CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Automatic Assessment of Spoken English

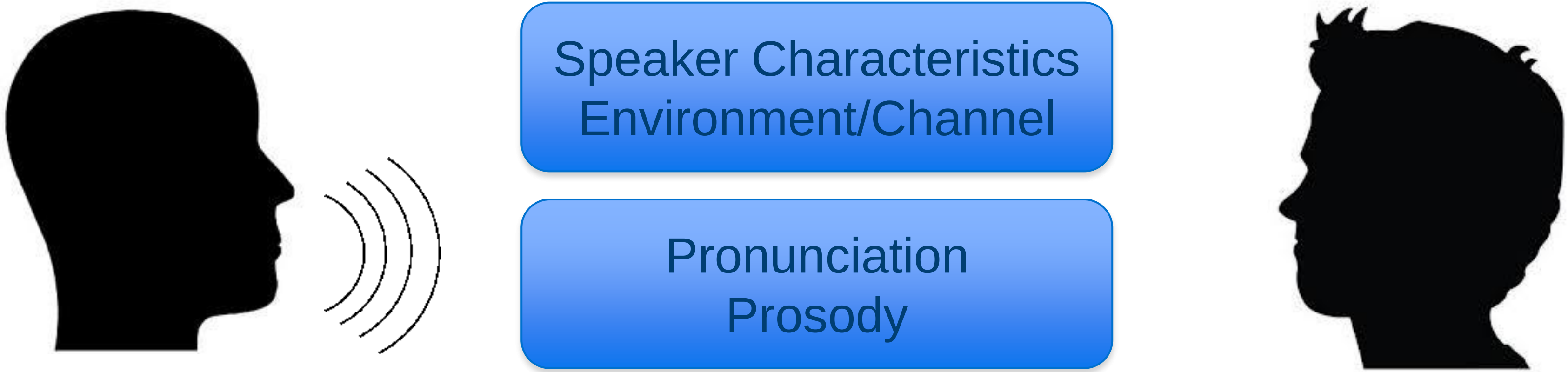
Challenges and Opportunities for Speech Technology

Mark Gales

University of Cambridge

100
CAMBRIDGE ENGLISH
CENTENARY 1913–2013

Spoken Communication



Speaker Characteristics
Environment/Channel

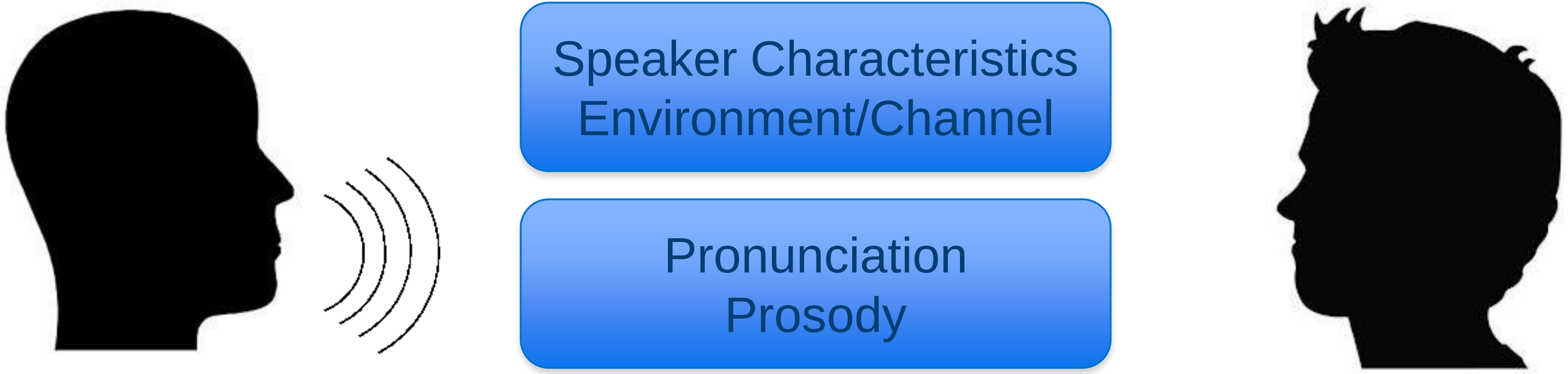
Pronunciation
Prosody

Message Construction

Message Realisation

Message Reception

Spoken Communication



Speaker Characteristics
Environment/Channel

Pronunciation
Prosody

Message Construction — Message Realisation — Message Reception

Spoken language is a very rich communication medium

Spoken Communication Requirements

Message Construction should consider:

- Has the speaker generated a coherent message to convey?
- Is the message appropriate in the context?
- Is the word sequence appropriate for the message?

Spoken Communication Requirements

Message Construction should consider:

- Has the speaker generated a coherent message to convey?
- Is the message appropriate in the context?
- Is the word sequence appropriate for the message?

Message Realisation should consider:

- Is the pronunciation of the words correct/appropriate?
- Is the prosody appropriate for the message?
- Is the prosody appropriate for the environment?

Spoken Communication Requirements

Message Construction should consider:

- Has the speaker generated a coherent message to convey?
- Is the message appropriate in the context?
- Is the word sequence appropriate for the message?

Message Realisation should consider:

- Is the pronunciation of the words correct/appropriate?
- Is the prosody appropriate for the message?
- Is the prosody appropriate for the environment?

Spoken Language Versus Written Language

ASR Output

yeah actually um i belong to a gym down here gold’s gym and uh i try to exercise five days a week um and now and then i’ll i’ll get it interrupted by work you know

Spoken Language Versus Written Language

ASR Output

yeah actually um i belong to a gym down here gold’s gym and uh i try to exercise five days a week um and now and then i’ll i’ll get it interrupted by work you know

Meta-Data Extraction (MDE) Markup

/{DM yeah actually} {F um} i belong to a gym down here / / gold’s gym / / and {F uh} i try to exercise five days a week {F um} / / and now and then [REP i’ll + i’ll] get it interrupted by work {DM you know } /

Spoken Language Versus Written Language

ASR Output

yeah actually um i belong to a gym down here gold’ s gym and uh i try to exercise five days a week um and now and then i’ ll i’ ll get it interrupted by work you know

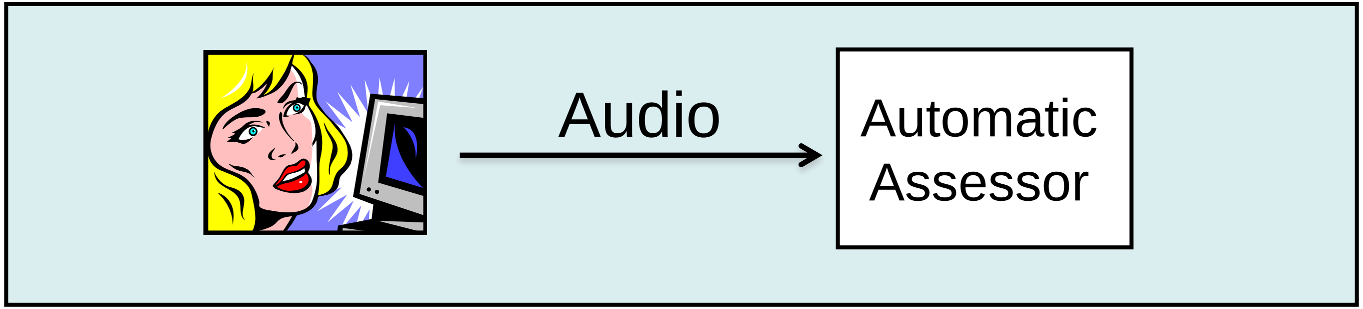
Meta-Data Extraction (MDE) Markup

/{DM yeah actually} {F um} i belong to a gym down here / / gold’ s gym / / and {F uh} i try to exercise five days a week {F um} / / and now and then [REP i’ ll + i’ ll] get it interrupted by work {DM you know } /

Written Text

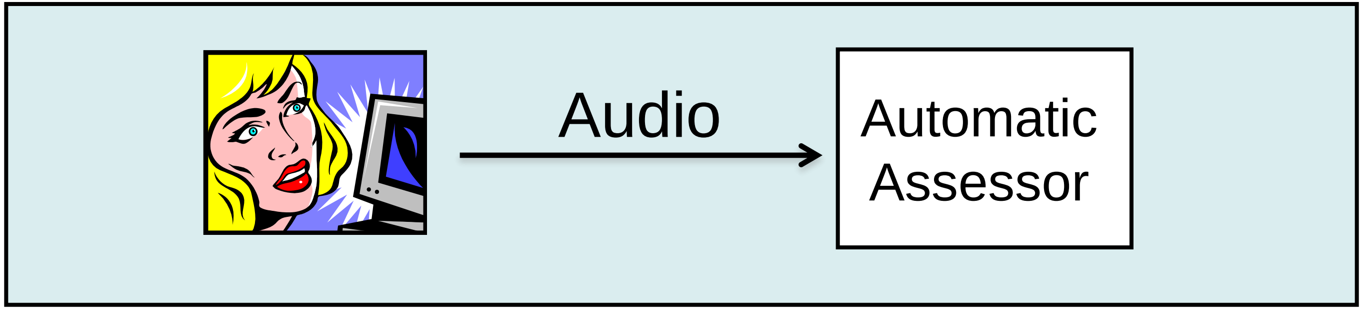
I belong to a gym down here. Gold’s Gym. And I try to exercise five days a week and now and then I’ll get it interrupted by work.

Automatic Spoken Language Assessment



Naive process – directly convert audio into grade

Automatic Spoken Language Assessment



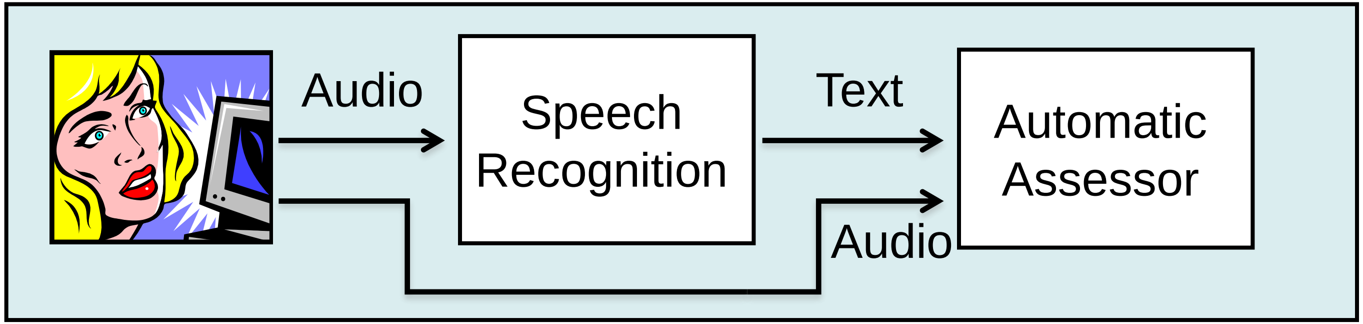
Audio

Automatic Assessor

Naive process – directly convert audio into grade

- Too little “structure” on the audio - insufficient information

Incorporating Speech Recognition



Incorporation of Speech Recognition System

- Adds structure to the audio
- Enables features based on the word-sequence to be used

Speech Recognition is Solved

Siri
Use your voice to send messages, set reminders, search for information, and more.
19

… possibly not

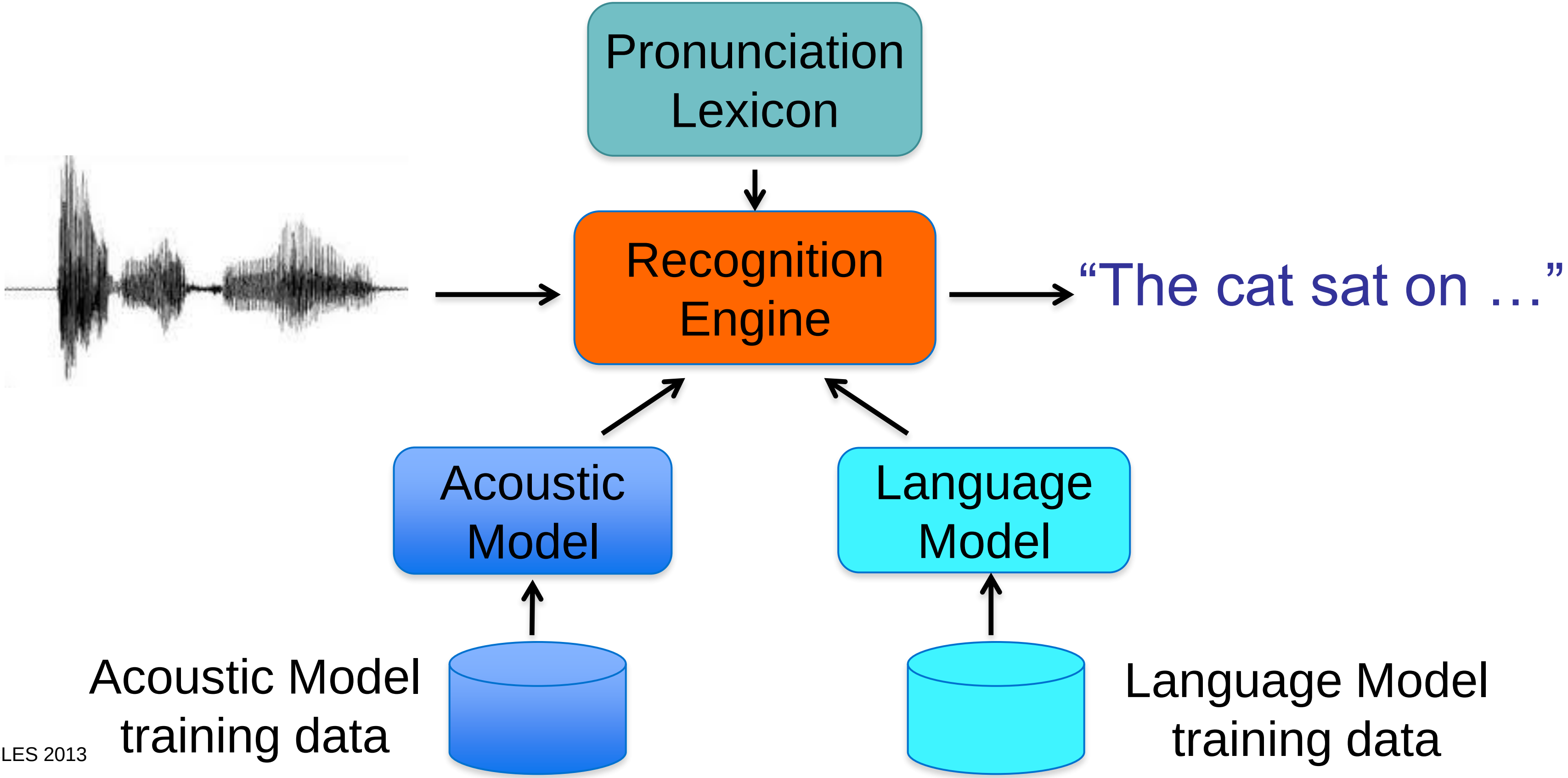
“Can you get the white Tielle please I’m coming home now”

… possibly not

“Can you get the white Tielle please I’m coming home now”

“… Nearly out long will you be home shortly hello Coxnet out long road be home shortly I can”

Automatic Speech Recognition Components



Pronunciation Lexicon

Recognition Engine

“The cat sat on …”

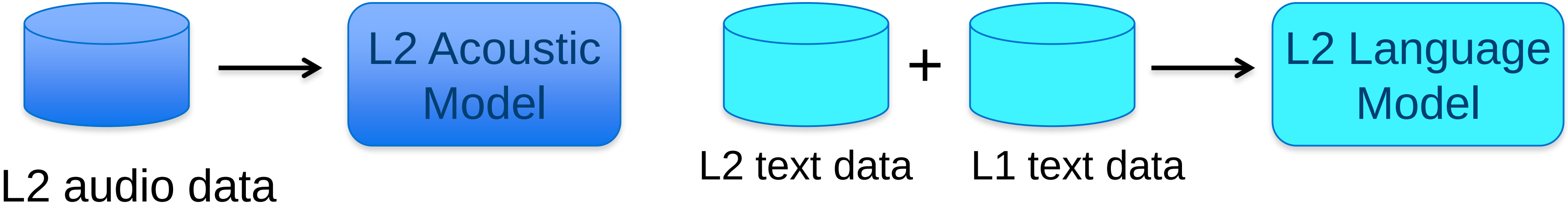
Acoustic Model

Language Model

Acoustic Model training data

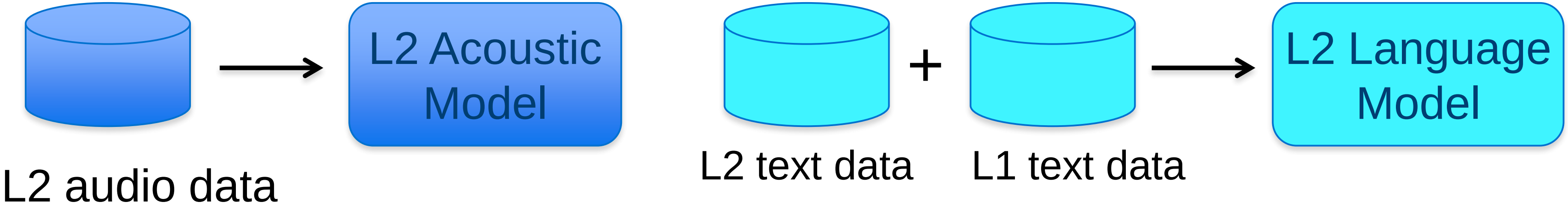
Language Model training data

Forms of Acoustic and Language Models

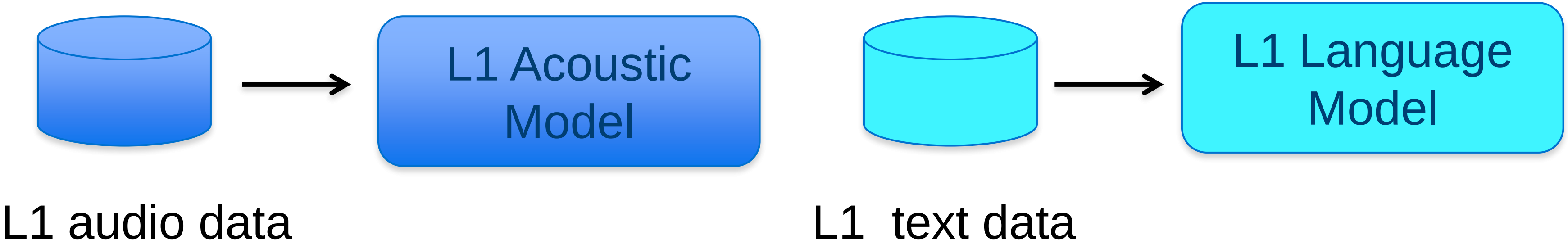


Used to recognise L2 speech

Forms of Acoustic and Language Models

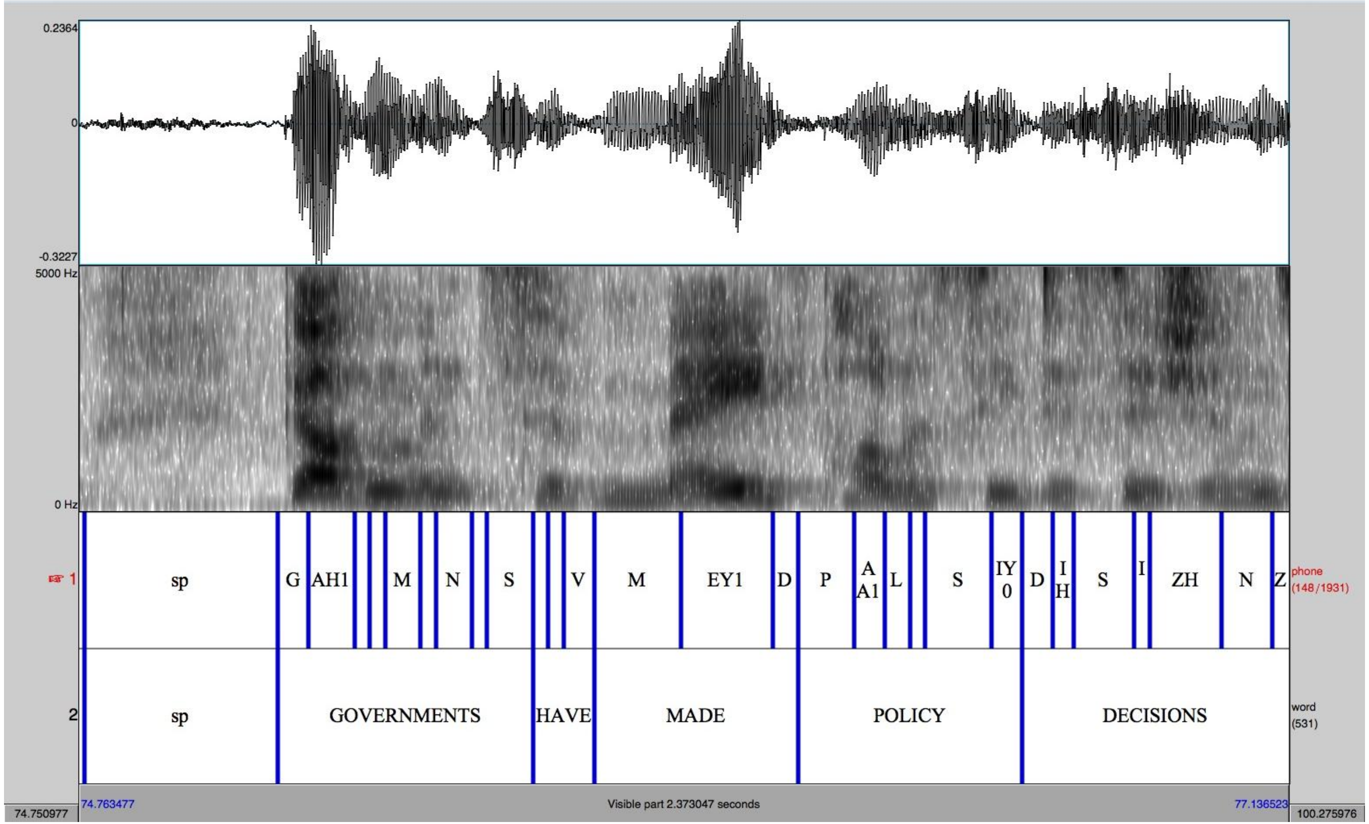


Used to recognise L2 speech



Useful to extract features

Aligning Speech and Text



Possible Features

Text and alignment features

- Word sequence – grammar and vocabulary

Possible Features

Text and alignment features

- Word sequence – grammar and vocabulary
- Disfluencies (hesitations and partial words) - fluency
- Speaker rate (phone/words per second) - fluency
- Pause durations/number of pauses - fluency

Possible Features

Text and alignment features

- Word sequence – grammar and vocabulary
- Disfluencies (hesitations and partial words) - fluency
- Speaker rate (phone/words per second) - fluency
- Pause durations/number of pauses - fluency

Audio features

- Energy/Pitch features

Possible Features

Text and alignment features

- Word sequence – grammar and vocabulary
- Disfluencies (hesitations and partial words) - fluency
- Speaker rate (phone/words per second) - fluency
- Pause durations/number of pauses - fluency

Audio features

- Energy/Pitch features

Richer Set of Possible Features than Written Text!

Speech Recognition Challenges

Mother tongue (L1) impacts speech of non-native English (L2)

- Pronunciation variations from L1 phonological rules
- Intonation (prosodic variations) imported from L1

Wide range of L2 speaking levels

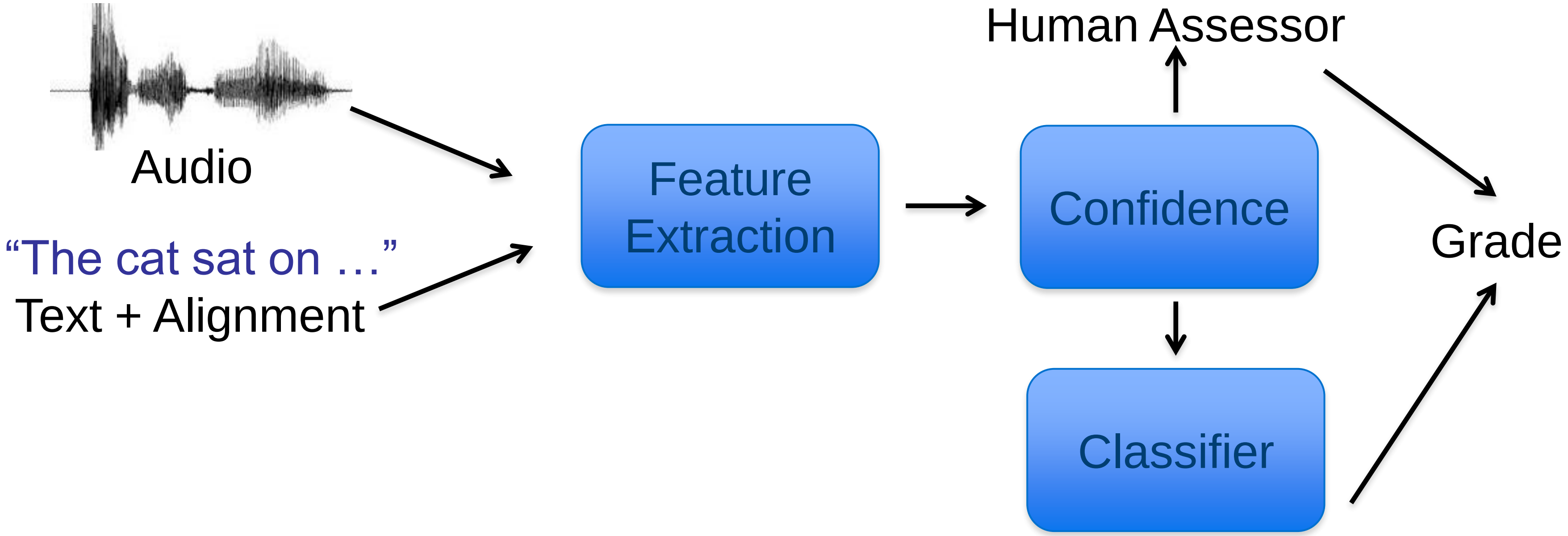
Minimal control over recording conditions

- Background speakers/noise

Limited (or no) language and acoustic model training data

- Useful for recognition system to transcribe disfluencies

Machine Learning for Assessment



Audio

“The cat sat on …”
Text + Alignment

Feature Extraction

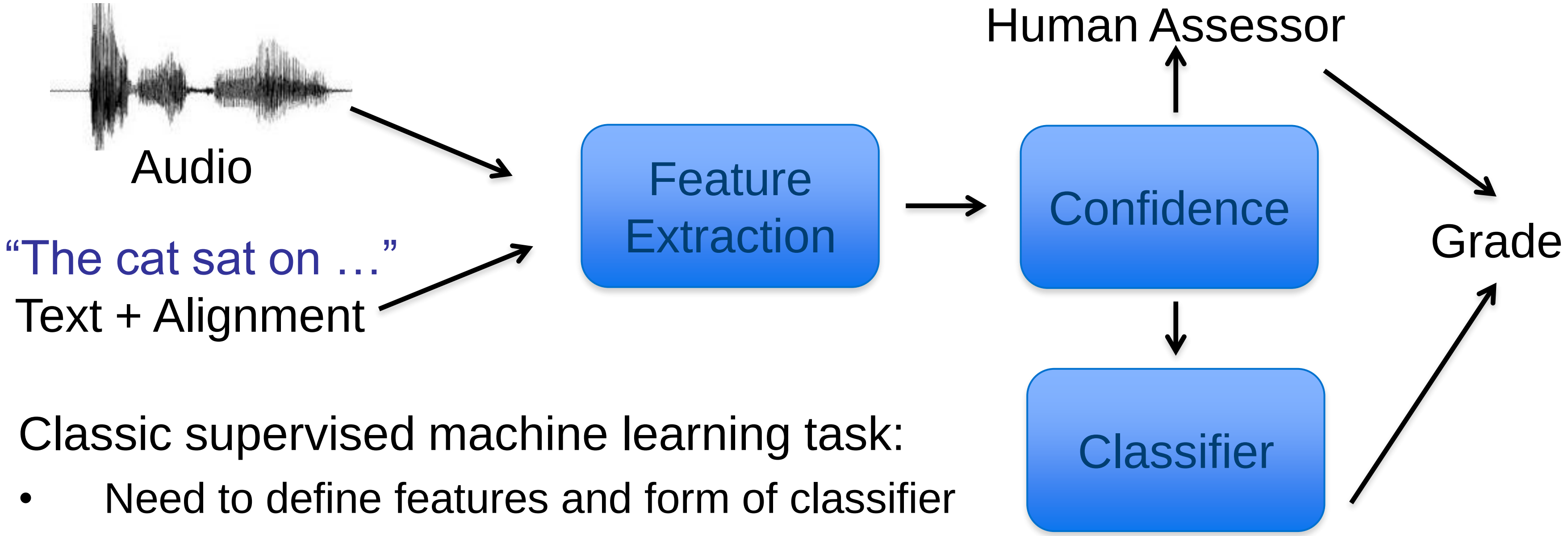
Confidence

Human Assessor

Classifier

Grade

Machine Learning for Assessment



Classic supervised machine learning task:

- Need to define features and form of classifier
- Detect “outliers” to pass to human assessor

Pronunciation Assessment

Assess how close pronunciation is to a native English speaker

- Mother tongue (L1) impacts speech of non-native English (L2)
- Phones from L2 missing from L1
- Pronunciation/prosody influence from L1

Pronunciation Assessment

Assess how close pronunciation is to a native English speaker

- Mother tongue (L1) impacts speech of non-native English (L2)
- Phones from L2 missing from L1
- Pronunciation/prosody influence from L1

Common form of current spoken language assessment

- Read sentences/limited domain responses
- Also used in Computer Aided Language Learning

Spoken Language Assessment

Currently domain of responses limited – short questions/story retelling

- Reduces recognition errors but limits spontaneity
- Limits ability to assess message construction (content)

Spoken Language Assessment

Currently domain of responses limited – short questions/story retelling

- Reduces recognition errors but limits spontaneity
- Limits ability to assess message construction (content)

Example features useful for assessing (unscripted) speech:

- Speaking rate (words per second)
- Mean duration of phones and silences between words
- Language model score (native)
- Acoustic model score

Challenges Moving Forward

Open question/discussion assessment – elicit spontaneous speech

- Speech recognition performance challenges

Currently extract general attributes of the word sequence

- Count/rate of words, number of unique words used
- Acoustic/language model scores

Does not assess:

- Construction of argument and coherence of response
- Relationship to topic to be discussed/described

Conclusions

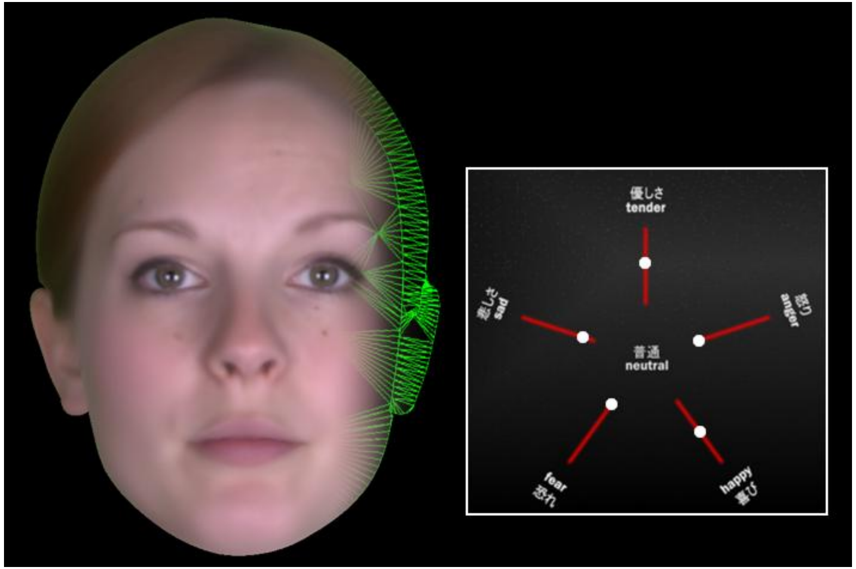
- Speech recognition is an essential component for automatic assessment of spoken language
- Current technology performance levels limits applications
  - Often fluency, not content, assessed
  - Only applicable to low-stake, practice, tests
- Spoken Language Processing technology development required
  - Not the same as Natural Language Processing!

Intelligent Interactive Agents for Assessment

System combines range of speech technologies:

- Spoken dialogue systems, speech recognition, expressive speech synthesis, audio-visual processing

(Image courtesy Toshiba CRL)


優しさ
tender
悲しさ
sad
怒り
anger
普通
neutral
fear
恐れ
happy
喜び