Joint-Autoencoder (JAE) consists of two Autoencoders (AEs): Source AE

\[ h_s = f_{sad}(W x + b_{sad}) \]
\[ x = f_{sad}(W h_s + b_{sad}) \]

Target AE

\[ h_t = f_{tad}(V y + c_{tad}) \]
\[ y = f_{tad}(V h_t + c_{tad}) \]

Cost function: Reconstruction cost + hidden layer similarity

\[ E = \alpha \| x - \hat{x} \|^2 + \alpha \| y - \hat{y} \|^2 + (1 - \alpha) \| h_t - \hat{h}_t \|^2 \]

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 initialize the DNN.

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 speakers to each source and target speakers to pre-train the mapping function.
- We also propose a new learning structure called Joint-Autoencoder.

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

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

EXPERIMENT: CONVERSION ACCURACY

<table>
<thead>
<tr>
<th>Config</th>
<th>Small</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>1</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>0.30</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Total of 40 listeners, each evaluating 48 sentence pairs

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