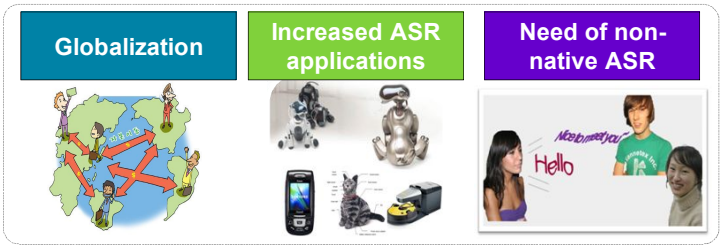
Frame-Level Selective-Decoding using Native and Non-native Acoustic Models for Robust Speech Recognition to Native and Non-native Speech

Yoo Rhee Oh, Hoon Chung, Jeom-ja Kang, and Yun Keun Lee
Speech Language Processing Team,
Electronics and Telecommunications Research Institute (ETRI), Daejeon 305-700, Korea.
E-mail: {yroh, hchung, jjkang, yklee @ etri.re.kr

1. INTRODUCTION

Non-native speech recognition

- Why we need non-native speech recognition
  - Globalization → Frequently use of the non-native language
  - Increasing demands for the ASR-based applications (ex. Language learning applications)
  - **Degradation of ASR performance** for non-native speech



- Regarded as **a mismatch problem** between the training and test conditions
  - **Training** condition: native speech
  - **Testing** condition: non-native speech
  - Widely used methods in speaker or environment adaptation
- **Research works dedicated to non-native ASR**
  - Acoustic modeling
  - Pronunciation modeling
  - Language modeling
  - Hybrid modeling
  - **Many researches uses a small amount of non-native speech**

What to do…

- Use a large amount of both native speech and non-native speech
- Obtain considerable performances for both native and non-native speech

2. SPEECH DATABASE & BASELINE ENGLISH ASR

Speech database

- Native speech database

	Language	Mother tongue	Databases	# of utterances
Training set	English	English	WSJ1, SiTEC DB, ETRI DB	331,527 (280 hours.)
Evaluation set			WSJ1	4,878

- Non-native speech database

	Language	Mother tongue	Databases	# of utterances
Training set	English	Korean	ETRI DB	205,879 (180 hours)
Evaluation set			ETRI DB	3,109

Baseline ASR

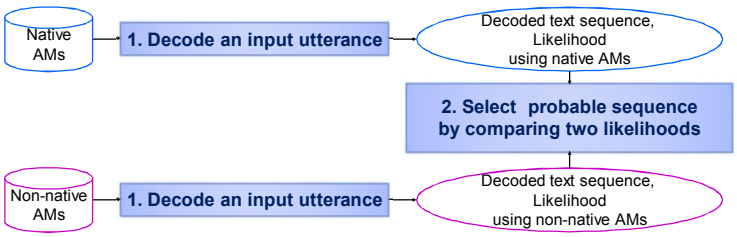
- **Acoustic models (AMs)**
  - 39-dimensional feature vector - 12 MFCCs + log energy, 1st, 2nd derivatives
    - Normalized using a Cepstral mean subtraction (CMS)
  - 3-state left-to-right, triphone-based hidden Markove models (HMMs)
    - Triphones-based HMMs are obtained from monophones-based HMMs
    - States of triphone-based HMMs are clustered using a decision tree
    - 8 Gaussian mixture density
- **Pronunciation models**
  - Native English pronunciations using ETRI grapheme-to-phoneme engine
- **Language models**
  - Native English evaluation set: bigram language models of WSJ1
  - Non-native English evaluation set: bigram language models of English education domain
- **Comparison of word error rates (WERs)**
  - {Native} AM: trained with native speech database
  - {Non-native} AM: trained with non-native speech database
  - {Native, Non-native} AM: trained with native & non-native speech databases – commonly used AMs

AMs	Native English	Relative WER reduction (%) compared to {Native} AM	Korean English	Relative WER reduction (%) compared to {Non-native} AM
{Native} AM	3.5	-	47.5	-112
{Non-native} AM	8.6	-146	22.4	-
{Native, Non-native} AM	5.7	-63	26.7	-19

3. FRAME-LEVEL SELECTIVE-DECODING

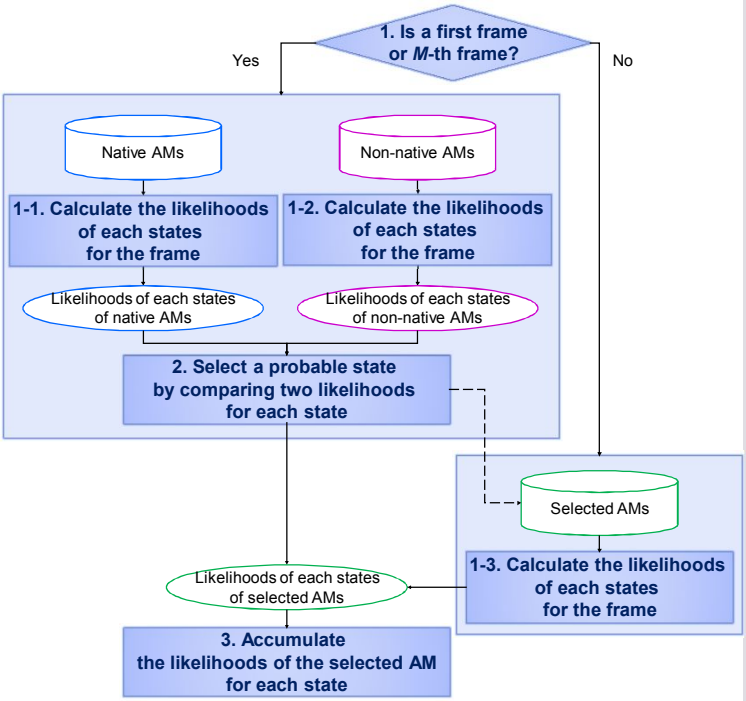
Utterance-level selective-decoding

- For an input utterance, parallel decoding
  - Decode a whole input utterance using native AMs
  - Decode a whole input utterance using non-native AMs
- Select more probable sequence by comparing likelihoods



Frame-level selective-decoding

- For each frame of an input utterance



- After decoding all frames of the input utterance, obtain the decoded text sequence

Comparison of word error rates (WERs)

- Frame-level selective-decoding method provides best performance when the selection is performed for each frame

Selective decoding	Native English	Relative WER reduction (%) compared to {Native} AM	Korean English	Relative WER reduction (%) compared to {Non-native} AM
Frame-level *M=1*	3.3	5.4	23.9	-6.5
Frame-level *M=2*	3.4	3.4	24.0	-6.8
Frame-level *M=3*	3.5	-0.9	24.2	-7.9
Frame-level *M=4*	3.6	-3.2	24.6	-9.6
Frame-level *M=5*	3.9	-10.6	25.4	-13.2
Utterance-level	5.6	-61.0	28.6	-27.6

4. CONCLUSION

Frame-level selective decoding has better performance than utterance-level selective-decoding

- Use well-trained native AMs and well-trained non-native Ams
- For each frame,
  - Decodes each AMs and selects more probable likelihoods