Use of Graphemic Lexicons for Spoken Language Assessment

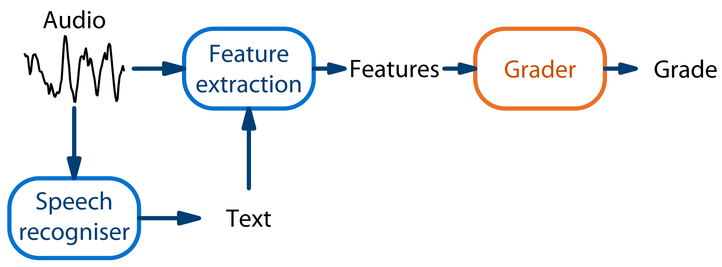
K. Knill, M.J.F. Gales, K. Kyriakopoulos, A. Ragni and Y. Wang
{kate.knill,mjfg,kk492,ar527,yw396}@eng.cam.ac.uk

ALTA Institute / Department of Engineering, University of Cambridge

1. Introduction

Assessment of spoken English of non-native (L2) learners:

- Many people are learning English → want official qualifications
- To help meet this demand:
  - Automatically assess unscripted responses of learners to prompts



Speech recognition is essential for assessment and error feedback

- Challenge: how to achieve good recognition accuracy?
  - wide variations from e.g. L1, proficiency level, recording conditions
  - spontaneous responses increase difficulty - disfluencies etc
  - 'off-the-shelf' systems don't work well
  - limited training data → 'Limited Resource Language (LRL)'
- Lower proficiency level → larger L1 affect on pronunciations
  - mismatch with pronunciation lexicons based on native speaker accents
  - very costly (infeasible?) to tune lexicons to L1s
- Proposal: use graphemic lexicons
  - consistent improvements over phonetic systems for LRLs e.g. Babel
  - BUT English is highly irregular so graphemic systems generally worse
  - can this be used in pronunciation assessment?

2. Graphemic Lexicons

Phonetic Lexicon

ABE'S	ey^I b^M z^F
ABLE	ey^I b^M ax^M l^F
ABOUT_%partial%	ae^I b^M aw^M t^F
ABOUT_%partial%	ax^I b^M aw^M t^F

Graphemic Lexicon

ABE'S	a^I b^M e^M;A s^F
ABLE	a^I b^M l^M e^F
ABOUT_%partial%	a^I b^M o^M u^M t^F;P

- Root graphemes:
  - 26 letters of alphabet
  - Hesitations modelled by graphemes /G00,G01/
- Attributes for context-dependent state tying:
  - word boundary information (I,M,F)
  - A apostrophe, P partial word
  - /a,e,i,o,u/ assigned to the vowel class
  - /vowel,y/ to the vowely class

3. Experimental Results: Automatic Speech Recognition

Data from Business Language Tests (BULATS)

- Up to 1 minute spontaneous responses to prompts
- ASR training data: 100 hours Gujarati L1 English speech
- Test sets: 225 speakers, A1-C grades
  - Gujarati L1 'Gujarati' and 6 mixed L1s 'Mixed'

Decode

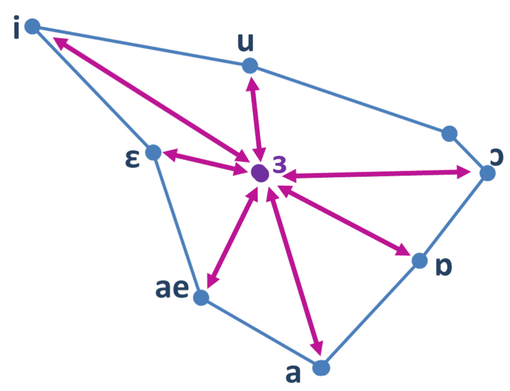
- Stacked hybrid DNN-HMM acoustic model with trigram LM (HTK)

Test Set	Ph/Gr	% WER	
		Viterbi	CNC
Gujarati	Ph	34.3%	33.7%
	Gr	33.3%	32.5%
	Ph ⊕ Gr	-	31.6%
Mixed	Ph	48.6%	47.5%
	Gr	47.2%	46.1%
	Ph ⊕ Gr	-	44.2%

- Graphemic system yields lower WER
- Phonetic and graphemic systems are complementary

4. Spontaneous Spoken Language Assessment

- Predict mark using Gaussian Process (GP) grader
- Standard grader features derived from audio and ASR hypothesis
  - e.g. mean energy, mean speaking rate, proportion disfluencies
- As proficiency improves, pronunciations become more native
  → add explicit pronunciation features to help assessment
- Challenges:
  - large cross-speaker variations → phone acoustic models not robust
  - no reference text for spontaneous responses
  - no reference native speaker speech of responses
- Solution: pronunciation distance features
  - define pronunciation of each phone relative to all other phones
  - full set of phone-pair distances characterises speaker's overall accent
  - more robust to cross-speaker variability



- Phone distance features model
  - Context-independent phone level ASR time alignment
  - Each phone is modelled by a single multivariate Gaussian
  - Model set trained on all the speech from a speaker
  - Symmetric K-L divergence computed for each phone-pair
- Graphemic lexicon system - replace phones with graphemes

5. Experimental Results: Automatic Assessment

Data from Business Language Tests (BULATS)

- Grader training data: 1000 speakers Gujarati L1 English speech
- Test sets: as for ASR plus read speech and short prompt sections
- Pron features: 47 phones/1081 distances, or 26 graphemes/326 distances

Decoder (word)	Gujarati %PER	Gujarati %GER	Mixed %PER	Mixed %GER
Ph	25.8	24.9	33.9	32.9
Gr	29.0	23.7	36.6	30.8

- GER is lower than PER
- PER increases with grapheme decode

Test Set	ASR Grd	Grader PCC Std	Pron	Std+Pron
Gujarati	Ph Ph	0.843	0.838	0.872
	Gr Gr	0.832	0.771	0.849
	Gr Ph	0.841	0.804	0.857
Mixed	Ph Ph	0.852	0.806	0.852
	Gr Gr	0.859	0.734	0.853
	Gr Ph	0.863	0.804	0.870

- Standard features robust to L1 and phone/grapheme recognition
- Pron. distance features less robust to L1 and grapheme-based models

6. Conclusions

- Non-native learner English: 'limited resource language':
  - large variation in spontaneous speech due to L1, proficiency etc
  - limited amount of training data
- Reduce lexical model mismatch with graphemic lexicon:
  - improves recognition performance
  - standard automatic grader performance equivalent to phonetic system
- Pronunciation distance features:
  - introduced to assess spontaneous speech pronunciations
  - graphemic-only system worse than phonetic system
  - phonetic features from graphemic output is successful