Subproject II: Robustness in Speech Recognition
Members (1/2)

Hsiao-Chuan Wang (PI)
National Tsing Hua University

Jeih-Weih Hung (Co-PI)
National Chi Nan University

Sin-Horng Chen
National Chiao Tung University

Jen-Tzung Chien (Co-PI)
National Cheng Kung University

Lin-shan Lee
National Taiwan University

Hsin-min Wang
Academia Sinica
Members (2/2)

Yih-Ru Wang
National Chiao Tung University

Yuan-Fu Liao
National Taipei University of Technology

Berlin Chen
National Taiwan Normal University
Research Theme

Signal Processing

Feature Extraction & Transformation

Speech Decoding including Word Graph Rescoring

Adaptive HMM Models

Adaptive Pronunciation Lexicon

Adaptive Language Models

Input speech

Output Recognition Results

Signal Level

Lexical Level

Model Level
Research Roadmap

Current Achievements

• Speech enhancement & wavelet processing
• Cepstral moment normalization & temporal filtering &
• Microphone array and noise cancellation approaches
• Discriminative adaptation for acoustic and linguistic models
• Maximum entropy modeling & data mining algorithm
• Robust language modeling

Future Directions & Applications

• Speech recognition in different adverse environments, e.g. car, home, etc.
• Robust broadcast news transcription
• Lecture speech recognition
• Spontaneous speech recognition
• Next generation automatic speech recognition
• Powerful machine learning approaches for complicated robustness problems
Signal Level Approaches

- **Speech Enhancement**
  - Harmonic retaining, perceptual factor analysis, etc.

- **Robust Feature Representation**
  - Higher-order cepstral moment normalization, data-driven temporal filtering, etc.

- **Microphone Array Processing**
  - Microphone array with post-filtering, etc.

- **Missing-Feature Approach**
  - Sub-space missing feature imputation and environment sniffing, mismatch-aware stochastic matching, etc.
Higher-Order Cepstral Moment Normalization (HOCMN) (1/3)

- Cepstral Feature Normalization Widely Used
  - CMS: normalizing the first moment
  - CMVN: normalizing the first and second moments
  - HEQ: normalizing the full distribution (all order moments)
  - How about normalizing a few higher order moments only?
  - Disturbances of larger magnitudes may be the major sources of recognition errors, which are better reflected in higher order moments
Higher-Order Cepstral Moment Normalization (HOCMN) (2/3)

- **Experimental results**: Aurora 2, clean condition training, word accuracy averaged over 0~20dB and all types of noise (sets A, B, C)

![Graph showing experimental results](image)

- (a) HOCMN[1,N] (full-utterance)
- (b) HOCMN[1,N](L=86)

(1st and N-th moments normalized)

N (even integer)
Experimental Results: Aurora 2, clean condition training, word accuracy averaged over 0~20dB for each type of noise condition.

- HOCMN is significantly better than CMVN for all types of noise.
- HOCMN is better than HEQ in most types of noise except for the “Subway” and “Street” noise.
Data-Driven Temporal Filtering

- Developed filters were performed on the temporal domain of the original features.
- These filters can be derived in a data-driven manner according to the criteria of PCA/LDA/MCE.
- They can be integrated with Cepstral mean and variance normalization (CMVN) to achieve further performance.
Microphone Array Processing (1/3)

- Integrated with Model Level Approaches (MLLR)

Model Adaptation

Speech Input Using Microphone Array

- Delay Estimator
- Delay-and-Sum Beamformer

Initial HMM Parameters → MLLR Adaptation → Adapted HMM Parameters

Speech Enhancement

- Using Time Domain Coherence Measure (TDCM)
- Enhanced signal → Speech Recognition

Speech Recognition

Result
Microphone Array Processing (2/3)

- Further Improved with Wiener Filtering and Spectral Weighting Function (SWF)

\[
\begin{align*}
\hat{s} &= X_1 \times \hat{W} \\
\hat{X} &= \hat{s} \times \hat{W} \\
\bar{X} &= \hat{X} \times \bar{W}
\end{align*}
\]
Microphone Array Processing (3/3)

- **Applications for In-Car Speech Recognition**
  - Power Spectral Coherence Measure (PSCM) used to estimate the time delay

![Physical configuration](image)

![Configuration in car](image)
Model Level Approaches

- Improved Parallel Model Combination
- Bayesian Learning of Speech Duration Models
- Aggregate \textit{a Posteriori} Linear Regression Adaptation
Aggregate a Posteriori Linear Regression (AAPLR)  (1/3)

- Discriminative Linear Regression Adaptation
- Prior Density of Regression Matrix is Incorporated to Construct Bayesian Learning Capabilities
- Closed-form Solution Obtained for Rapid Adaptation

AAPLR

Discriminative criterion

Prior information of regression matrix

Bayesian Learning

Closed form solution
Aggregate a Posteriori Linear Regression (AAPLR) (2/3)

- **MAPLR**
  \[ J_{MAPLR}(\hat{W}) = R(\hat{W}|W) = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \log \frac{p(X_{m,n}|\hat{W}_r, \lambda_m)g(\hat{W}_r)}{p(X_{m,n})} \]

- **AAPLR**
  \[ J_{AAPLR}(W) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \frac{p(X_{m,n}|W_r, \lambda_m)P_m g(W_r)}{p(X_{m,n})} \]

  —aggregated over all model classes \( m \) with probabilities \( P_m \)

- **Discriminative Training**
  \[ J_{AAPLR}(W) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \ell(d_{AAPLR}^{m}) \]

  \[ d_{AAPLR}^{m} = g_m(X; \lambda_m, W_{r(m)}) - \log \left\{ \frac{1}{M-1} \sum_{j \neq m} \exp[\eta g_j(X; \lambda_j, W_{r(j)})] \right\}^{1/\eta} \]

  \[ g_m(X; \lambda_m, W_r) = \log \{ p(X_{m,n}|W_r, \lambda_m)g(W_r) \} \]
## Comparison with Other Approaches

<table>
<thead>
<tr>
<th></th>
<th>Estimation Criterion</th>
<th>Discriminative adaptation</th>
<th>Bayesian learning</th>
<th>Closed-form solution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
<td>MAP</td>
<td>MCE</td>
<td>MMI</td>
</tr>
<tr>
<td>MLLR</td>
<td>○</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPLR</td>
<td>○</td>
<td>○</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCELR</td>
<td></td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>CMLLR</td>
<td>○</td>
<td></td>
<td></td>
<td>○</td>
</tr>
<tr>
<td>AAPLR</td>
<td></td>
<td>○</td>
<td></td>
<td>○</td>
</tr>
</tbody>
</table>
Lexical Level Approaches

- Pronunciation Modeling for Spontaneous Mandarin Speech
- Language Model Adaptation
  - Latent Semantic Analysis and Smoothing
  - Maximum Entropy Principle
- Association Pattern Language Model
Pronunciation Modeling for Spontaneous Mandarin Speech

- Automatically Constructing Multiple-pronunciation Lexicon using a Three-stage Framework to Reduce Confusion Introduced by the Added Pronunciations

Stage 1: Automatically generating possible surface forms but avoiding confusion across different words

Stage 2: Ranking the pronunciations to avoid confusion across different words

Stage 3: Keeping only the necessary pronunciations to avoid confusion across different words

Multiple-Pronunciation Lexicon

Training Corpus
Association Pattern Language Model (1/5)

- N-grams Consider only Local Relations

- Trigger pairs Consider Long-distance Relations, but only for Two Associated Words

- Word Associations Can Be Expanded for More than Two Distant Words

- A New Algorithm to Discover Association Patterns via Data Mining Techniques
Association Pattern Language Model (2/5)

- Bigram & Trigram

- Trigger Pairs
Association Pattern Language Model (3/5)

- **Association Patterns**

```
Sept. 11 George Bush Twin Towers
```

Diagram:
```
Sept. \(\rightarrow\) 11 \(\rightarrow\) George \(\rightarrow\) Bush \(\rightarrow\) Twin \(\rightarrow\) Towers
```

Arrows indicate association patterns.
Association Pattern Language Model (4/5)

Association Pattern Mining Procedure

<table>
<thead>
<tr>
<th>Candidate 1-word subset</th>
<th>Frequent 1-word subset</th>
<th>Candidate 2-word subset</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>L₁</td>
<td>C₂</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>Taiwan</td>
<td>Bush ∪ Taiwan</td>
<td>0.22</td>
</tr>
<tr>
<td>Bush</td>
<td>President</td>
<td>Bush ∪ President</td>
<td>0.16</td>
</tr>
<tr>
<td>President</td>
<td>Taiwan ∪ President</td>
<td>Taiwan ∪ President</td>
<td>0.07</td>
</tr>
<tr>
<td>China</td>
<td>President ∪ Bush</td>
<td>President ∪ Bush</td>
<td>0.04</td>
</tr>
<tr>
<td>Alliance</td>
<td>Palestine</td>
<td>Taiwan ∪ Bush</td>
<td>0.03</td>
</tr>
<tr>
<td>Israel</td>
<td>People</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Army</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prime Pass: Taiwan → Bush

Join Pass: Taiwan ∪ President

Frequent 2-word subset

Prime Pass: Bush → Taiwan

Join Pass: Bush, President → Taiwan

Candidate 3-word subset

First Prime Pass: Bush, President ∪ Taiwan

Refined 3-word subset

Second Prime Pass: Bush, President ∪ Taiwan

Frequent 3-word subset
Association Pattern Language Model (5/5)

- Association Pattern Set $\Omega_{AS}$ Covering Different Association Steps Constructed

- Merge Mutual Information of All Association Patterns

$$MI(W_{a-1}^q \rightarrow w_j) = \log \frac{p(W_{a-1}^q, w_j)}{p(W_{a-1}^q)p(w_j)}$$

$$\log p_{AS}(W) = \sum_{q=1}^{L} \log p(w_q) + \sum_{s=1}^{S} \sum_{W_{a-1}^s \rightarrow w_j^s \in \Omega_{AS}} \text{MI}(W_{a-1}^s \rightarrow w_j^s)$$

- Association Pattern $n$-gram Estimated

$$\log \tilde{p}(W) = a_1 \log p_{AS}(W) + a_2 \log p(W)$$
Future Directions

- **Robustness in Detecting Speech Attributes and Events**
  - Detection-based processing for Next Generation Automatic Speech Recognition
  - Robustness in sequential hypothesis test for acoustic and linguistic detectors

- **Beyond Current Robustness Approaches**
  - Maximum entropy framework is useful for building robust speech and linguistic models
  - Develop new machine learning approaches, e.g. ICA, LDA, etc, for speech technologies
  - Build powerful technologies to handle complicated robustness problem

- **Application of Robustness Techniques in Spontaneous Speech Recognition**
  - Robustness issue is ubiquitous in speech areas
  - Towards robustness in different levels
  - Robustness in establishing applications and systems