A Neural Network Based Nonlinear Feature Transformation for Speech Recognition

Hongbing Hu, Stephen A. Zahorian
Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902, USA
hongbing.hu@binghamton.edu, zahorian@binghamton.edu

Introduction
- Accurate Automatic Speech Recognition (ASR)
  - Highly discriminative features
    - Incorporate nonlinear frequency scales and time dependency
  - Efficient recognition models
    - HMMs: good time alignment capability and convenient mechanisms for incorporating language models
    - Neural Networks: good discriminatory power
- Nonlinear Feature Transformation for Speech Recognition
  - Combination of Neural Networks and HMMs
  - A neural network based Nonlinear Principal Component Analysis (NLPCA) is used as a dimensionality reduction approach for speech features

Experimental Evaluation
- Database
  | TIMIT database | Reduced 39 phone set mapped from the TIMIT 62 phone set |
  | Target | Training data | Testing data |
  | 4620 sentences (460 speakers) | 1680 sentences (168 speakers) |

NLPCA for HMM Recognition
- Nonlinear Principal Components Analysis (NLPCA)
  - Based on a bottleneck neural network
  - Activations from the middle hidden layer are used as the reduced dimensionality data
  \[ x \rightarrow \phi(x) \]
  \[ \phi: \mathbb{R}^n \rightarrow \mathbb{R}^m \] M dimension feature space
- Phoneme HMMs for Phonetic Recognition
  - Dimensionality reduced features are recognized as phonemes using HMMs with Gaussian Mixture Model
- Parameters of the HMMs are trained by the Baum-Welch algorithm, independent of the NLPCA training

Experiment 1
- NLPCA was evaluated with various dimensions in the reduced feature space
- HMMs were trained with 1, 3 and 5 states, and with 1, 2, 5 and 10 mixes per state
- Recognition accuracies of 1-state (left) and 3-state HMMs (right) with various reduced feature dimensions
  - For both the 1-state and 3-state HMMs, best accuracy was obtained with feature dimensionality between 10 and 30
  - NLPCA is able to represent the complexity of original feature in a reduced dimensionality space
  - The reduced features result in high accuracy using a small number of mixes and states in HMMs

Experiment 2
- Various block lengths in DCTC-DCSC feature calculation were evaluated for optimal block length
- 20-dimensional NLPCA features were used to compare with the original 91-dimensional features
- Recognition accuracies of 1-state (left) and 3-state HMMs (right) using the original and reduced features
  - NLPCA features are better able to represent speech information over longer segment
  - NLPCA can account for some of the temporal information accounted with HMMs, thus potentially simplifying the HMM configuration

Experiment 3
- 50% of the training data was used for the NLPCA training and the other 50% of the data for HMMs
- Neural network in NLPCA
  - 3 hidden-layers with 500 nodes in the first and third hidden layers and varied nodes in the second layer
  - Input layer with 91 nodes and output layer with 39 nodes
- Comparing results with full training data, the best 1 state HMMs results are 2% lower and the best 3 state results are slightly lower using the partitioned training data
  - Only a small degradation due to reduced size of the training data

Conclusions
- A neural network based nonlinear feature transformation (NLPCA) is incorporated with an HMM recognition model for continuous speech phonetic recognition
- Recognition accuracies with NLPCA reduced dimensionality features are higher than that with original features, especially for a small number of states and mixtures
- NLPCA features are able to well represent spectral-temporal information in segments as long as 200ms, thus potentially reducing HMM model complexity
- The entire recognition system could benefit from low dimensional features in terms of processing time and recognition accuracy