NTT Speech recognizer with OutLook On the Next generation: SOLON

Takaaki Hori
NTT Communication Science Laboratories, NTT Corporation

Outline
- Introduction
- What is SOLON?
- Acoustic modeling in SOLON
- WFST-based efficient decoding in SOLON
- Online speech summarization system using SOLON
- Summary

Introduction
- Progress of speech recognition technology
  - Dictation systems, IVR systems, etc.
- What is expected of the next generation speech recognizers?
  - Recognize real-world speech
    - Open-vocabulary and open-domain
    - Speaking-style independent
    - In any environment
  - Work alongside brilliant applications
    - Something beyond recognition
- Aim of our research
  - Develop basic technologies to achieve real-world speech recognition
  - Build the next generation speech recognizer by integrating these developed technologies
  - Propose new applications using this speech recognizer
**What is SOLON?**

- Research achievements
- Acoustic modeling
- Spoken Language Processing
- Discriminative Training
- Decoding Algorithm
- Evaluation
- Application
- Real world speech

**Architecture of SOLON**

- Recognizer (WFST decoder)
- Trainer (ML/MCE/VBEC)
- Speech input for recognition
- Speech input for training
- Transcription
- Recognition output
- Acoustic model

**Acoustic modeling in SOLON**

- Acoustic model training
  - Conventional Maximum Likelihood (ML) Training
  - Minimum Classification Error (MCE) Training
  - Variational Bayesian Estimation and Clustering (VBEC)
- Model selection by VBEC
  - Automatic selection of the best model according to the amount of training data

**Minimum Classification Error (MCE) Training**

Discriminative training of acoustic models

- | Maximum Likelihood (ML) | Discriminative Training |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Represents distribution of features strictly according to data</td>
<td>Clarifies boundary of classes according to data</td>
</tr>
</tbody>
</table>

MCE Training [Juang & Katagiri 1992]
String-level MCE [McDermott & Katagiri 1987]
VBEC

- Variational Bayes (VB): approximates posterior calculation for practical use based on a variational method [Attias 1999...]
- Applying VB to speech recognition [Watanabe 2002]
  
  VBEC (Variational Bayesian Estimation & Clustering for speech recognition)

Total Bayesian framework for speech recognition holds

\[ \begin{align*}
  (1) & \text{Robust classification based on distribution estimation} \\
  (2) & \text{Model selection according to the amount of training data}
\end{align*} \]

Evaluation in CSJ task

- Corpus of Spontaneous Japanese (CSJ)
- Acoustic model
  - Sampling rate: 16kHz, 25ms window/10ms shift, MFCC(12)+ΔMFCC(12)+Δlog-energy(1)
  - Training data: Academic presentations (187 hour)
- Language model
  - Training data: Transcription of 2592 presentations
  - 3-gram, vocabulary size: 30,000
- Evaluation set
  - CSJ Benchmark Test 1 (academic presentations)

Performance of VBEC model selection

Word accuracy for male lecture speech
Speech recognition in SOLON

Weighted Finite-State Transducer (WFST) in Speech Recognition

- **Advantages**
  - Yield a unified framework for describing models
  - Integrate different models into a single model via composition operations
  - Improve search efficiency via optimization algorithms

- **Problems**
  - Composition of complex models generates a huge WFST
  - Search space increases, and huge memory is necessary

- **Solution**
  - Efficient algorithm using on-the-fly composition

Definition of a WFST

- State
- Final state
- State transition

- Morphological analysis [Pereira 1994]
- Machine translation [Oncina 1994]
- Syntactic analysis [Alshawi 1996]

Composition of WFST’s

- Morphological analysis [Pereira 1994]
- Machine translation [Oncina 1994]
- Syntactic analysis [Alshawi 1996]
WFST-based speech recognition

On-the-fly composition

A pair of WFST's used in on-the-fly composition

Traditional on-the-fly composition
**Proposed on-the-fly composition**

Hypotheses of the first WFST

On-the-fly rescoring pass

**Results of the CSJ task**

CSJ Benchmark test 1 (10 academic presentations)

- **Results of the very large vocabulary task**

- 2,000 utterances in spoken interactive QA domain
- Vocabulary size: 65K, 200K, 1M, 1.8M

**Speech summarization using SOLON**

- Spontaneous Speech Recognition
  - Automatic generation of captions, minutes, etc.
  - Problem:
    1. Recognition errors (ex. 30% word error rate)
    - Large variation, no sufficient training data
    2. Recognition results are difficult for readers to understand
      - Spoken language
      - Redundant information
        - filled pauses, repetitions, repairs, and word fragments
Approach

- Solution: Speech summarization
  - Speech Recognition
    - Transcribe speech signal
  - Paraphrasing
    - Translate spoken language into written language
  - Sentence Compaction
    - Extract important phrases excluding recognition errors

WFST-based integration of processings

System Architecture

Experiments

- Acoustic model
  - Trained with speech from 338 lectures (100 hours)
  - Consists of 3000 states, each of which has 16 Gaussians
- Language model for spontaneous speech
  - Trained with transcriptions of 680 lectures (2-M words)
  - Vocabulary size: 20,000
- Language model for written language
  - Newspaper articles (35-M words), WWW (180-K words), Automatically translated text (280-K words)
- Word substitution rules
  - Extracted from a parallel corpus consisting of 6 lectures
  - 3317 substitution rules
- Decoder: SOLON

Example of a speech summarization result

Transcription: あの一部会員の在席・出席というとまと
名前に引きずられるんですよですねけど

Paraphrased result: コーパスベースの音声合成というとまと
名前に知られるのとは思うけど

Summarized result: コーパスベースの音声合成の
名前を知るのだけど

Recognition error

Written style
**Summary**

- **SOLON**
  - The integration of speech recognition research achievements in the CS labs
  - Acoustic modeling in SOLON
    - Automatic model selection by VBEC
  - WFST-based efficient decoding
    - Real-time recognition on over one million word vocabulary tasks
  - Speech summarization system using SOLON

**Summarization Accuracy**

![Graph showing summarization accuracy]