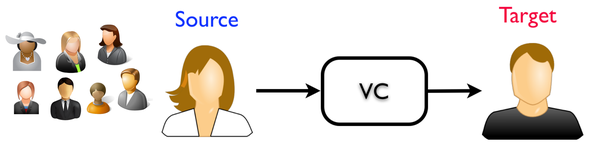
VOICE CONVERSION USING DEEP NEURAL NETWORKS WITH SPEAKER-INDEPENDENT PRE-TRAINING

Seyed Hamidreza Mohammadi, Alexander Kain
Oregon Health & Science University

VOICE CONVERSION PROBLEM

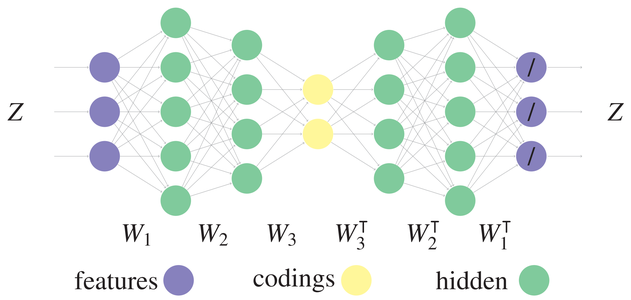
- Voice Conversion (VC): How to make a source speaker's speech sound like a target speaker
- VC procedure:
  - Analyze speech and get features (MCEP)
  - Align source features, X, and target features, Y
  - Map X to Y
  - Synthesize using Y
- Different mappings:
  - Gaussian Mixture Models (GMM)
  - Frame Selection (FS), which is similar to unit-selection in Text-to-speech Synthesis
  - Deep Neural Networks (DNN)



- Question: Can we apply an semi-supervised approach to improve the supervised learning of the mapping?
- Approach: We propose to first train a deep autoencoder on unlabeled speakers Z and use those weights as part of pre-training a DNN mapping.

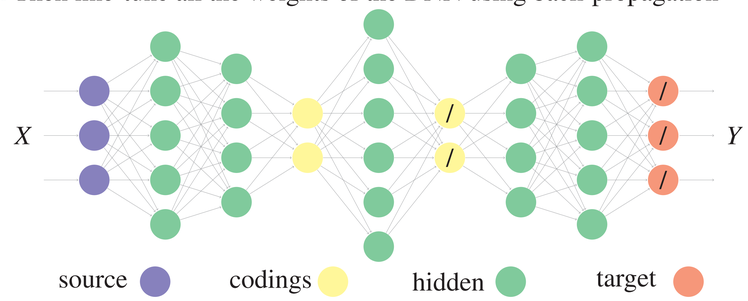
AUTOENCODER

- Train the AE to compute a compact representation of features
- The deep AE is trained layer-wise
- Input layer: 24^{th} order MCEP
- AE 1: Denoising AE with coding layer of **100**
- AE 2: Contractive AE with coding layer of **40**
- AE 3: Contractive AE with coding layer of **15**
- The encoding and decoding weights of AEs are tied
- The activation functions are sigmoid (except the last decoding layer)



THE DEEP NEURAL NETWORK

- First build a DNN by plugging in the AE encoding/decoding weights
- Add two middle layers in between the encoding and decoding layer
- Train the middle layers using back-propagation while keeping the AE weights fixed, i.e. we train on compact features
- Then fine-tune all the weights of the DNN using back-propagation

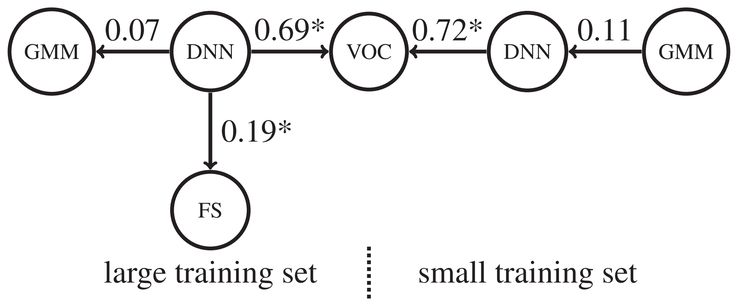


EXPERIMENTS

- Seven speakers for training the AE, approx. 2 hours of speech total
- Four speakers for VC: two male (M1, M2), two female (F1, F2)
- Four Conversions: two intra-gender (M1 $\rightarrow$ M2, F2 $\rightarrow$ F1) and two cross-gender (M2 $\rightarrow$ F2, and F1 $\rightarrow$ M1)
- Small / Large training set: 2 / 70 sentences
- Amazon mechanical turk listeners evaluate

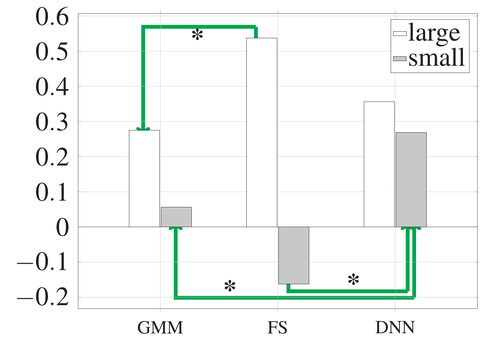
EXPERIMENT: SPEECH QUALITY

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



EXPERIMENT: CONVERSION ACCURACY

- Total of 40 listeners, each evaluates 40 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 more positive scores
- Diff case: we play converted target and a different speaker (with same gender as target), we hypothesize more negative scores
- $score = -1 \times diff\text{-}score + same\text{-}score$



- 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.