UNIVERSITY OF CAMBRIDGE

Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features

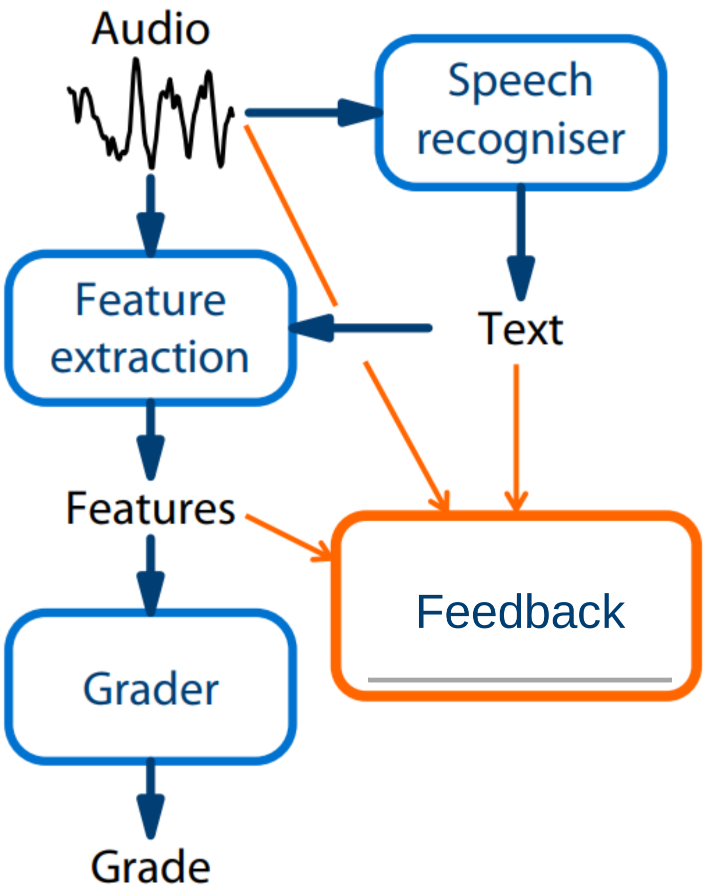
K. Kyriakopoulos, M.J.F. Gales, K.M. Knill

SLaTE 2017

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Automatic Pronunciation Assessment and Feedback

- Motivation is CALL/CAPT
- Features:
  - Matched to specific aspects of proficiency
  - Used for both grading and feedback
  - Indicative of how wrong you are AND how you are wrong



Two Concepts of Bad Pronunciation

- Pronunciation: Rendering of a word as a series of phones

1. Bad pronunciation as individual lexical errors:

e.g. *subtle:*

/s/ /ʌ/ /t/ /ə/ /l/ => /s/ /ʌ/ /b/ /t/ /ə/ /l/ (insertion error)

vs.

2. Bad pronunciation as general property of how one speaks

e.g. Spanish speaker confusing /b/ and /v/

French speaker rendering all /r/ as [χ]

Two Concepts of Bad Pronunciation

- Pronunciation: Rendering of a word as a series of phones

1. Bad pronunciation as individual lexical errors:

e.g. *subtle:*

/s/ /ʌ/ /t/ /ə/ /l/ => /s/ /ʌ/ /b/ /ə/ /l/ (insertion error)

VS.

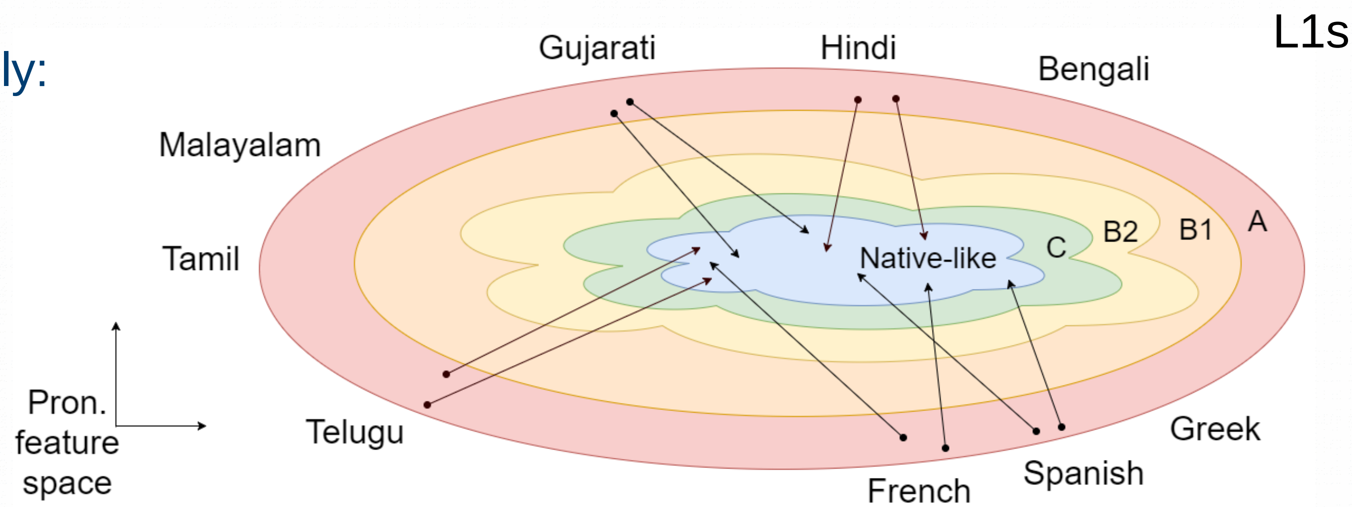
2. Bad pronunciation as general property of how one speaks

e.g. Spanish speaker confusing /b/ and /v/

French speaker rendering all /r/ as [χ]

Hypothesis: Pronunciation Learning Path

Ideally:



- Features should be able to predict:
  1. English speaking proficiency (automarking)
  2. Speaker's L1
- L1 prediction accuracy should decrease with increased proficiency

Key Constraints

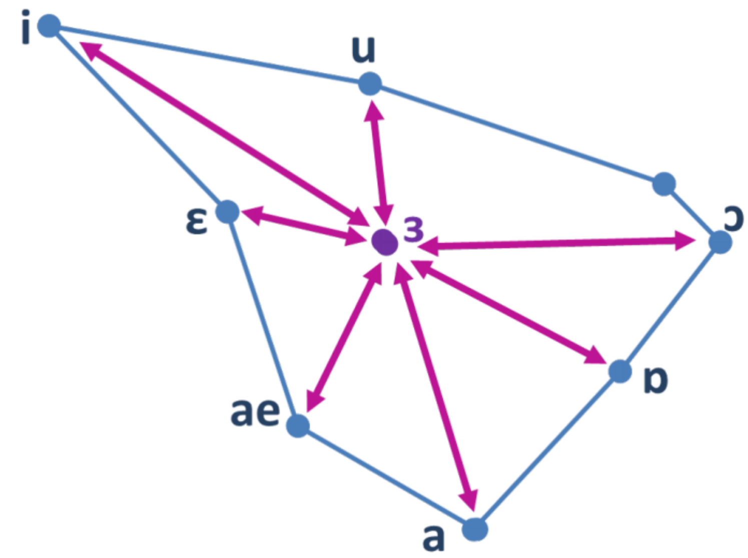
- Spontaneous speech
  - No native models with identical text
  - ASR word (and phone) error rate
- Only have native speaker lexicon (broad not narrow transcription)

e.g. *riot* : /ɹ/ /a/ /ɪ/ /ə/ /t/ and /ɾ/ /a/ /ɪ/ /ə/ /ʔ/ => /r/ /a/ /ɪ/ /ə/ /t/

- Variability in speaker attributes

Phone Distance Features

- Each phone characterised relative to others
- Independence to speaker attributes
- Train model for speaker’s pronunciation of each phone
- Calculate distance between each pair of models



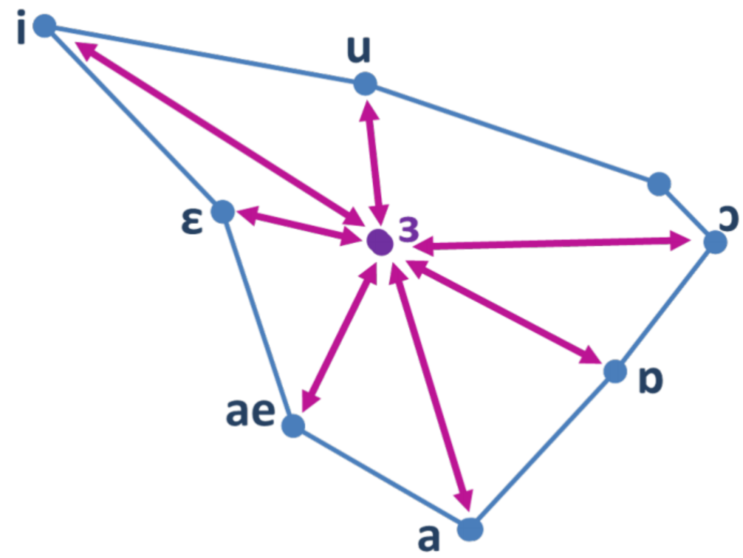
Phone Distance Features

- 47 English phones
- 47 English phones
- 1081 distances
- 1081 distances
- Gaussian model for phone predicts PLP features :
- Gaussian model $p_i(\boldsymbol{o}|\phi_i)$ for phone ϕ_i predicts PLP features $\boldsymbol{o}$:

$$p_i(\boldsymbol{o}|\phi_i) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$

- Distance metric is symmetric K-L divergence:
- Distance metric is symmetric K-L divergence:

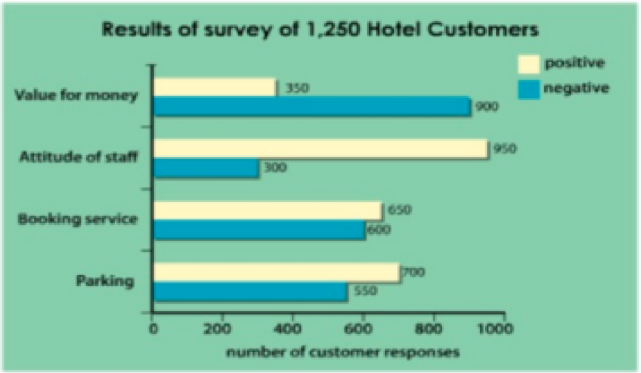
$$D_{JS}(p_i, p_j) = \frac{1}{2}[D_{KL}(p_i||p_j) + D_{KL}(p_i||p_j)]$$



Data: BULATS Speaking Test

- BUsiness Language Testing Service (BULATS) Spoken English Test

  A. Introductory Questions: where you are from
  B. Read Aloud: read specific sentences
  C. Topic Discussion: discuss a company that you admire



  D. Interpret and Discuss Chart/Slide: example above
  E. Answer Topic Questions: 5 questions about organising a meeting

Data: Training and Testing Sets

- 21 L1s
- Balanced gender and proficiency levels
- Varying numbers of speaker per L1
- EVL1 with more L1s to test L1 classification
- EVL2 with expert assigned grades to test score prediction

L1	Set		
	TRN	EVL1	EVL2
Spanish	4502	2156	-
Tamil	1468	790	-
Gujarati	1015	230	94
French	291	115	36
Polish	258	69	39
Vietnamese	245	67	37
Dutch	173	47	32
Thai	144	43	36
Oriya	65	26	-

Table: Nos. of speakers for select L1s

Data: Training and Testing Sets

- 21 L1s
- Balanced gender and proficiency levels
- Varying numbers of speaker per L1
- Spanish speakers from 3 countries

L1	Set	
	TRN	EVL1
Spanish	4502	2156

Country	Set	
	TRN_S	EVL1_S
Colombia	798	296
Mexico	3208	1578
Spain	359	220

Table: Nos. of speakers for Spanish speaking countries

Experimental Setup: ASR

- Acoustic model: Hybrid-Si DNN-HMM
- Language model: N-gram
- Trained on separate set of BULATS Gujarati L1 speakers
- Evaluated using word error rate (WER) and phone error rate (PER)

Correct: Today I ran so far /t/ /ə/ /d/ /e/ /ɪ/ /aɪ/ /r/ /æ/ /n/ /s/ /əʊ/ /f/ /a/

Recognised: Today Iran sofa /t/ /ə/ /d/ /e/ /ɪ/ /aɪ/ /r/ /æ/ /n/ /s/ /əʊ/ /f/ /ə/

Experimental Setup: ASR

- Acoustic model: Hybrid-Si DNN-HMM
- Language model: N-gram
- Trained on separate set of BULATS Gujarati L1 speakers
- Evaluated using word error rate (WER) and phone error rate (PER)

Correct: Today I ran so far /t/ /ə/ /d/ /e/ /ɪ/ /a/ /ɪ/ /r/ /æ/ /n/ /s/ /ə/ /ʊ/ /f/ /a/

Recognised: Today Iran sofa /t/ /ə/ /d/ /e/ /ɪ/ /a/ /ɪ/ /r/ /æ/ /n/ /s/ /ə/ /ʊ/ /f/ /ə/

WER: 80% PER: 6.67%

Experimental Setup: ASR

- Overall WER 47.5%, PER 33.9%
- WER drops with increasing proficiency
- Pronunciation L1 classification accuracy thus trade-off – expect best accuracy in middle, lower at extremes

	Spanish	Arabic	Dutch
A1	69.8	69.7	78.7
A2	58.7	67.4	45.7
B1	48.6	47.2	41.3
B2	47.1	47.3	40.3
C	48.8	48.6	43.1
All	50.9	52.0	42.5

Table: ASR WER(%) by CEFR level for select L1s

Experimental Setup: Automarking

- State of the art DNN
- Trained to minimise MSE for score prediction
- Evaluated using MSE and PCC
- Baseline fluency, vocabulary and simple prosody feature set

Results: Score Prediction

	PCC	MSE
Base	0.737	26.4
PDF	0.751	23.6
Base+PDF	0.832	15.8

Table: PCC and MSE between expert assigned and predicted grades (EVL2)

- Pronunciation features performs better than baseline
- Pronunciation features different and complementary to fluency

Experimental Setup: L1 classification

- Same DNN configuration
- Replace output layer with softmax closed-task classification layer
- Trained for minimum cross entropy
- Evaluated using % Accuracy
- Same baseline fluency, vocabulary and simple prosody feature set

Results: L1 classification

	Accuracy (%)	
	EVL1	EVL2
Base	53.1	31.9
PDF	69.0	61.2
Base+PDF	66.5	60.0

Table: Accuracy of 21-way L1 classifier

- Phone Distance Features better than baseline
- Baseline not complementary to Phone Distance Features

Results: L1 classification

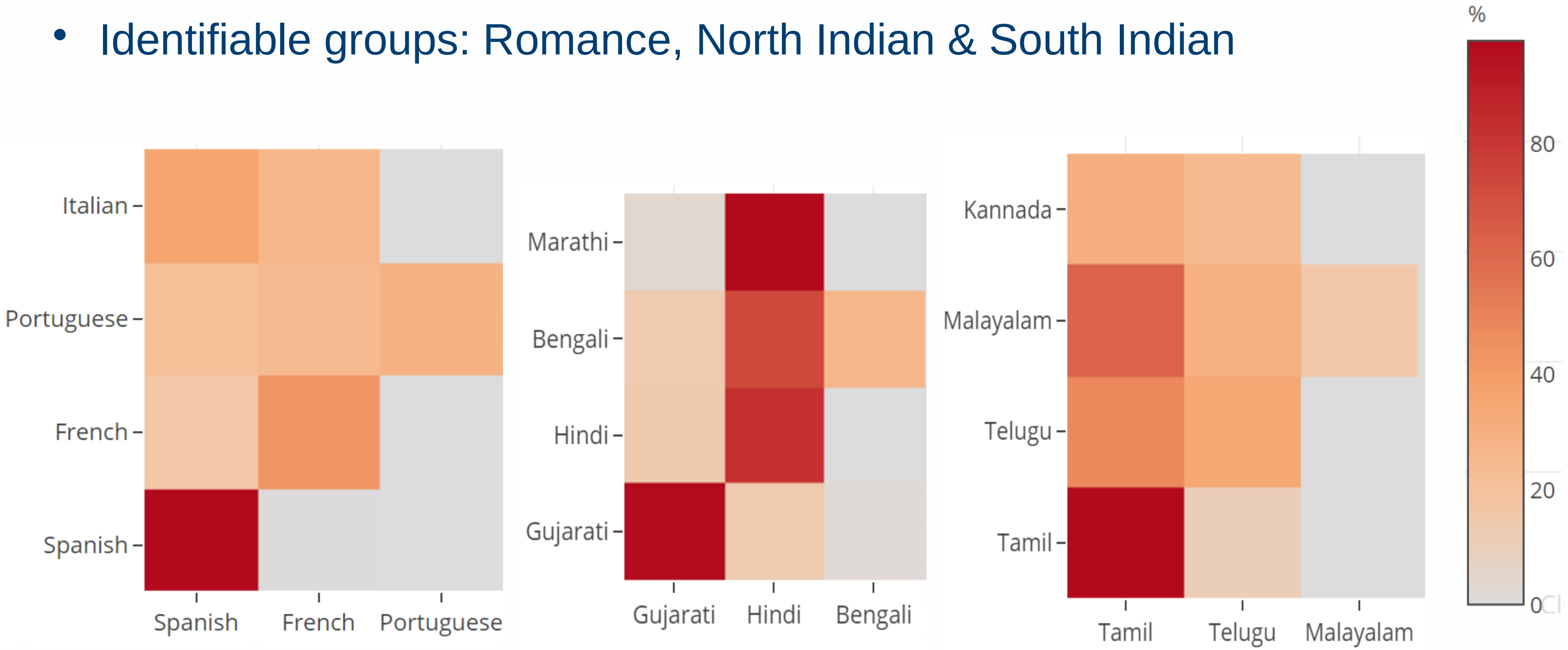
	% Correct L1	% # Speakers in TRN	Most confused
Overall	66.5	-	-
Spanish	97.7	4502	Portuguese
Tamil	76.7	1468	Telugu
Gujarati	74.5	1015	Hindi
Hindi	62.3	563	Telugu
Marathi	0.0	106	Hindi
Italian	2.4	107	Spanish

Table: L1 classifier accuracy for select L1s

- Classifier seems biased to common training data speakers

Results: L1 classification Error Analysis

- L1s most commonly confused with L1s of similar language group
- Identifiable groups: Romance, North Indian & South Indian



Results: Country of origin classification

	Accuracy (%)
Base	77.3
PDF	85.5
Base+PDF	87.0

Table: Country of origin classifier accuracy

- Spanish speakers accurately classified between Spaniards, Mexicans and Columbians

Results: Classification Accuracy by Grade

- PDF accuracy increases then decreases as expected

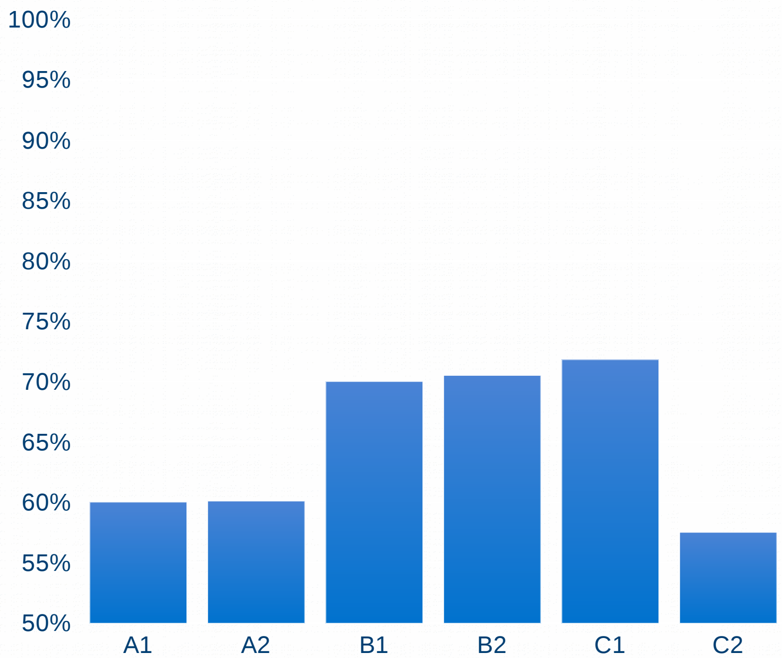


Figure: L1 Accuracies

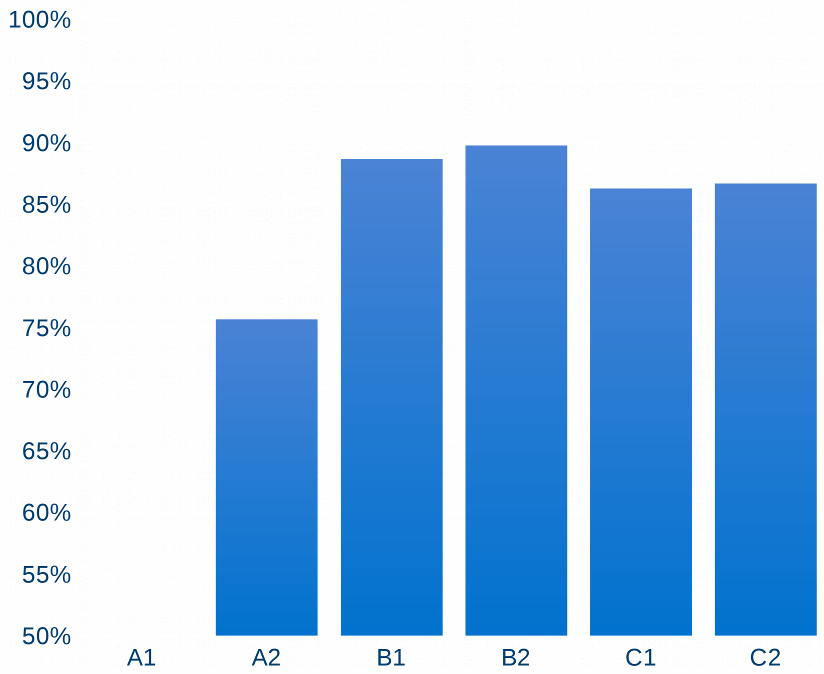


Figure: Country Accuracies

Conclusions

- Phone distance features indicative of:
  - Proficiency
  - Source accent (defined by L1 and country of origin)
- Proficient speakers' source accents are harder to distinguish using PDFs than intermediate speakers
- Features are sensitive to ASR performance

UNIVERSITY OF CAMBRIDGE

Any questions?

Cambridge ALTA

Institute for Automated Language Teaching and Assessment

Variational Autoencoder Projections of Features