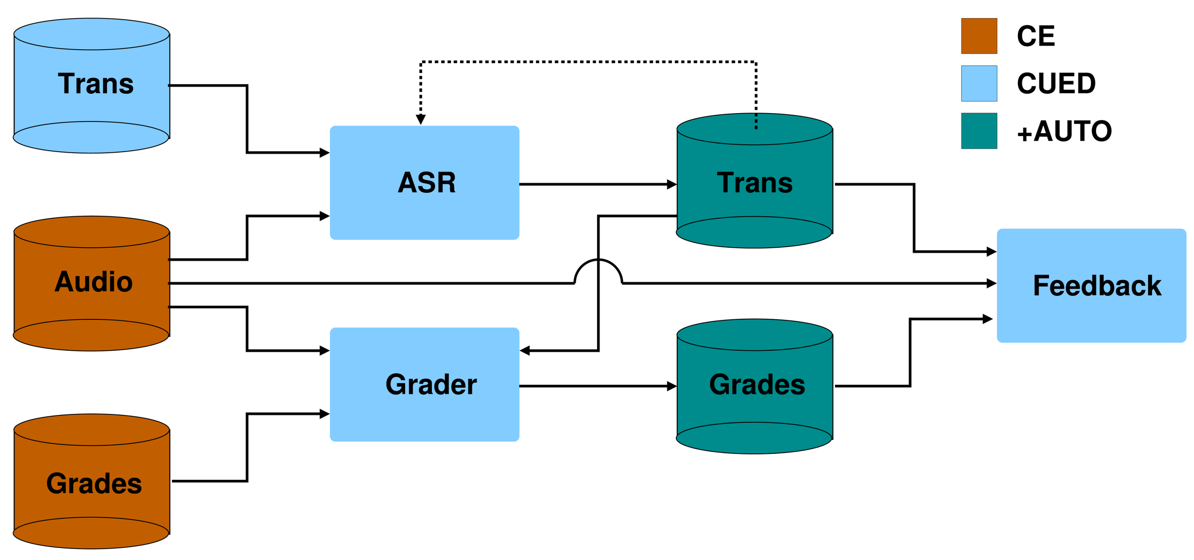
Improving Automatic Speech Recognition for Spoken Language Assessment

Y. Wang, A. Ragni, K. M. Knill and M. J. F. Gales

UNIVERSITY OF CAMBRIDGE

ALTA Institute / Department of Engineering, University of Cambridge

1. Introduction



Automatic speech recognition (ASR) is essential for assessment and feedback

- Grader is trained to be robust to ASR errors
- Feedback is sensitive to ASR errors

However, it is challenging to achieve good recognition accuracy

- Wide variations from e.g. L1, proficiency level, recording
- Spontaneous responses increase difficulty, e.g. disfluencies
- Transcribing is challenging ⟶ inter-annotator error rate about 24.7%

2. Semi-supervised and Supervised Training

Data from Business Language Testing Service (BULATS)

- Section A: short response to prompted questions, Section B: read aloud sentences
- Section C-E: up to 1 minute spontaneous responses to prompts

Trn1 set (108 hours) is comprised of 1000 Gujarati L1 speakers

- Crowd-sourced transcriptions
- Speaker-independent stacked hybrid system build in HTK

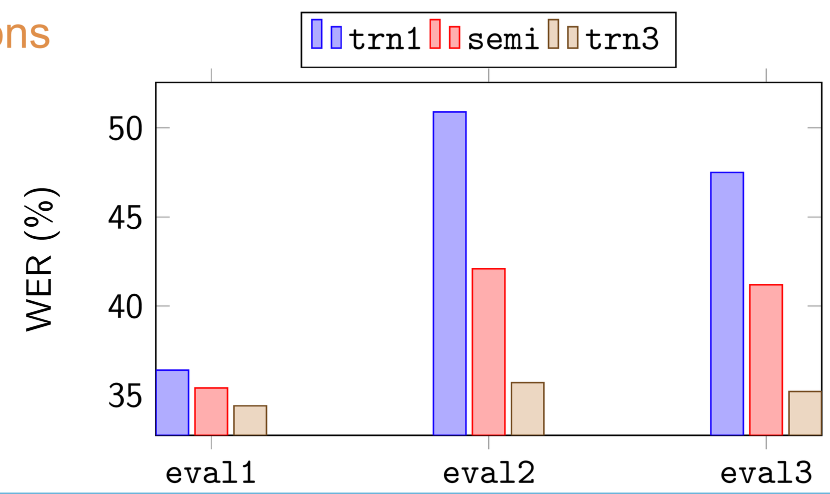
Eval{1,2,3} sets (about 13 hours) contain spontaneous speech from 200 speakers with Gujarati, LA Spanish and mixed L1s, respectively

- Eval3 incluses Polish, Arabic, Vietnamese, French, Thai, Dutch
- Crowd-sourced for spontaneous sections

Semi set contains trn1 and 675 hours unsupervised spontaneous speech

Trn3 set contains trn1 and 200 hours selected from the unsupervised set

- From middle range of confidence
- Contains more than 30 L1s



3. Graphemic Lexicon

Standard ASR uses phonetic lexicon to derive pronunciations

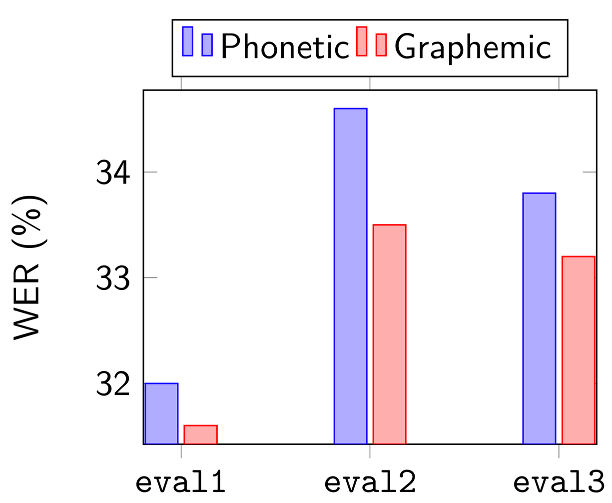
- Reflects standard native pronunciation

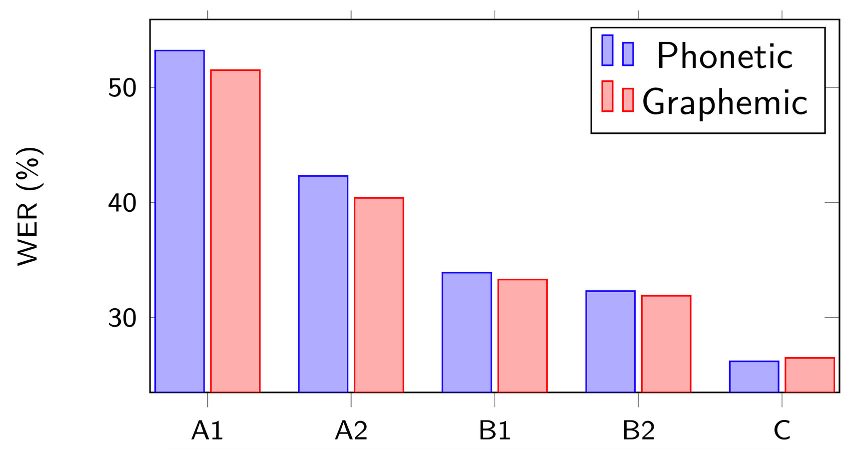
Non-native pronunciations

- Strongly accented, odd pronunciations
- Resort to orthography when in doubts

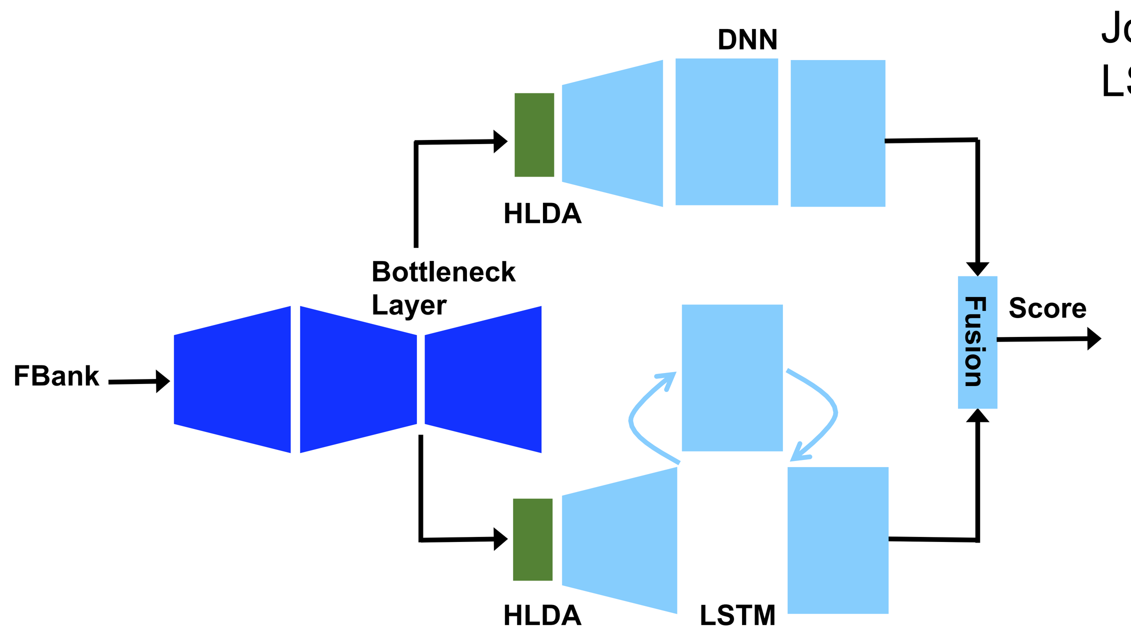
Use graphemic lexicon to yield orthographic pronunciations

- Suitable for lower grade levels





4. Improved ASR System



Joint decoding of SI DNN and LSTM hybrid systems

- Trained on trn3 dataset
- Using a graphemic lexicon
- Built in Kaldi

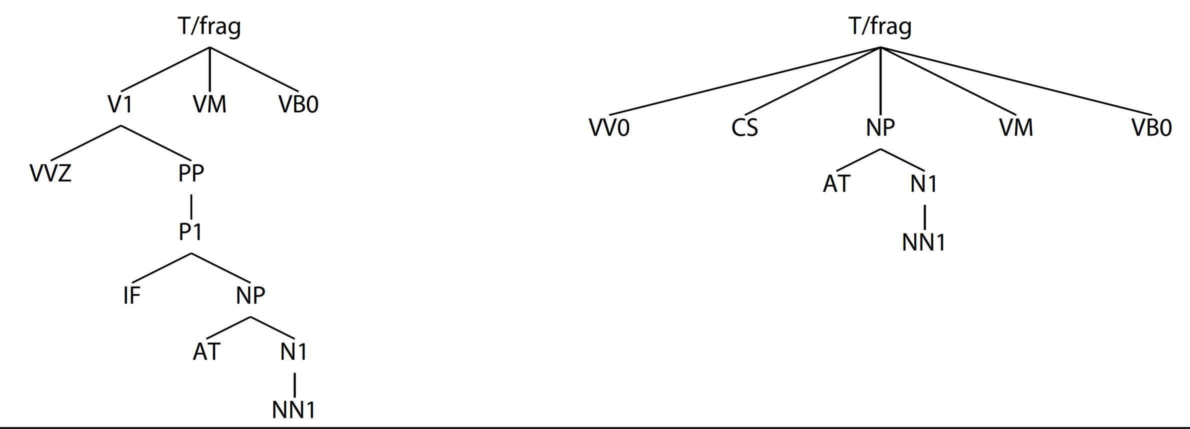
Test set	%WER	
	trn1	trn3
eval1	36.4	30.1
eval2	50.9	30.8
eval3	47.5	30.4

5. Parse Tree

Parse trees represent the syntactic structure of a sentence using context-free grammars

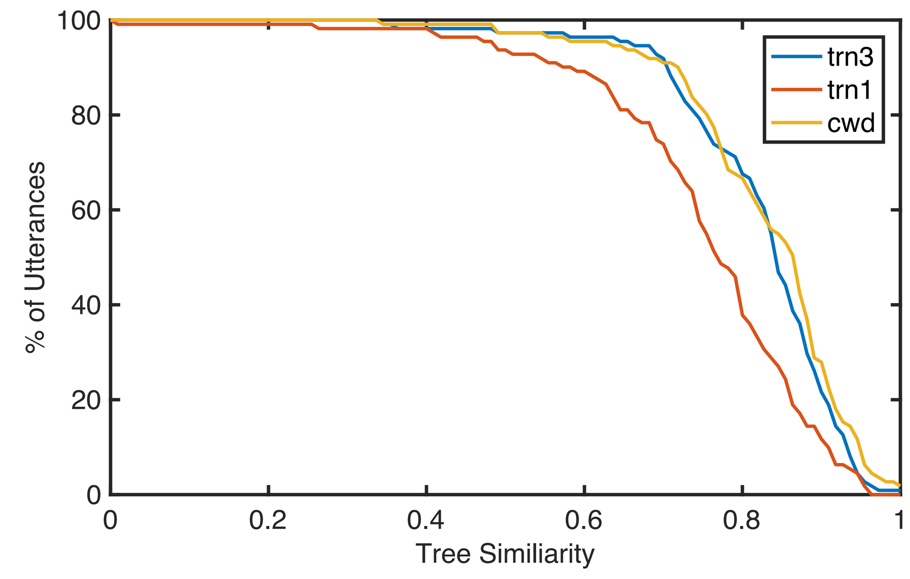
- Sensitive to ASR errors
- Smaller subtrees and leaves are fairly robust

Correct	Speech recognition
advocates for the supplier must be	advocate so the supplier must be



By comparing the parse trees generated on ASR hypothesis against those from a gold standard manual reference, we can get an idea of their suitability for parsing

- Tree similarities are calculated using Convolution Tree Kernels
- Calculated for spontaneous sections
- Hypothesis from trn3 performs similarly to crowd-sourced transcription



Area under the curve:
trn1: 0.750
trn3: 0.822
Crowd-sourced: 0.830

6. Auto-Marking (Grading)

Part-of-Speech (PoS) tags can be extracted from leaf nodes of parse trees

- Reflect relations between words, important for grading and feedback
- More robust than pase trees to ASR errors
- PoS tag error rate calculated by Levenshtein distance
  ⟶ trn1: 42.8, trn3: 30.9

Predict scores using Gaussian Process (GP) grader

- Grader training data: 1000 speakers Mixed L1 data, with standard grades
- Test data: eval3, with expert grades
- Standard grader features derived from audio and ASR hypothesis
  - e.g. mean energy, mean speaking rate, proportion disfluencies
  - robust to ASR errors
- PoS features are extracted as the TFIDF of each PoS tag

Features	PCC	
	trn1	trn3
Baseline	0.854	0.849
POS	0.792	0.830
Baseline + POS	0.847	0.860

7. Conclusion

- ASR for non-native learner English needs data that covers large variations resulting from e.g. L1s, proficiency levels
- Graphemic lexicon can improve the ASR performance
  - Reduce the lexical mismatch
  - Especially suitable for lower grade levels
- Hypothesis from improved ASR has significantly better tree similarities with gold standard transcriptions
  - More syntactically close to manual transcriptions
- PoS features can be extracted from parse trees for GP grader
  - When there are less errors in the PoS tags generated from the hypothesis, PoS features can improve the GP grader.