MULTI-LINGUAL SPEECH RECOGNITION SYSTEM FOR SPEECH-TO-SPEECH TRANSLATION

S. Nakamura, K. Markov, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, G. Kikui

ATR Spoken Language Translation Research Laboratories, Kyoto, Japan

IWSLT Workshop, Oct. 1, 2004
OUTLINE

- S2ST and Speech Recognition
- Overview of the ATR ASR System
  - MDL-SSS Acoustic Model
  - Multi-Dimensional Class N-gram LM
- BTEC Corpus Description
- Evaluation:
  - Japanese ASR
  - English ASR
  - Chinese ASR
- Conclusion
Speech-To-Speech Translation System

- **Speech Recognition Module:**
  - Provides text input for translation module
  - Can provide additional information:
    - Word POS tags
    - Word Confidence scores
    - Out-of-domain utterance control
Minimum Description Length (MDL)
Criterion for Model Selection

\[ L_i(x) = -\log P(x | \hat{\theta}^{(i)}) + \frac{\alpha_i}{2} \log N_T + \log I \]

\[ \log \text{likelihood} \quad \# \text{ of parameters} \quad x \log \# \text{of samples} \]

\( x = \{x_1, \ldots, x_{N_T}\} \) : observation data
\( \{1, \ldots, i, \ldots, I\} \) : a set of models
\( \alpha_i \) : the number of free parameters of model \( i \)
\( \hat{\theta}^{(i)} \) : the maximum likelihood estimate of model \( i \)
Gain Function of MDL-SSS

A gain function can be derived from the difference of the MDL criteria between before splitting and after splitting.

For contextual splitting:

\[
G_c^{(MDL)}(S_i) = -G_c^{(ML)}(S_i) + C_c \left( \frac{\alpha'_c - \alpha_c}{2} \right) \log N_{all}
\]

For temporal splitting:

\[
G_t^{(MDL)}(S_i) = -G_t^{(ML)}(S_i) + C_t \left\{ \frac{\alpha'_t}{2} \log N_{all}' - \frac{\alpha_t}{2} \log N_{all} \right\}
\]

\(C_c, C_t\): adjust differences between the 1\textsuperscript{st} term and the 2\textsuperscript{nd} term.
MDL-SSS Algorithm

Initial state

For all states

Temporal splitting

Select the splitting with min criterion,

\[ G_t^{(MDL)} \]

or

\[ G_c^{(MDL)} \]

Contextual splitting

Yes

\[ G_t^{(MDL)} > 0 \text{ and } G_c^{(MDL)} > 0 \]?

No

Parameter re-estimation

Finished

“HMnet”

(T. Jitsuhiro et al., 2004)
Multi-Class N-gram LM

Conventional Class 2-gram

\[ P(w_i | w_{i-1}) \approx P(c(w_i) | c(w_{i-1}))P(w_i | c(w_i)) \]

Class assignment of an and a:
- Same class -> less accurate
- Different class -> less reliable

This is an animal.

This is a bird.

Multiple class assignment depends on direction:

\[ P(c^f(w_i) | c^p(w_{i-1}))P(w_i | c^f(w_i)) \]

(H. Yamamoto et al., 1999)
Basic Travel Expression Corpus (BTEC)

- Covers utterances in the travel domain:
  - Sentences extracted from bi-lingual phrase-books.
  - Revised to reduce context dependence.
  - Out of domain and special sentences removed.

- Divided into 4 parts – BTEC 1,2,3 and 4:
  - In total: ~600,000 sentences

- Available in 3 languages:
  - Japanese
  - English
  - Chinese
Japanese ASR - Experiment

- **Training data for acoustic models:**
  - Pseudo-dialogs: Travel Arrangement (TRA)
  - Phonetic balanced sentences (BLA)
  - Total 30 hours
  - 407 speakers

- **Training data for language models:**
  - BTEC: 160k sentences with 1.2 M words
  - 37K word dictionary

- **Evaluation data**
  - BTEC test set 01: 510 sentences
  - 20 males and 20 females
Japanese ASR - Performance

Word accuracy [%]

# of states

ML-SSS (3 states max)
ML-SSS (4 states max)
MDL-SSS
English ASR - Experiment

- Training data for acoustic models:
  - Wall Street Journal (WSJ) corpus
  - 284 speakers (WSJ-284)
  - Total ~60 hours

- Training data for language models:
  - BTEC: 160k sentences with 1.2 M words
  - 22K word dictionary

- Evaluation data
  - BTEC test set 01: 200 sentences
  - 10 males and 10 females
English ASR - Performance

Word Accuracy (%)

| Language Model       | Word Accuracy (%) |
|----------------------|-------------------|
| Word bi-gram         |                   |
| Multi-class bi-gram  |                   |

Tri-gram rescoring
- **No**
- **Yes**
Chinese ASR - Experiment

- **Basic subword units:** 21 Initials and 37 Finals
- **Training data for acoustic models:**
  - ATR phonetically rich Putonghua (General domain)
  - 140 speakers with a total of 54 hours of speech.
- **Training data for language models:**
  - 200k BTEC Chinese sentences
  - 16.5k word dictionary
- **Evaluation data:**
  - BTEC: 12 000 sentences
  - 20 males and 20 females
Chinese ASR - Results

- **Acoustic model**
  - ML-SSS HMnet
  - 1200 states

- **Language model**
  - Multi-class bi-gram
  - Tri-gram
Conclusions

- **ATR multi-lingual ASR system:**
  - Uses advanced modeling technologies – MDL-SSS, Multi-class N-gram, etc.
  - Achieves high performance (about 8% WER) in all languages: Japanese, English and Chinese

- **Ongoing development work:**
  - Implementation of noise and channel robust techniques
  - Adaptation to various accents of Japanese, English and Chinese
  - Field trial in real environment