proposed Solution, Part 1:
  - We propose a new architecture
  - The goal is to train two separate autoencoder
  - The autoencoders are joined by the encoding layer, using the cost function
  - Goal: The two autoencoders have similar values

Joint Autoencoder



$$\mathbf{h}_x = f_{hid}(\mathbf{W}\mathbf{x} + \mathbf{b}_{hid})$$
$$\hat{\mathbf{x}} = f_{vis}(\mathbf{W}^\top \mathbf{h}_x + \mathbf{b}_{vis})$$

$$\mathbf{h}_y = f_{hid}(\mathbf{V}\mathbf{y} + \mathbf{c}_{hid})$$
$$\hat{\mathbf{y}} = f_{vis}(\mathbf{V}^\top \mathbf{h}_y + \mathbf{c}_{vis})$$

Joint Autoencoder

- Source AE

$$\mathbf{h}_x = f_{hid}(\mathbf{W}\mathbf{x} + \mathbf{b}_{hid})$$
$$\hat{\mathbf{x}} = f_{vis}(\mathbf{W}^\top \mathbf{h}_x + \mathbf{b}_{vis})$$

- Target AE

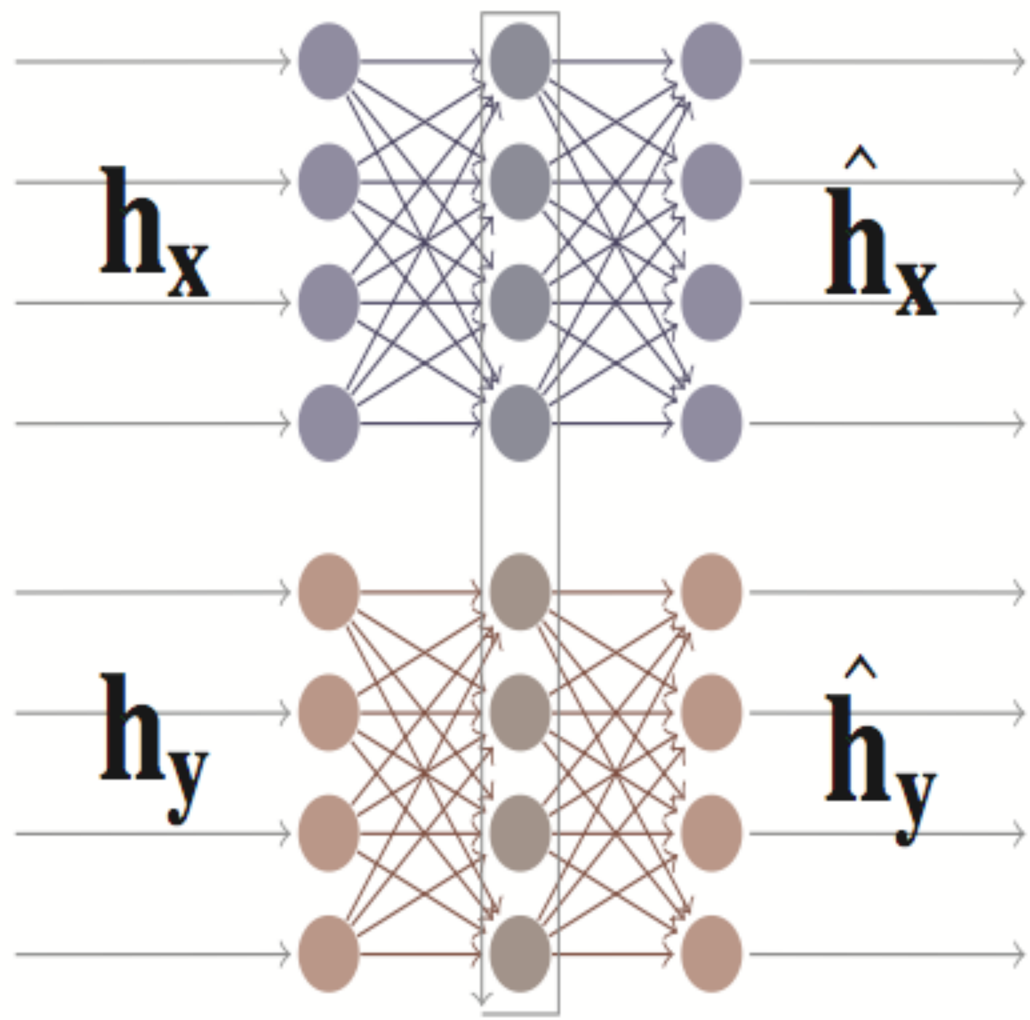
$$\mathbf{h}_y = f_{hid}(\mathbf{V}\mathbf{y} + \mathbf{c}_{hid})$$
$$\hat{\mathbf{y}} = f_{vis}(\mathbf{V}^\top \mathbf{h}_y + \mathbf{c}_{vis})$$

- Cost function: Reconstruction cost + hidden layer similarity

$$E = \alpha \left\|\mathbf{x} - \hat{\mathbf{x}}\right\|^2 + \alpha \left\|\mathbf{y} - \hat{\mathbf{y}}\right\|^2 + (1 - \alpha) \left\|\mathbf{h}_x - \mathbf{h}_y\right\|^2$$

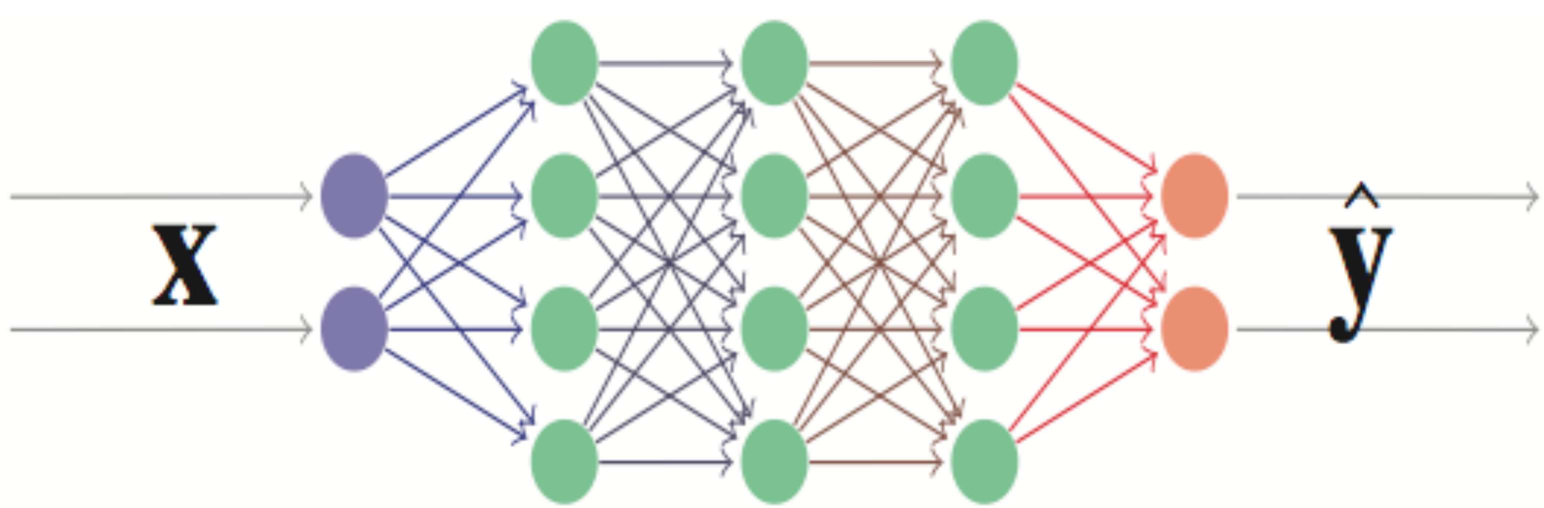
Joint Autoencoder

- Second Layer



Joint Autoencoder

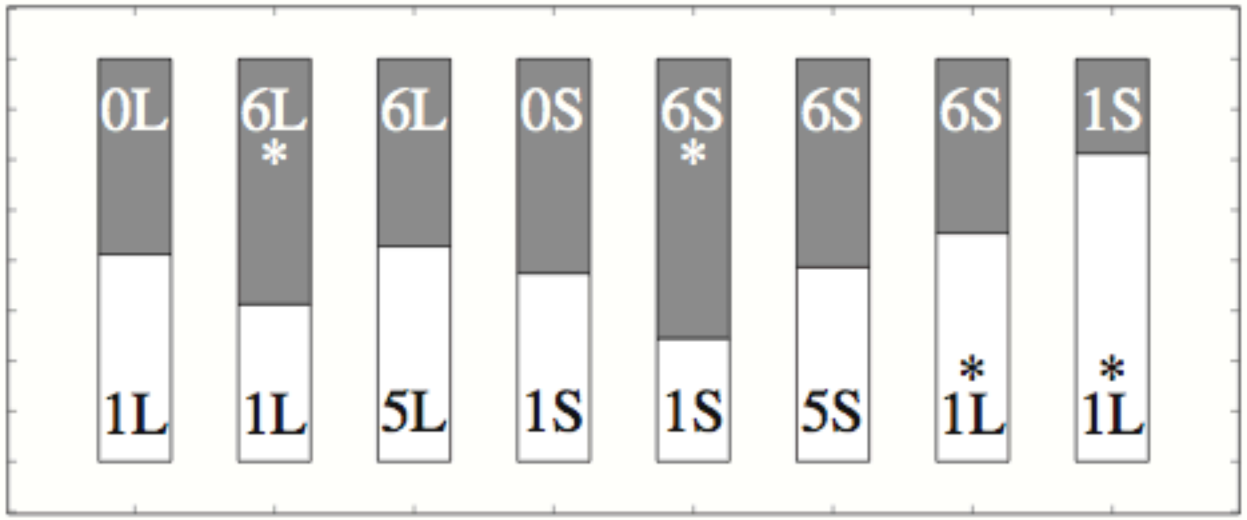
- Initialize the DNN using the joint Autoencoder weights



Evaluation: Quality

- Four CMU-arctic speakers for VC
- Two Conversions: CLB-to-SLT (females), and RMS-to-BDL (males)
- Small (S)/Large (L) training set: 5/100 sentences
- Amazon Mechanical Turk listeners evaluate
- Total of 40 listeners, each evaluating 20 sentence pairs
- Comparative MOS scores, from much worse (-2) to much better (+2)

Evaluation: Quality

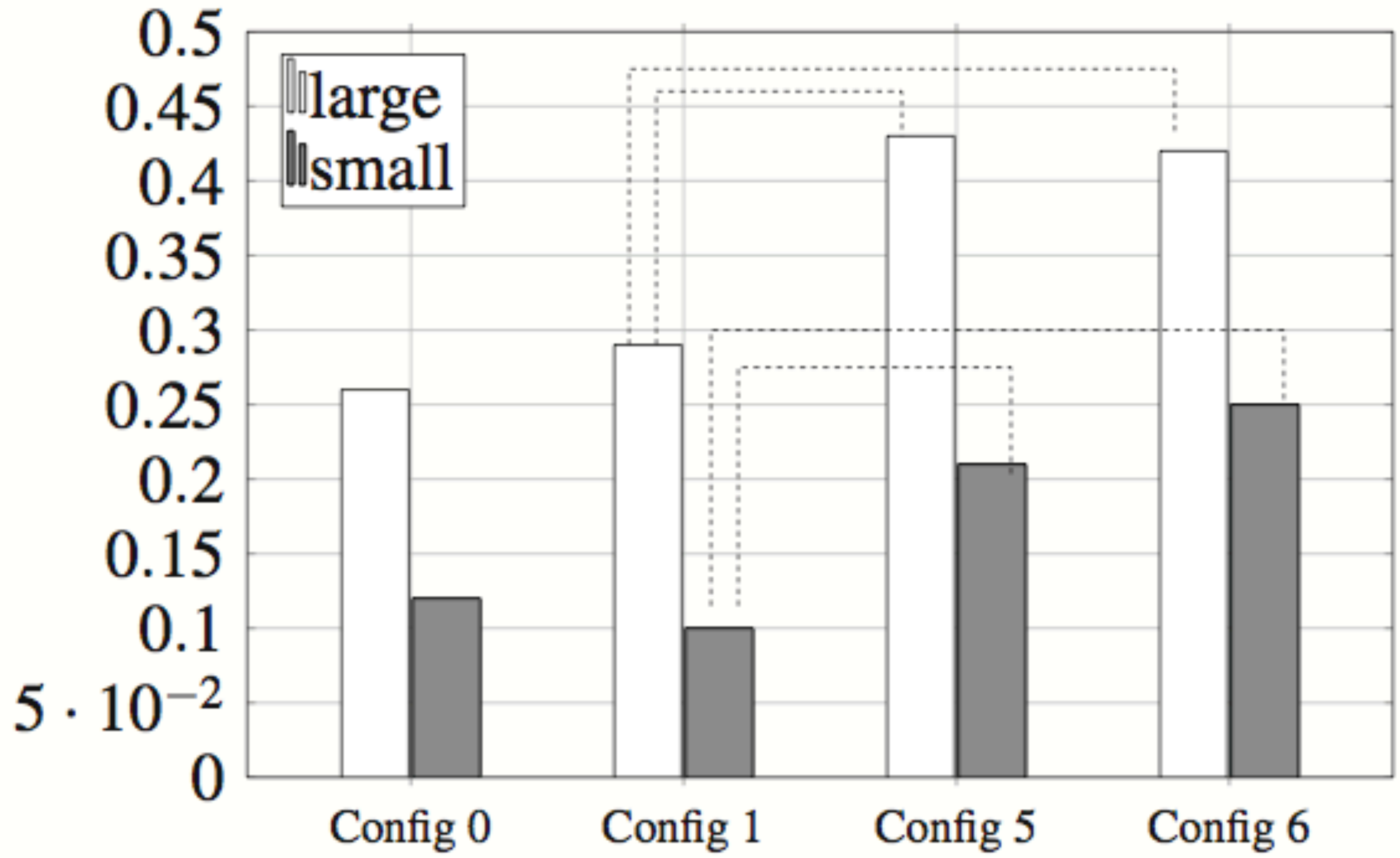


- Configurations: **(0)** GMMs with 1 frame, **(1)** DNN with 1 frame, **(5)** DNN pre-trained with 15 frames, **(6)** DNN pre-trained with 1 frame

Evaluation: Similarity

- Total of 40 listeners, each evaluating 48 sentence pairs
- Listeners hear two stimuli and score whether they are uttered by the same speaker, from definitely (+2) to definitely not (-2)
- Same case: we play converted target and real target, we hypothesize *positive* scores
- Diff case: we play converted target and a different speaker (with same gender as target), we hypothesize *negative* scores
- Final score is *same-score* − *diff-score*

Evaluation: Similarity



Configurations: (**0**) GMMs with 1 frame, (**1**) DNN with 1 frame, (**5**) DNN pre-trained with 15 frames, (**6**) DNN pre-trained with 1 frame

Questions?