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



٠

٠

•••

•••

*

•••

٠

AMs

{Native} AM

{Non-native} AM

{Native, Non-native} AN

Relative WER reduction (%)

compared to

{Native} AM

-146

-63

Korear

Englis

47.5

22.4

26.7

compared to

{Non-native} AM

-112

-19

Native

Englis

3.5

8.6

5.7

Speech Language Processing Team, Electronics and Telecommunications Research Institute (ETRI), Daejeon 305-700, Korea. E-mail: {yroh, hchung, jjkang, yklee @ etri.re.kr 3. FRAME-LEVEL SELECTIVE-DECODING **1. INTRODUCTION** Utterance-level selective-decoding Non-native speech recognition Why we need non-native speech recognition For an input utterance, parallel decoding Globalization \rightarrow Frequently use of the non-native language Decode a whole input utterance using native AMs Increasing demands for the ASR-based applications (ex. Language learning Decode a whole input utterance using non-native AMs applications) Select more probable sequence by comparing likelihoods Degradation of ASR performance for non-native speech Decoded text sequence. Native Increased ASR Need of non-1. Decode an input utterance Likelihood Globalization AMs using native AMs native ASR applications 2. Select probable sequence by comparing two likelihoods Hello Decoded text sequence Non-native 1. Decode an input utterance Likelihood AMs using non-native AMs Frame-level selective-decoding Regarded as a mismatch problem between the training and test For each frame of an input utterance conditions Training condition: native speech 1. Is a first frame Testing condition: non-native speech or M-th frame? No Yes Widely used methods in speaker or environment adaptation Research works dedicated to non-native ASR Acoustic modeling Native AMs Non-native AMs Pronunciation modeling Language modeling Hybrid modeling 1-1. Calculate the likelihoods 1-2. Calculate the likelihoods of each states of each states Many researches uses a small amount of non-native speech for the frame for the frame What to do... Use a large amount of both native speech and non-native speech Likelihoods of each states Likelihoods of each states of native AMs of non-native AMs Obtain considerable performances for both native and non-native Т speech 2. Select a probable state by comparing two likelihoods 2. SPEECH DATABASE & BASELINE ENGLISH ASR for each state Speech database Native speech database Mother tongue Databases # of utterances Language Selected AMs WSJ1, SITEC DB, ETRI DB 331,527 (280 hours.) Training set English English 1-3. Calculate the likelihoods Evaluation set WSJ1 4.878 Likelihoods of each states of each states of selected AMs Non-native speech database for the frame Language Mother tongue Databases # of utterances 3. Accumulate **FTRI DB** the likelihoods of the selected AM 205,879 (180 hours) Training set English Korean for each state Evaluation set ETRI DB 3,109 After decoding all frames of the input utterance, obtain the Baseline ASR decoded text sequence Acoustic models (AMs) 39-dimensional feature vector - 12 MFCCs + log energy, 1st, 2nd Comparison of word error rates (WERs) derivatives Normalized using a Cepstral mean subtraction (CMS) Frame-level selective-decoding method provides best 3-state left-to-right, triphone-based hidden Markove models (HMMs) performance when the selection is performed for each frame Triphones-based HMMs are obtained from monophones-based HMMs Relative WER **Relative WER** States of triphone-based HMMs are clustered using a decision tree reduction (%) reduction (%) Native English Korean English 8 Gaussian mixture density Selective decoding compared to compared to Pronunciation models {Native} AM {Non-native} AM Native English pronunciations using ETRI grapheme-to-phoneme engine Frame-level M=1 3.3 23.9 -6.5 5.4 Language models Frame-level M=2 3.4 -6.8 3.4 24.0 Native English evaluation set: bigram language models of WSJ1 Frame-level M=3 3.5 -0.9 24.2 -7.9 Non-native English evaluation set: bigram language models of English Frame-level M=4 3.6 -3.2 24.6 -9.6 education domain Frame-level M=5 3.9 -10.6 25.4 -13.2 Comparison of word error rates (WERs) Utterance-level 5.6 -61.0 28.6 {Native} AM: trained with native speech database -27.6 {Non-native} AM: trained with non-native speech database 4. CONCLUSION {Native, Non-native} AM: trained with native & non-native speech databases commonly used AMs

Relative WER reduction (%) Frame-level selective decoding has better performance than utterance-level selective-decoding

- Use well-trained native AMs and well-trained non-native Ams ٠ For each frame,
 - Decodes each AMs and selects more probable likelihoods