HYBRID HMM/MLP SPEECH RECOGNITION

Horacio Franco, Victor Abrash, Michael Cohen
Ananth Sankar, Mitch Weintraub

Speech Technology and Research Program
SRI International
Menlo Park, California

OVERVIEW

- **Port of Hybrid HMM/MLP Technology to New Tasks**

  Experimental Results on Switchboard Credit Card Transcription Task:
  - Develop hybrid MLP/Genonic models.
  - Explore MLP architecture parameters to increase performance.

- **Speaker Adaptation in WSJ Dictation Task**
  - Develop supervised adaptation using different transformation models.
  - Explore straightforward adaptation of speaker independent MLP.
  - Explore combined adaptation.

- **Summary**

DESCRIPTION OF SRI'S LVCSR SYSTEM

- **Large Vocabulary Recognition System**
  - NNet experiments use 5,000 words.
- **Gender-Specific Acoustic Models**
- **Genonic Acoustic Models (Digalakis '93)**
  - Automatically balances the resolution/robustness trade-off to optimize performance.
- **Hybrid MLP/HMM Acoustic Models (Cohen '92)**
  - Discriminative training.
  - Capturing feature correlation within & across frames.
- **Progressive-Search Strategy (Murveit '93)**
  - Multiple-pass strategy using simpler/faster initial models to prune search for later passes.
- **Fast Large-Vocabulary Algorithms (Murveit '93)**
  - Use lexical tree for back-off to unigrams.
  - Gaussian clustering, Shortlists.

SRI International

SWITCHBOARD TRAINING/TESTING

- **Acoustic Training**
  - Credit card subset of Switchboard.
    - 2.6 hours male speech.
  - SRI Partyline (predecessor of Macrophone).
    - Read TIMIT, ATIS, WSJ0 sentences from 2600 talkers.
    - 11.0 hours male speech.
- **Language Training**
  - 2200 conversations.
  - 3.2 M words of training text.
- **Development Test Set**
  - Credit card task.
  - 1065 sentences from 11 male speakers.

HYBRID HMM/MLP RESULTS

- **Results with 5K Bigram Grammar and Non-Crossword Models.**
- **CD-MLP Uses 9 Generalized Left and Right Context Classes.**

System	# Input Frames	# Hidden Units	Word Error
Baseline Genone System	-	-	54.7
CI Hybrid MLP/Genone	9	1000	53.5
CI Hybrid MLP/Genone	17	2000	53.0
CI Hybrid MLP/Genone	27	3000	52.6
CD Hybrid MLP/Genone	9	1000	52.3
CD Hybrid MLP/Genone	17	2000	52.1
CD Hybrid MLP/Genone	27	3000	51.2

- **Combining Genone and MLP Produces**
  - 3.5% Absolute Improvement,
  - 6.4% Relative Improvement.
- **Increasing MLP Input Window, Hidden Layer, Context-Dependence Helped Performance.**
- **Cross-Word Experiments In Progress.**

SPEAKER ADAPTATION

- **Transformation-Based Approach (Watrous '93)**

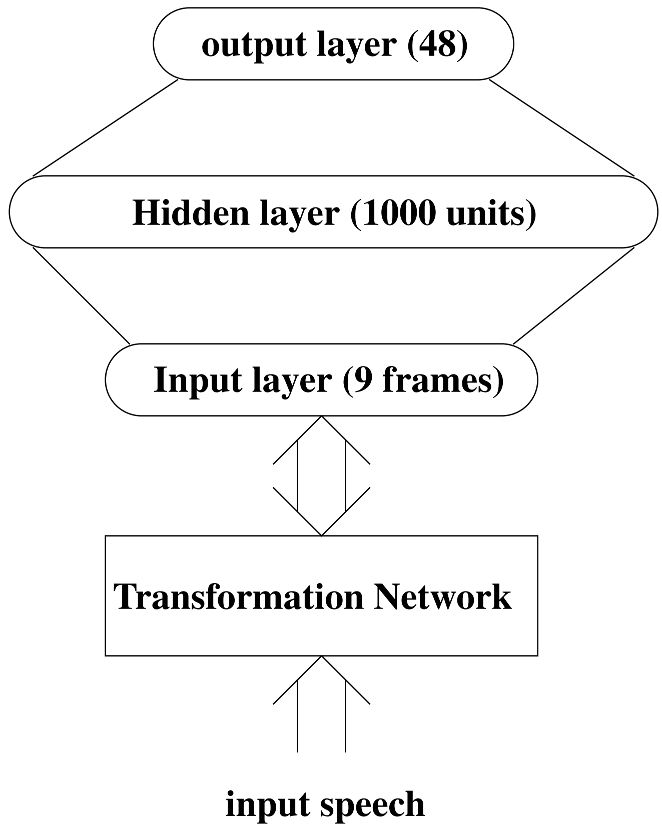
  Motivations:
  - Transform SI models to better match new speaker.
  - Separate sub-networks for modeling speaker dependent and speaker independent characteristics.
  - More reliable training of few parameters.
  - Multiple speaker compact representation.

  Characteristics:
  - Same transformation applied to all input frames.
  - Small number of parameters to adapt.

- **Cross-validation-based Adaptation of the MLP**
  - Comparison with transform-based approach.

SRI International

ADAPTATION METHOD



- Supervised adaptation from HMM forced alignments.
- Cross-validation-based Training.
- Transformation is tied across input frames.
- Adapt the transformation network only.
- Adapt MLP and transformation network.

SRI International

PRELIMINARY ADAPTATION RESULTS

- **30 sentences adaptation**
- **10 sentence cross-validation**
- **40 sentences test**
- **Context-independent MLP/HMM hybrid.**

Word error rate on 5 non-native speakers from WSJ dictation task

Experiment	4n0	4n3	4n5	4n9	4na	avg
Hybrid CI baseline	54.1	39.2	31.6	24.9	20.1	34.0
$y_i=a_i x_i+b_i$	43.3	23.3	28.8	21.0	18.3	26.9
y=Ax+b	33.1	25.9	25.9	22.2	12.7	24.0
y=Ax+b and MLP	25.3	17.7	22.2	19.6	13.7	19.7
MLP	29.4	18.3	26.5	17.5	15.9	21.5

SRI International

NEXT STEPS

- Iterative joint training of MLP and transformation networks.
- Acoustic-region-dependent transformation.

SRI International

SUMMARY

- **Transcription Results on Switchboard Credit Card Task.**
  - Earlier hybrid HMM/MLP scheme helped in Switchboard corpus.
  - From 54.7% to 51.2% word error rate when including context-dependent MLP.
- **Speaker Adaptation Results in Wall Street Journal Dictation Task.**
  - MLPs with a large number of parameters can be adapted with a small amount of data using cross-validation.
  - Small transformation networks can capture speaker dependent characteristics.
- **Ongoing Work on Wall Street Journal Dictation Task and ATIS.**