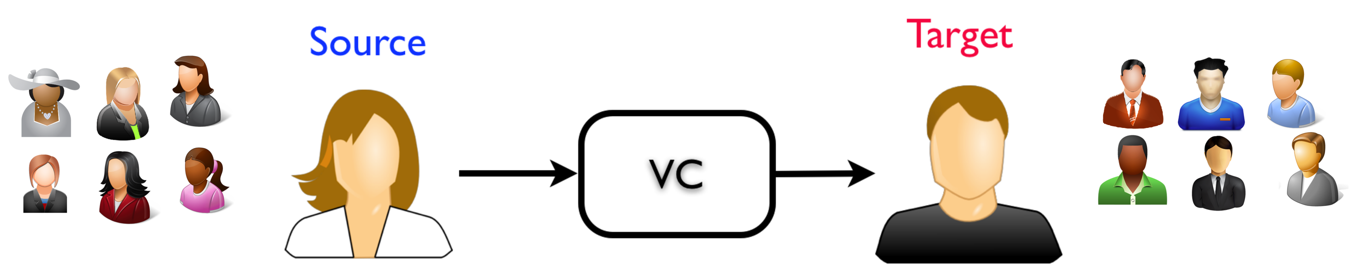
SEMI-SUPERVISED TRAINING OF A VOICE CONVERSION MAPPING FUNCTION USING A JOINT-AUTOENCODER

Seyed Hamidreza Mohammadi, Alexander Kain
Oregon Health & Science University

OREGON HEALTH & SCIENCE UNIVERSITY

VOICE CONVERSION

- Processing a *source* speaker's speech to sound like a *target* speaker
- A typical Voice Conversion system:
  1. Given source and target speakers' training sentences, extract MCEP features
  2. Parallelize and align source features, X, and target features, Y
  3. Train a mapping function $\mathcal{F}$ that predicts Y from X
  4. Given a test sentence, extract features X_{test}
  5. Map X_{test} using the mapping function to $\hat{Y}_{test} = \mathcal{F}(\mathcal{X})$
  6. Synthesize a new waveform from $\hat{Y}_{test}$



- Questions:
  - Does a semi-supervised approach improve VC performance?
  - Does using multiple frames improve performance?
- Approach:
  - We propose to first train a deep autoencoder on unlabeled TIMIT speakers and use those weights as part of pre-training a DNN mapping.
  - We propose to find several similar speaker to each source and target speakers to pre-train the mapping function.
  - We also propose a new learning structure called Joint-Autoencoder.

JOINT AUTOENCODER

- We can train two separate autoencoders on the source and target speakers' features — the source encodings and the target encodings are unlikely to be correlated
- We propose to maximize the similarity of the encoding values and thus reduce the complexity
- The Joint-Autoencoder (JAE) consists of two Autoencoders (AEs):
  - 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

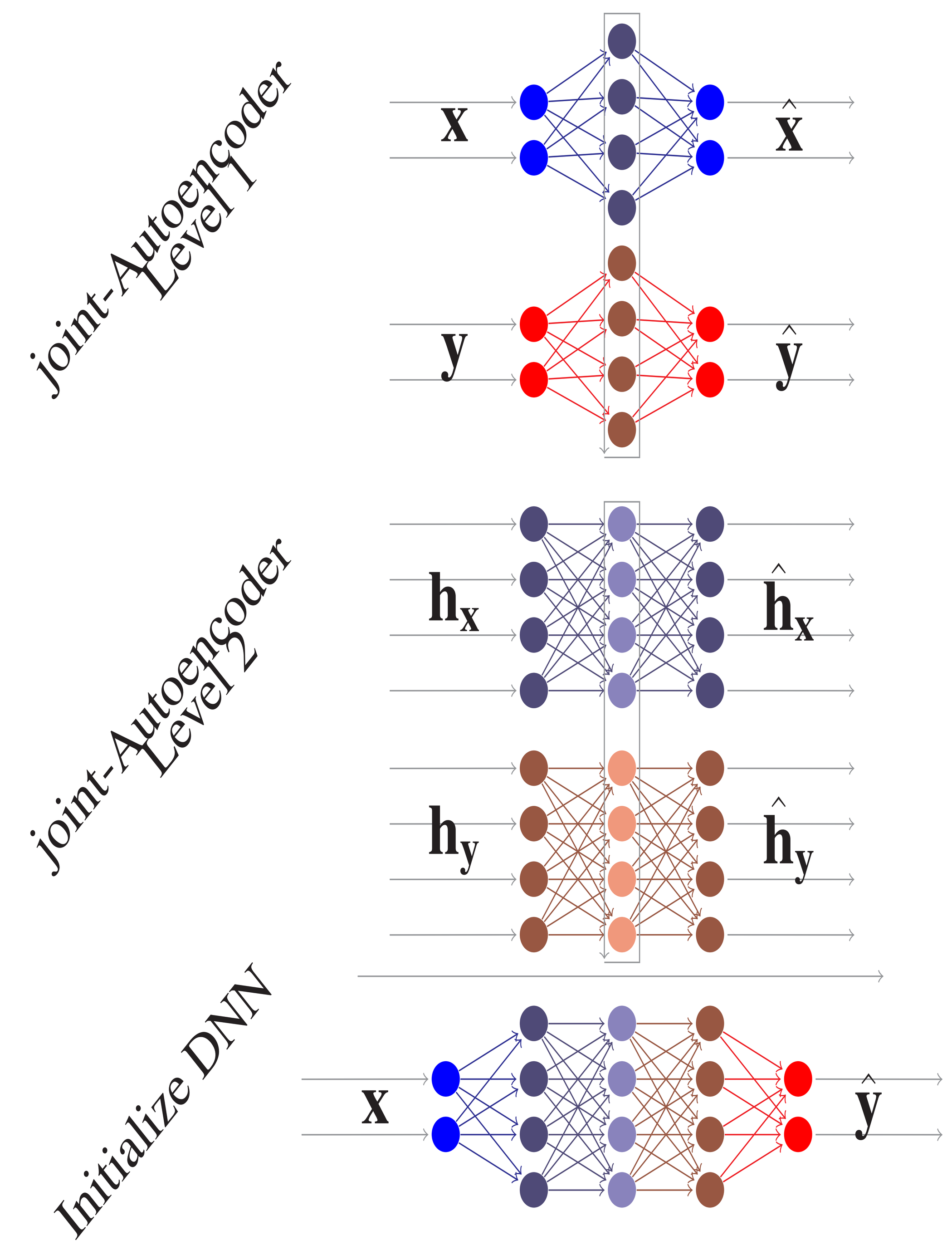
$$\mathbf{h}_y = f_{hid}(\mathbf{V}\mathbf{y} + \mathbf{c}_{hid})$$
$$\hat{\mathbf{y}} = f_{vis}(\mathbf{V}^\top \mathbf{h}_y + \mathbf{c}_{vis})$$

- Cost function: Reconstruction cost + hidden layer similarity

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

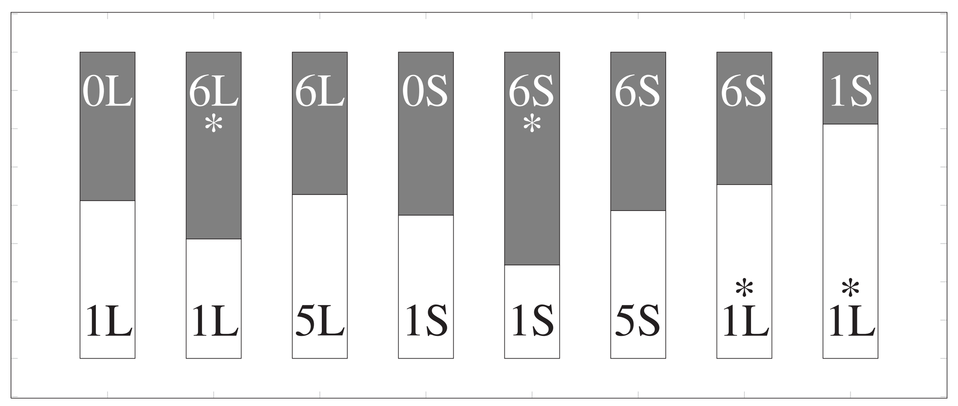
- For phonetically similar speech segments from source and target speakers, the encoding values from the two AEs are similar.
- Encoding layers of the source AE is followed by the Decoding layers of the target AE to initializes the DNN.

STACKED JOINT AUTOENCODER (SJAE)



EXPERIMENT: SPEECH QUALITY

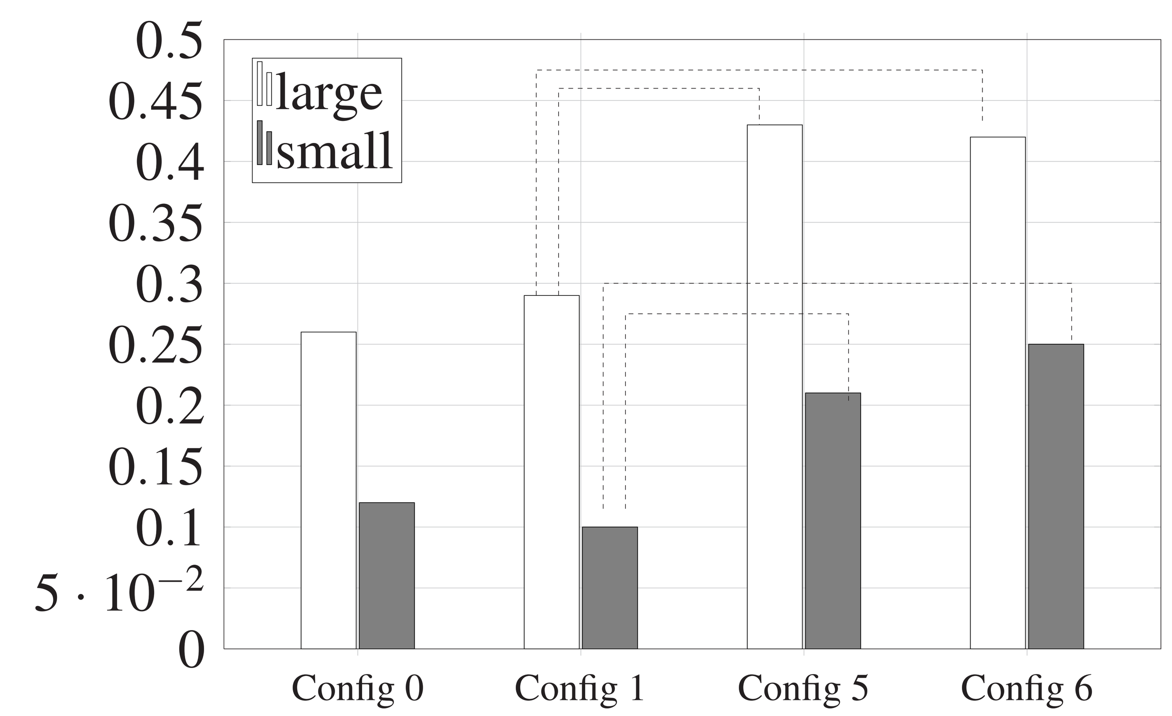
- All 630 TIMIT speakers for training a SAE
- Four CMU-arctic speakers for VC
- Two Conversions: CLB-to-SLT (females), and RMS-to-BDL (males)
- Small (**S**)/Large (**L**) training set: 5/100 sentences
- Amazon Mechanical Turk listeners evaluate
- Total of 40 listeners, each evaluating 20 sentence pairs
- Comparative MOS scores, from much worse (-2) to much better (+2)



- 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

EXPERIMENT: CONVERSION ACCURACY

- 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*



- multi-frame DNN performed better than single-frame DNN
- pre-trained DNN performed better than randomly-initialized DNN

CONCLUSIONS

- We proposed a Stacked-Joint-Autoencoder architecture, which aims to find a common encoding of parallel source and target features
- We found similar speakers in TIMIT corpus for source and target speaker
- We used this Stacked-Joint-Autoencoder to pre-train a DNN
- A pre-trained DNN performed better than a non pre-trained DNN for both quality and similarity
- We did not find a significant improvement in the subjective scores using multiple frames

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0964468.

Center for Spoken Language Understanding, Oregon Health & Science University — Portland, Oregon USA   email: mohammah@ohsu.edu, kaina@ohsu.edu   www: ohsu.edu/cslu