# Automatic assessment and error detection of non-native English speech using phone distance features

UNIVERSITY OF CAMBRIDGE


Konstantinos Kyriakopoulos, Kate Knill, Mark Gales

{kk492,kmk1001,mjfg}@eng.cam.ac.uk

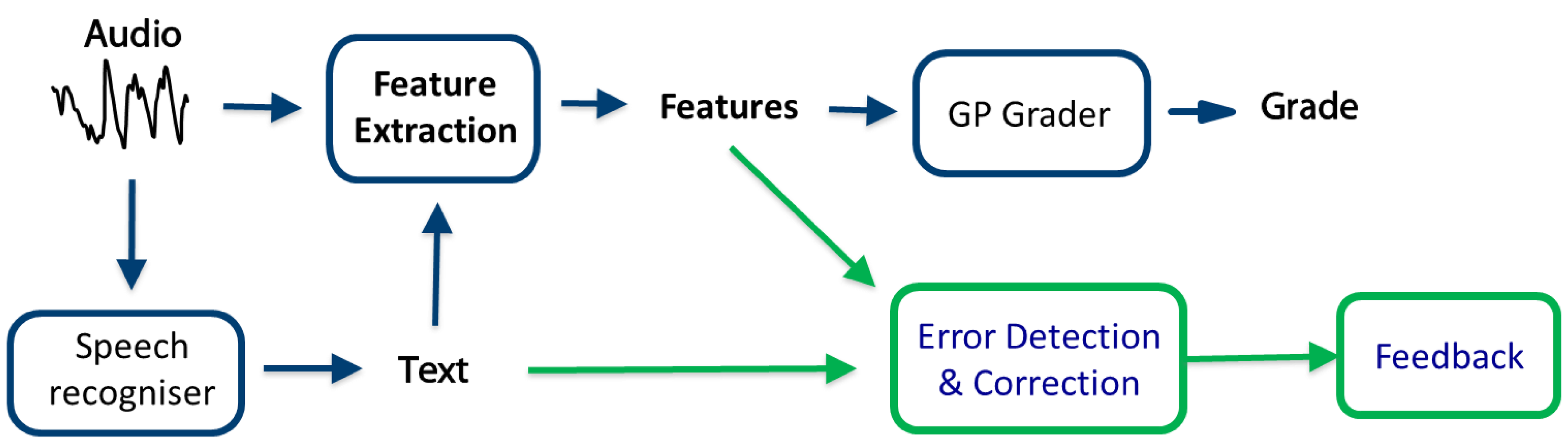
ALTA Institute / Department of Engineering, University of Cambridge


## Introduction

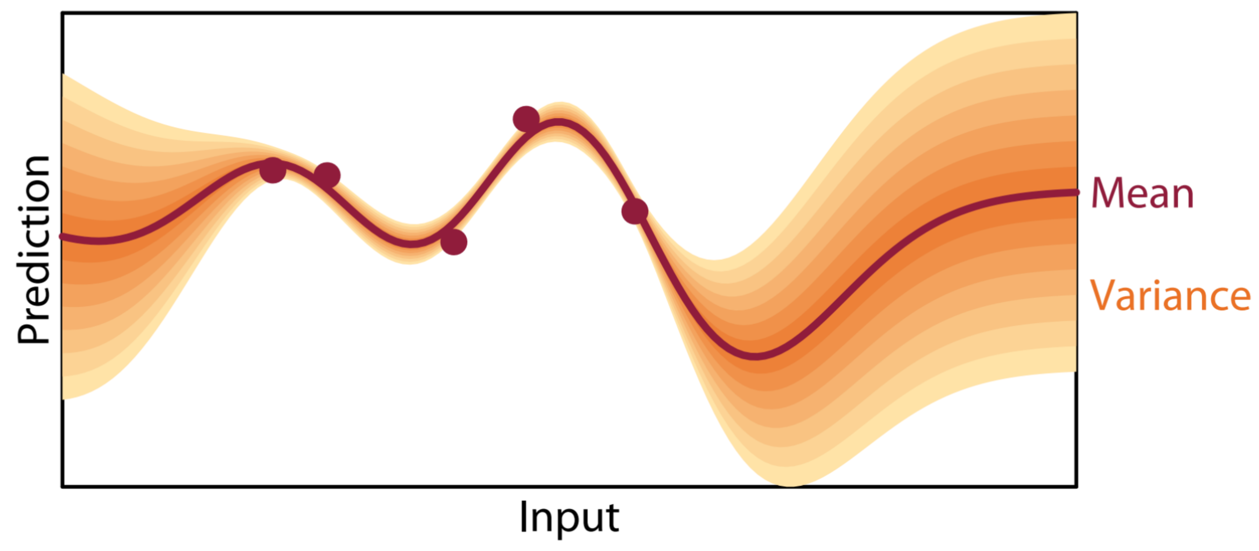
Millions are learning English worldwide and millions yearly take tests

- Automatic assessment assigns grades to candidates
- Error detection identifies and interprets localised mistakes
- Feedback results to user to improve their pronunciation

ALTA system works with unstructured, spontaneous speech

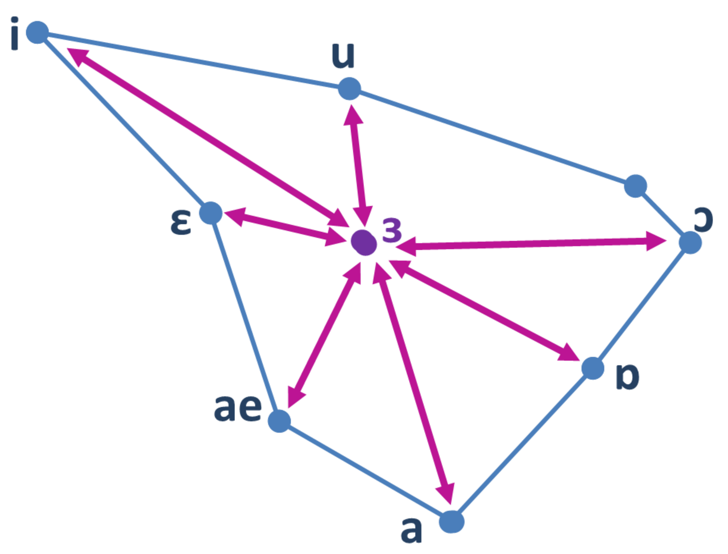



- Speech Recogniser Word Error Rate: 37.6 %
- Gaussian Process (GP) based grader:



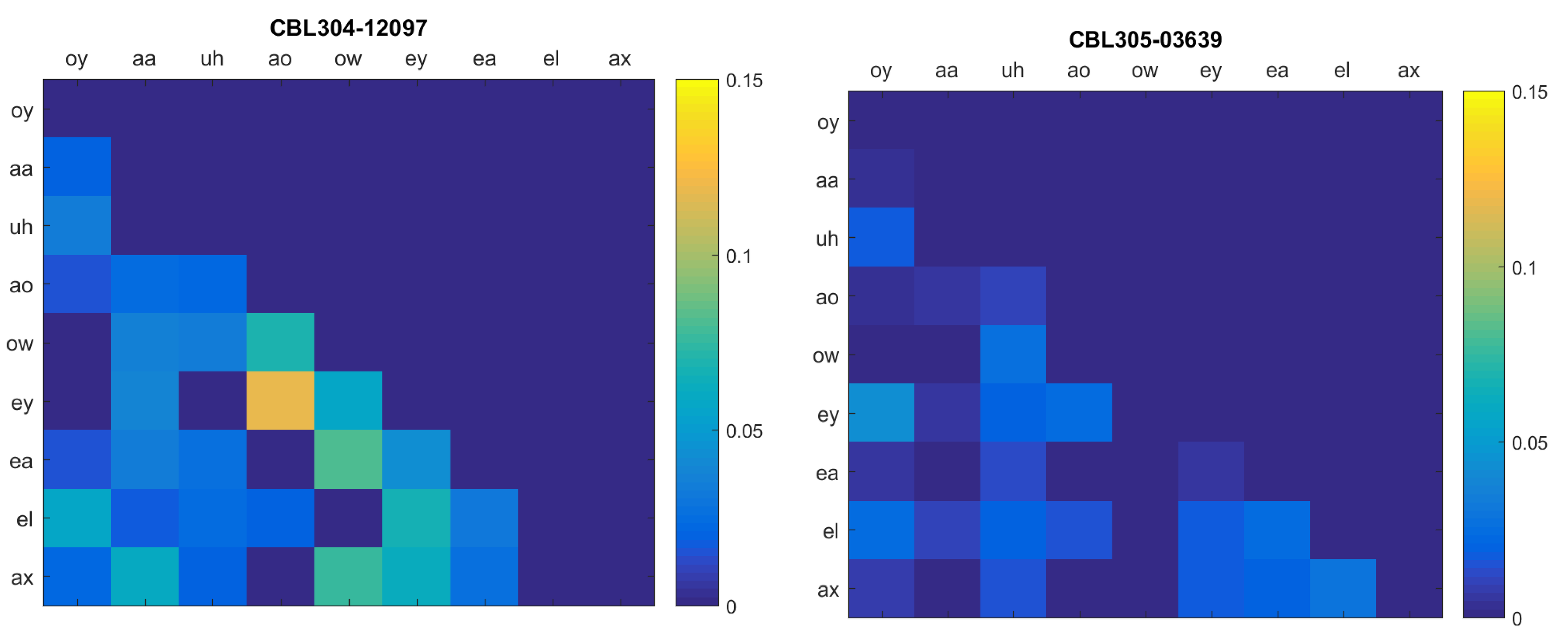

## Phone Distance Features

- Idea: Distances characterise accent but independent of voice quality




- Builds on principle behind vowel formant approach
- All phone pairs - not just vowels
- Full HMM acoustic representation – not just formants
- 1081 distance features for 47 English language phones

Select features for bad speaker (left) and good speaker (right):




Strongest correlations between distance and score are negative i.e. better speakers pronounce phones more similarly

Method:

- Train an HMM on all instances of each phone for each speaker
- When insufficient data for direct training use CMLLR model adaption
- Distance is relative entropy (K-L divergence) between pair of models
- For HMMs with multiple emitting states use variational upper bound

## Assessment Performance

- Train and test on recorded answers to BULATS speaking test
- Unstructured, spontaneous, unlabelled speech

Pearson correlations of grader output with expert grades:

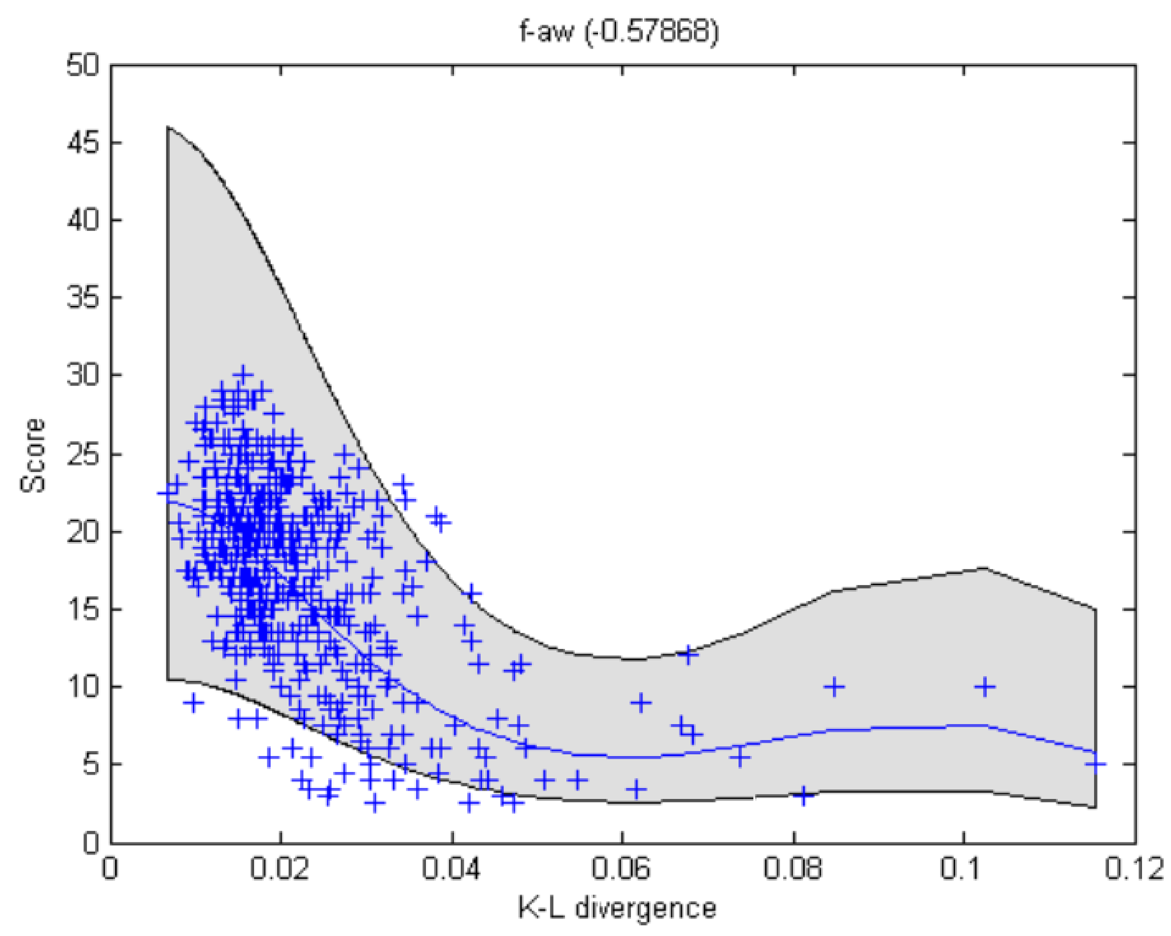
| Data sources | Baseline features | Baseline + pronunciation features |
|---|---|---|
| Only Gujarati speakers | 0.816 | 0.872 |

## Accent Evaluation and Error Detection

GPs to relate score to distance of each phone to all other phones.

f

f-aa
f-ae
f-ah
**f-aw**
f-ax
…
f-uw
f-uy
f-y
f-z
f-zh




Phones that best predict score for different L1s:

| L1ls | Predictive power (Pearson correlations) of top ten phones |
|---|---|
| Spanish | ɪ (0.638), **ɲ** (0.604), u (0.589), **h** (0.587), ə (0.580), æ (0.558), **j** (0.547), ɜː (0.532), **tʃ** (0.505), **b** (0.502) |
| Gujarati | **θ** (0.419), aɪ (0.417), k (0.409), p (0.402), eə (0.395), tʃ (0.379), **w** (0.377), eɪ (0.366), f (0.363), ɪ (0.350) |
| French | ɪə (0.585), l (0.500), **ɲ** (0.442), ɑ (0.442), ʌ (0.442), i (0.442), **r** (0.442), ɜː (0.442), s (0.442), ʃ (0.442) |
| Thai | d (0.724), aɪ (0.711), dʒ (0.701), b (0.685), k (0.674), ɪə (0.651), g (0.643), ɜː (0.632), æ (0.607), ɪ (0.605) |

- Can now score speakers on pronunciation of each phone
- Use to characterise accent relative to L1 and proficiency
- Identify problem phones for feedback to the speaker
- Use to distinguish accent errors from lexical errors

e.g. for a Spanish speaker

| | | | | |
|---|---|---|---|---|
| yes: | jɛs | => | dʒɛs | accent error |
| subtle: | sʌtl | => | sʌbtl | lexical error |

for a French speaker

| | | | | |
|---|---|---|---|---|
| near: | nɪə(r) | => | nɪʁ | accent error |
| grader: | greɪdə(r) | => | gʁædəʁ | accent + lexical error |

- Detect lexical errors with phone substitution and insertion models

## Conclusion

- phone distance features significantly increase grader performance over baseline audio and fluency features
- Performance is stronger for known L1
- Discriminating power is greater for lower scores
- Promising potential for use in error detection