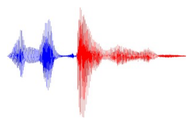
A First Speech Recognition System For Mandarin-English Code-Switch Conversational Speech

N.T. Vu, D.C. Lyu, J.Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. Chng, T. Schultz, H. Li

Karlsruhe Institute of Technology

NANYANG TECHNOLOGICAL UNIVERSITY

1. Challenges



我 吃 chocolate

- Co-articulation effects at code-switches
- No dictionary for Singaporian and Malaysian EN
- Limited training data for AM
- Lack of text data for LM training

2. SEAME Corpus

	Train	Dev	Eval	Total
#Speakers	139	8	19	157
Duration (hrs)	58.4	2.1	2.3	62.8
#Utterances	48,080	1,943	2,162	52,245

- Distribution: EN:CN:SIL:Others = 44:26:21:7
- #Code-Switches = 2.6/Utt
- Very short monolingual segments
- > 82% EN, >73% CN segments < 1 sec
- Average duration of EN segments= 0.67 sec
- Average duration of CN segments = 0.81 sec

3. Baseline system

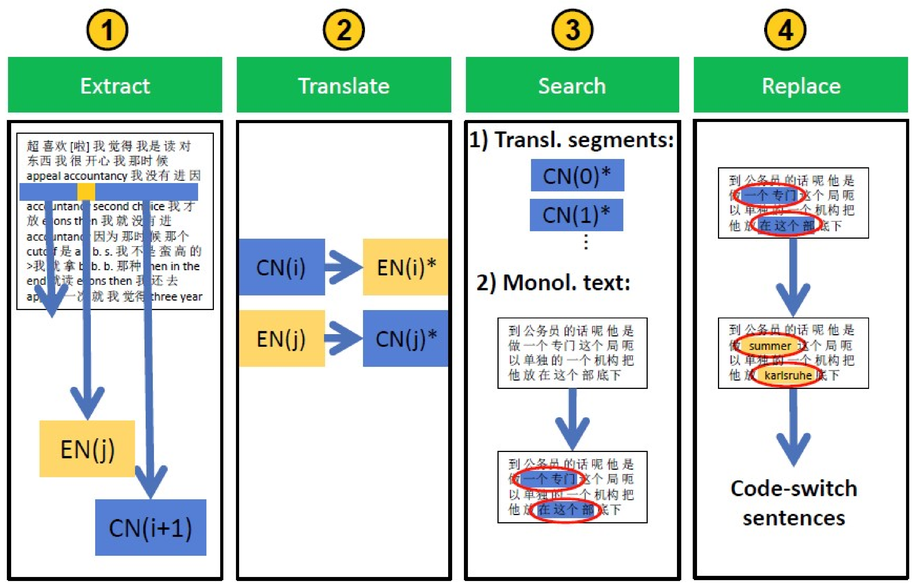
- 143 MFCC (adjacent frames) → 42 (LDA)
- HMM – GMM, SAT + bMMIE training
- CMU EN dictionary + MAN dictionary + pronunciation variations (rules-based)
- n-gram LM:Linear interpolation of CN LM and EN LM with LM trained with CS transcriptions
- OOV rate: 1.21%, PPL: 489.4
- 2-pass system (cMLLR + MLLR)
- MER: 37.3%

4. Code-Switching Acoustic Models

- Phone merging:
  → Knowledge based: IPA
  → Data-driven: using Battacharya Distance
- Apply Discriminative Training on top

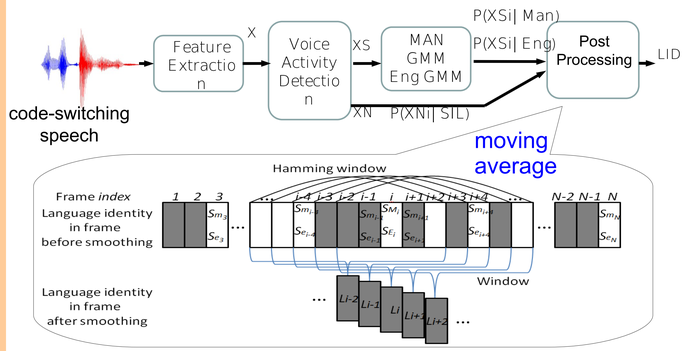
System	Baseline	Knowledge-based	Data-driven
SAT	39.7	39.6	39.6
+ bMMIE	37.3	37.1	37.2

5. SMT-based Language Modeling



- Analysis of approaches to decide which segments to translate
  → based on the code-switch behavior in the SEAME training transcriptions (simple Search&Replace, replacement only if segment occurs at least twice, using information of Trigger Words and Trigger POS tags, adapt frequency of code-switch segments, combinations)
- Best approach:
  Set maximum number of replacements per segment, based on the segment frequency in the SEAME training text
  → Improvement in MER by 0.2% absolute → MER: 36.9%
  → Small improvement due to little data for reliable estimates of code-switch behavior for our approaches (48k utterances in SEAME training text)

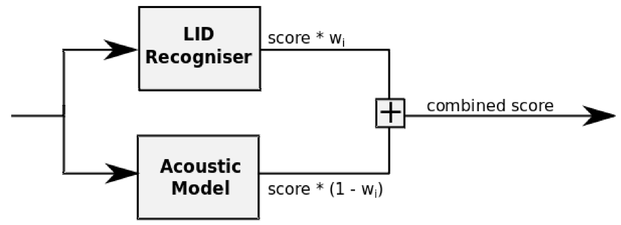
6. Language Identification (LID)



Voice activity detection: 5.88% Frame Error Rate
Two languages LID: 16.64 % FER, Others: 7% FER

7. LID integration into Decoding

Multistream approach



- Language Tag (CN, EN) in dictionary
- LID decision tree, LID weight = 0.1
- MER: 36.5 % (Oracle experiment: 34.4 % MER)

8. Conclusion

Systems	MER(%)
Baseline	37.3
+ CS AM	37.1
+ SMT based LM	36.9
+ LID Integration	36.5
+ Oracle LID	34.4