We propose to first train a deep autoencoder on unlabeled speech data to create a compact representation of features. Then, we fine-tune the weights of the autoencoder using back-propagation to improve the mapping.

Question: Can we apply a semi-supervised approach to improve the supervised learning of the mapping?

Approach: We propose to first train a deep autoencoder on unlabeled speech data and use those weights as part of pre-training a DNN mapping.

**EXPERIMENT: SPEECH QUALITY**

- Total of 40 listeners, each evaluates 20 sentence pairs
- Comparative MOS scores, from much worse (-2) to much better (+2)
- Arrows point to the better performing configuration
- VOC: Vocoder speech (without modifying parameters)

**EXPERIMENT: CONVERSION ACCURACY**

- The significant differences are shown by green lines marked by asterisk
- FS-Large performs better than GMM-Large
- DNN-Small performs better than GMM-Small
- DNN-Small is performing similar to DNN-Large

**CONCLUSIONS**

- We created an Autoencoder from unlabeled speech data
- We used this Autoencoder to pre-train a DNN
- A pre-trained DNN trained with 2 sentences performed similarly to a GMM trained with 70 sentences
- Frame Selection performs best with 70 training sentences

**ACKNOWLEDGEMENT**

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