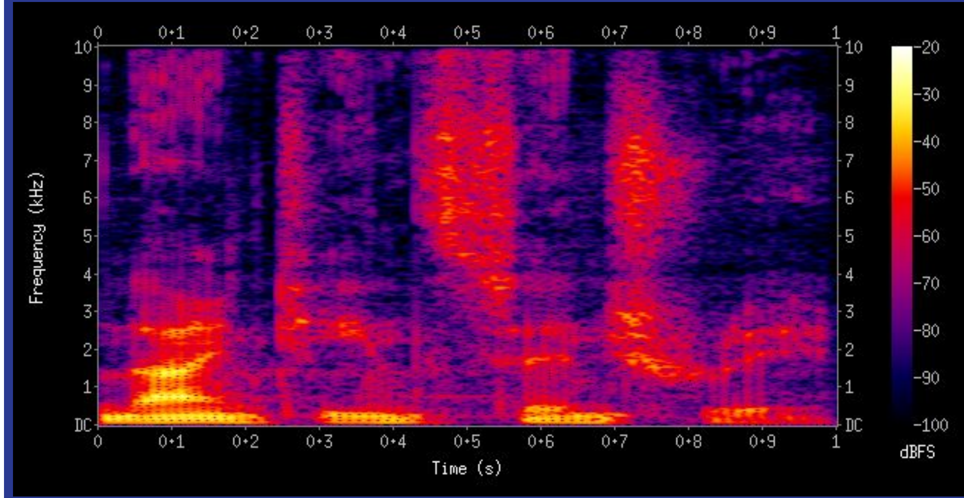
Automated assessment of spoken English

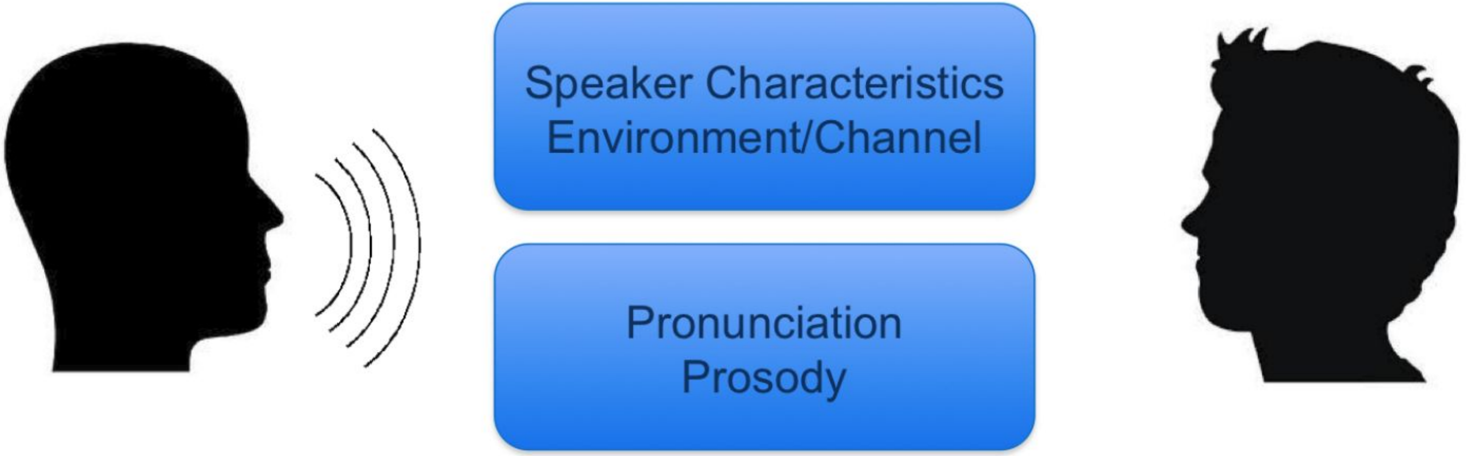
Helen Yannakoudakis



Source: https://commons.wikimedia.org/wiki/File:Spectrogram-19thC.png

Kate Knill, Mark J.F. Gales, Kostas Kyriakopoulos, Andrey Malinin, Anton Ragni, Yu Wang, Rogier van Dalen, Mohamamd Rashid (2017; 2018)

Spoken Communication



Message Construction | Message Realisation | Message Reception

Spoken communication is a very rich communication medium

Spoken Communication Requirements

Message Construction:

- Is the speaker’s topic development appropriate?
- Has the speaker generated a coherent message?
- Is the speaker using language correctly?

Message Realisation:

- Is word pronunciation correct?
- Is the prosody appropriate for the message to convey / for the environment?

Spoken Language vs. Written Language

da I live in <unknown> err I live in a flat room err it's about 20 heh err it's about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat yeah wo and two one kitchen

Spoken Language vs. Written Language

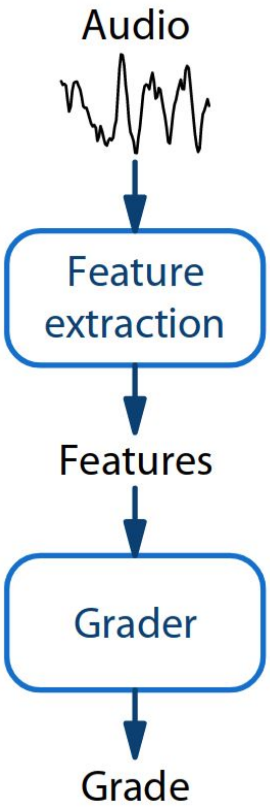
da I live in <unknown> err I live in a flat room err it's about 20 heh err it's about 200 err uh uh quarter meters and there are 4 room in my room err there are 4 rooms in my flat yeah wo and two one kitchen

I live in <unknown>
I live in a flat room
it's about 200 quarter meters and there are 4 rooms in my flat
yeah
and one kitchen

So how can we do automated spoken language assessment?

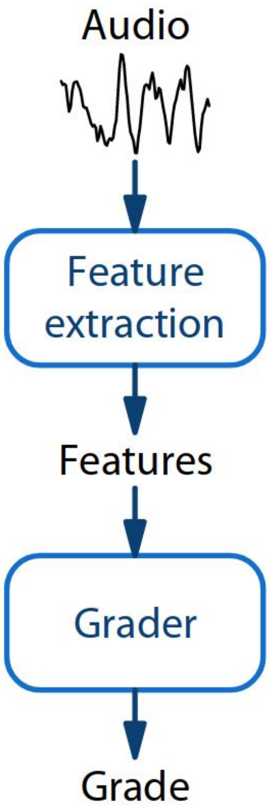
Automated Spoken Language Assessment

Automated Spoken Language Assessment

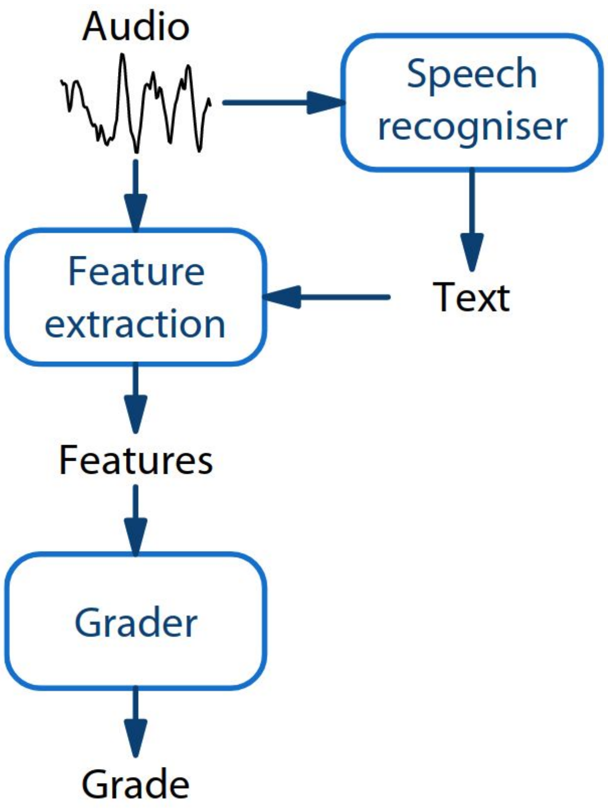


Automated Spoken Language Assessment

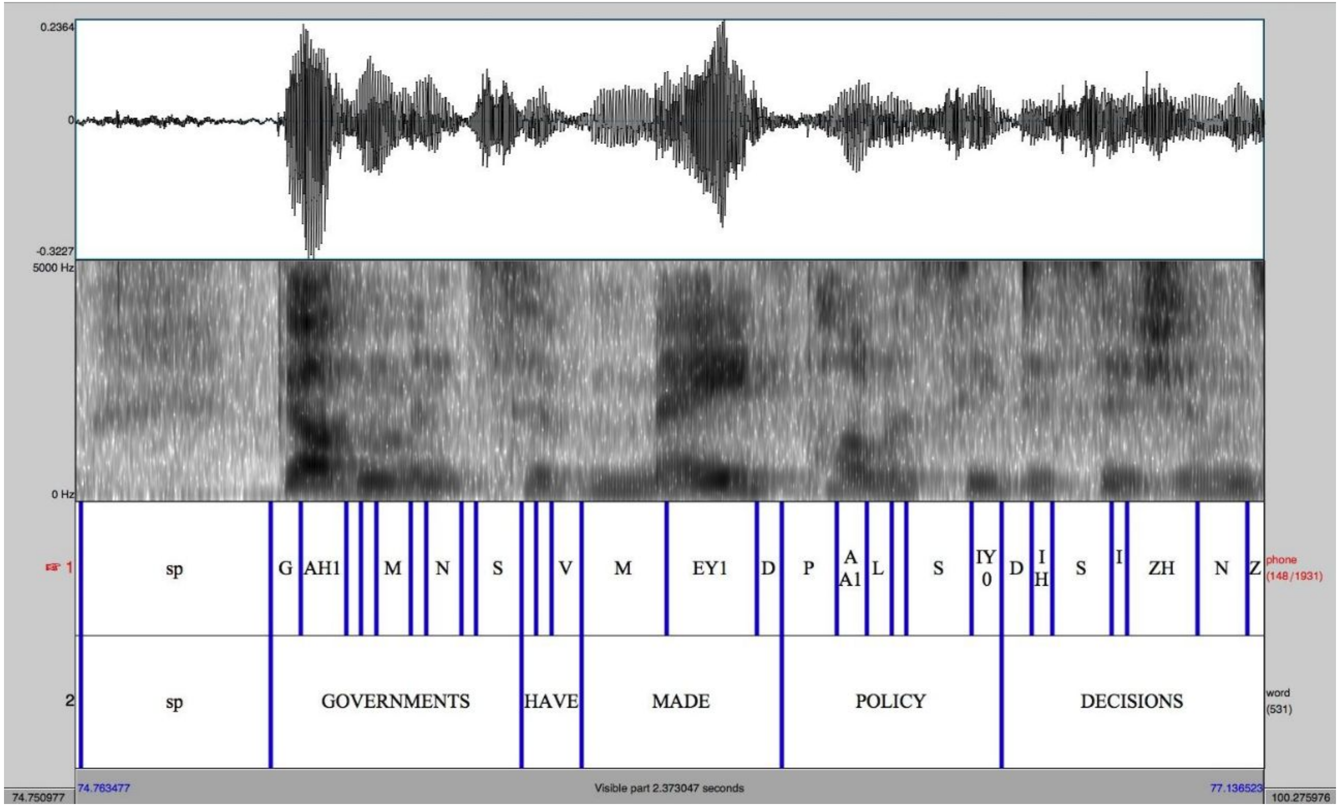
Insufficient information for assessment: no structure, nor information about the message



Automated Spoken Language Assessment



Aligning Speech and ASR Text



Audio and Fluency Features

Audio features	
Fundamental frequency	mean mean-weighted: minimum, maximum, extent, mean absolute deviation
Energy	mean, standard deviation mean-weighted: minimum, maximum, extent, mean absolute deviation
Fluency features	
Long silence	number
Long silence duration	mean, standard deviation, median, mean absolute deviation
Silence duration	mean, standard deviation, median, mean absolute deviation
Disfluencies	number
Words	number, number per second, mean duration
Phones	mean, standard deviation, median, mean absolute deviation

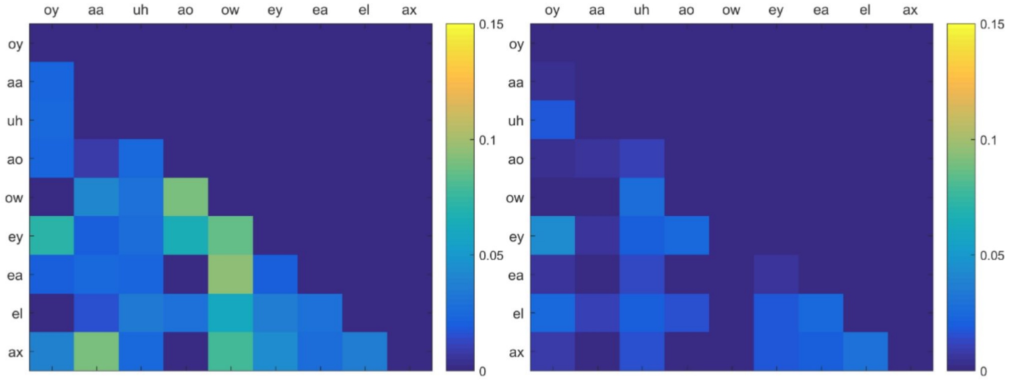
Audio and Fluency Features

Item	Feature	PCC
Audio features		
Energy	mean	-0.05
	standard deviation	-0.03
Fluency features		
Silence	duration mean	-0.34
	duration standard deviation	-0.52
Long silence	duration mean	-0.52
Words	number	0.70
	frequency	0.66
Phone duration	duration mean	–-0.54
	duration median	-0.53

Pronunciation Features

- Train model for a speaker's pronunciation of each phone (Gaussian models for each phone)
- Calculate distance between each pair of models (symmetric KL divergence)
- Features: phone-to-phone distances
    - Phone distance features: distances between acoustic models more robust to speaker variability (though still depend on speaker's L1)

Vowel pairs of poor vs good speaker:

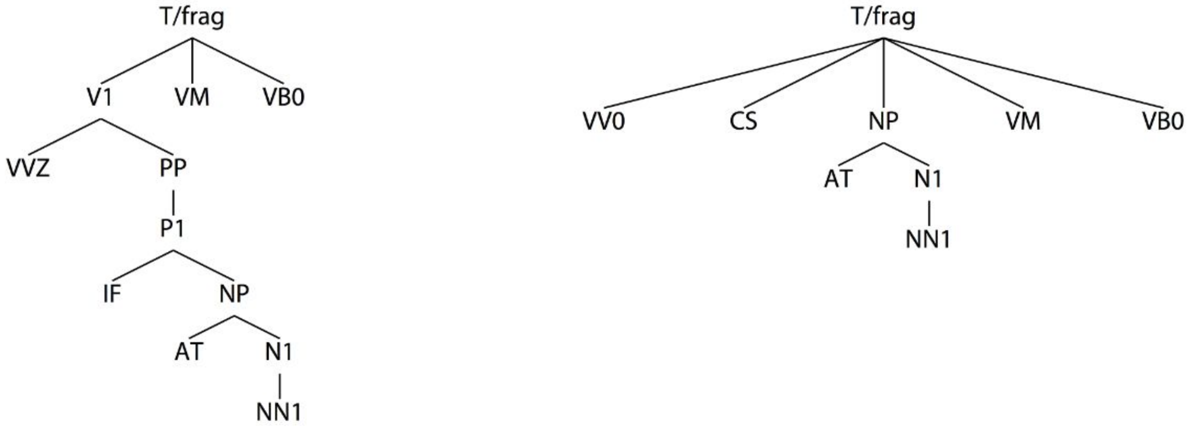


More sophisticated features?

Linguistic features: parse trees

Manual vs ASR transcriptions:

advocates for the supplier must be   advocate so the supplier must be



Challenge: is the parser able to capture the syntactic structure with a high enough level of accuracy?

Linguistic features: parse trees

- Are parse trees sufficiently robust to extract linguistic features?
- Can we determine the “quality” of the parse trees from ASR transcriptions?

Linguistic features: parse trees

- Are parse trees sufficiently robust to extract linguistic features?
- Can we determine the “quality” of the parse trees from ASR transcriptions?
- Calculate the similarity between the ASR-based parse trees and trees from manual transcriptions (using Convolution Tree Kernels)

Convolution Tree Kernels

- Let *n* be the number of unique subtrees in the training set
- We can then represent a tree by an *n* dimensional feature vector $\mathbf{h}(\mathcal{T})$
    - Each element contains the frequency of a subtree

$$\mathbf{h}(\mathcal{T}) = [h_1(\mathcal{T}), h_2(\mathcal{T}), \ldots, h_n(\mathcal{T})]^{\mathrm{T}}$$

- The tree kernel is then defined as the inner product between two trees:

$$k(\mathcal{T}_1, \mathcal{T}_2) = \mathbf{h}(\mathcal{T}_1) \cdot \mathbf{h}(\mathcal{T}_2)$$

- The tree kernel similarity score for the entire set then is:

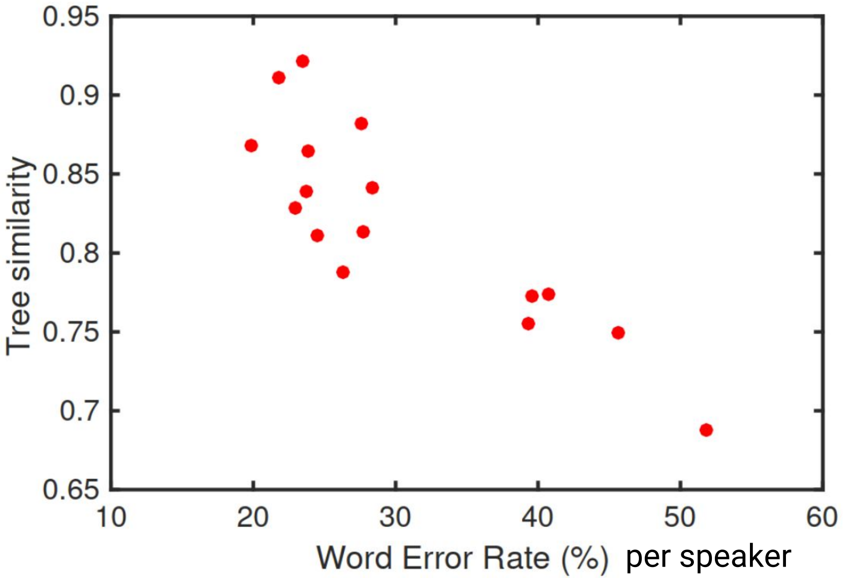
$$\epsilon = \frac{\sum_i k\left(\widetilde{\mathcal{T}}_i, \widehat{\mathcal{T}}_i\right)}{\sum_i \sqrt{k\left(\widetilde{\mathcal{T}}_i, \widetilde{\mathcal{T}}_i\right) k\left(\widehat{\mathcal{T}}_i, \widehat{\mathcal{T}}_i\right)}}$$

Linguistic features: parse trees

- Are parse trees sufficiently robust to extract linguistic features?
- Can we determine the “quality” of the parse trees from ASR transcriptions?
- Calculate the similarity between the ASR-based parse trees and trees from manual transcriptions (using Convolution Tree Kernels)

Linguistic features: parse trees

- Are parse trees sufficiently robust to extract linguistic features?
- Can we determine the “quality” of the parse trees from ASR transcriptions?
- Calculate the similarity between the ASR-based parse trees and trees from manual transcriptions (using Convolution Tree Kernels):



POS features

POS unigram features more robust to ASR errors (though first need to remove eg, partial words and hesitations):

Feature Name	Feature Description	PCC
NN2	plural common noun (e.g. books)	0.66
NN1	singular common noun (e.g. book)	0.65
RR	general adverb	0.61
II	general preposition	0.59
AT	article (e.g. the, no)	0.57

Any features that might help us wrt ASR errors?

ASR Confidence features

Confidence wrt whether a phone/word/utterance has been correctly recognised.

Low confidence reasons:

- Unclear/incorrect pronunciation
- Strong accented speech
- Grammatical errors and disfluencies

Thus, “better” speakers would have higher confidence scores.

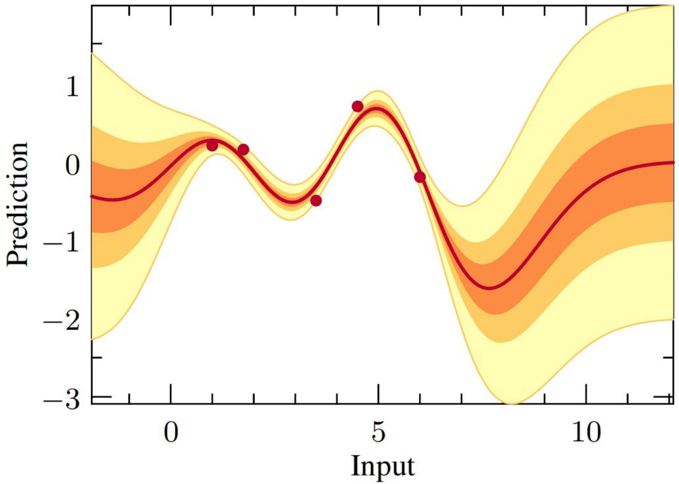
Feature: word posterior probabilities as the confidence score of word hypotheses

Automated grading: Gaussian Process

- A non-parametric Bayesian model for approximating an unknown function
  - Function that maps feature vector into a score/grade
- Provides a measure of the uncertainty around this estimate
  - Variance of function used to assign a measure of confidence to a score/grade
- Parameterised by a mean function and a covariance function:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

- A Gaussian Process trained on five data points:
  - Bands: predicted Gaussian distribution
  - Middle line: mean
  - Coloured band: variance contours



Automated grading: evaluation

Features	PCC	MSE
Baseline	0.843	12.0
+ Conf	0.855	10.9
+ RASP	0.850	11.2
+ Pron	0.854	11.3
+ RASP+Conf	0.860	10.4
+ RASP+Conf+Pron	**0.865**	**10.1**

Room for improvement?