Speech Processing

LASC10061 (UG) & LASC11158 (PG)

Catherine Lai

21 Sep 2023

Course Variants

Undergraduate course code: LASC10061 (20 credits)

Postgraduate course code: LASC11158 (20 credits)

All course materials and assessments are the same for both versions.

- Different expectations for marking purposes

Course Materials: https://speech.zone/courses/speech-processing/

Emails/announcements, assignment submissions, online tests, library resource list on **Learn**

Teaching Staff

Lecturers

- Dr Catherine Lai (Course Organizer)
  - c.lai@ed.ac.uk
  - In-person office hour: Mon 3-4pm
  - http://homepages.inf.ed.ac.uk/clai/meetings/
- Dr Rebekka Puderbaugh
  - r.puderbaugh@ed.ac.uk
  - https://bit.ly/3wjSqGE
- Prof Simon King
  - simonk@ed.ac.uk
  - https://speech.zone/meet-me

You can always schedule a 1-1 meeting with us!

Tutors

- Atli Sigurgeirsson
- Zihang Peng

Talk to the tutors in the labs!

What is this course about?

This course is about computational methods for processing speech for the developing speech technologies

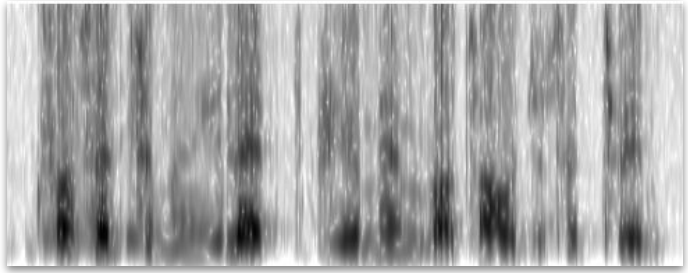
- Automatic speech recognition
- Speech synthesis

To understand the basic problems and current solutions, we need to understand what speech is and how we can capture it on a computer. This means learning about:

- Articulatory and acoustic phonetics
- Signal processing

How does automatic transcription work?

We capture speech as a (digital) waveform



We transform the waveform into a useful feature representation using signal processing and machine learning

We use statistical modelling/machine learning to map speech features to words

```
i said to him when you left do you
remember i told you i said to him don't
forget dave if you ever get in trouble
give us a call you never know your luck
```

From Audio to Text

We're going to focus on Hidden Markov Model based ASR in this course.

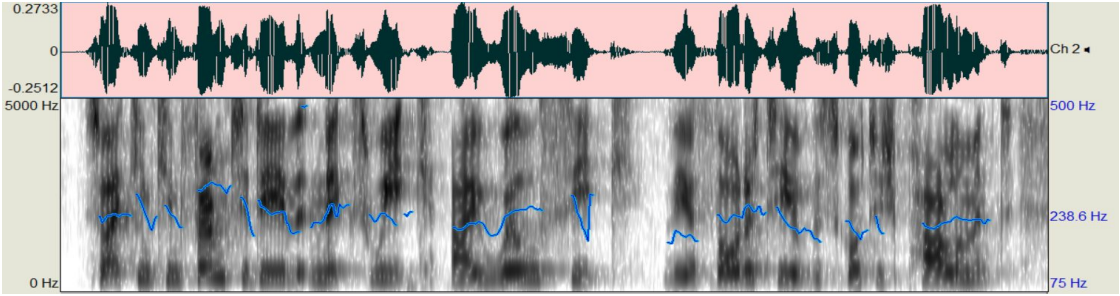
Feature Extraction

- signal processing

Statistical modeling, machine learning

- Acoustic model: Learn acoustic properties of speech sounds
- Language model: Learn probability of potential word sequences
- Use structured models to combine this acoustic and language model to get words

→ Automatic Speech Recognition (ASR)

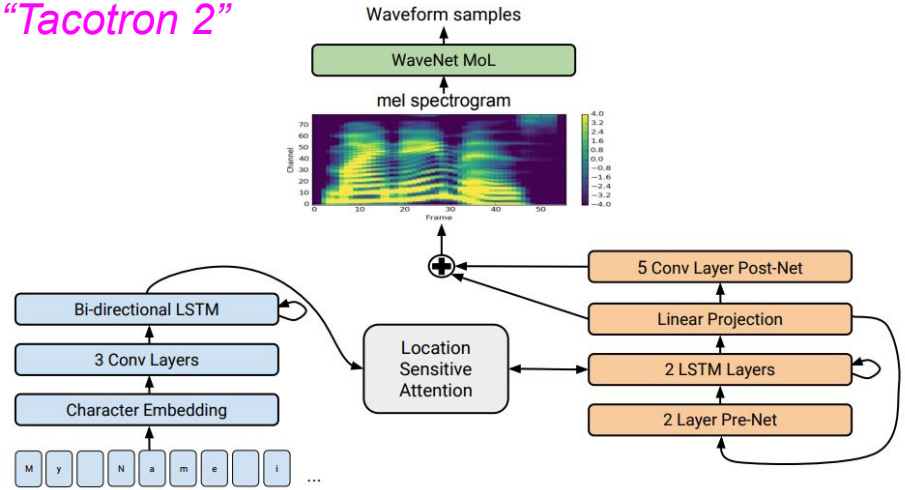


I CAN CONFIRM TODAY THAT NEXT WEEK I WILL SEEK THE AUTHORITY OF THE SCOTTISH PARLIAMENT TO **GIE** WITH THE U.K. GOVERNMENT THE DETAILS OF A SECTION THIRTY ORDER THE PROCEDURE THAT WILL ENABLE THE SCOTTISH PARLIAMENT TO LEGISLATE FOR AN INDEPENDENCE REFERENDUM THE DETAILED ARRANGEMENTS FOR THE REFERENDUM INCLUDING ITS TIMING MUST BE FOR THE SCOTTISH PARLIAMENT TO DECIDE THESE CONSIDERATIONS LEAD ME TO THE CONCLUSION THAT IF SCOTLAND IS TO HAVE A REAL CHOICE WHEN THE **THAMES OF BREAK AT A NORMAN** BUT BEFORE IT IS TOO LATE TO CHOOSE **OUT** OWN COURSE THEN THAT CHOICE MUST BE OFFERED BETWEEN THE AUTUMN OF NEXT YEAR TWENTY EIGHTEEN AND THE SPRING OF TWENTY NINETEEN

How can computers generate speech from text?

Statistical Parametric/Neural speech synthesis: Learn a statistical relationship between text features and acoustic features

E.g. "Tacotron 2"



Map from speech features to a waveform

Use machine learning methods to map from the text representation to a speech representation

Learn a statistically driven representation of text

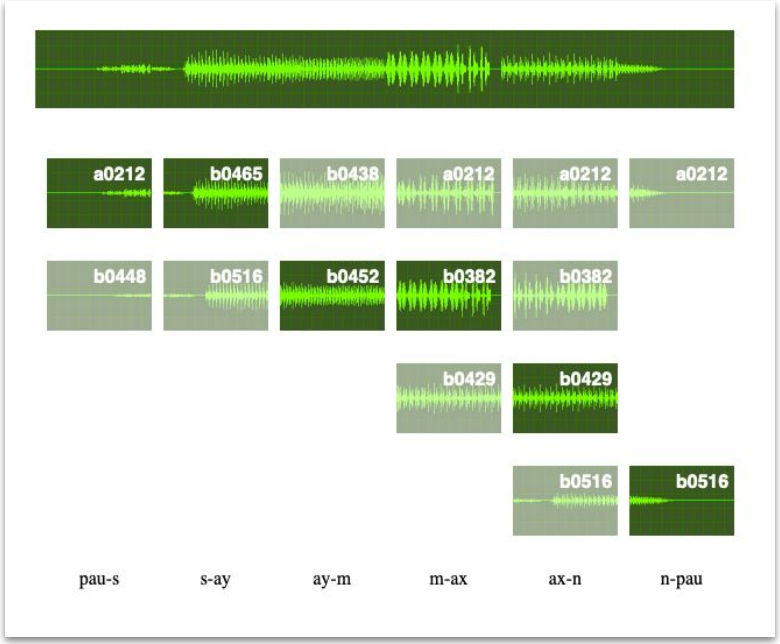
Shen et al. "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions." ICASSP 2018

From Text to Speech (a simpler method?)

Text-to-Speech (TTS) using Concatenative synthesis/unit selection:

- Concatenate bits of pre-recorded speech (units) together
- But how to select which bits?
- How to automate this? Use machine learning?
- https://speech.zone/interactive-unit-selection/

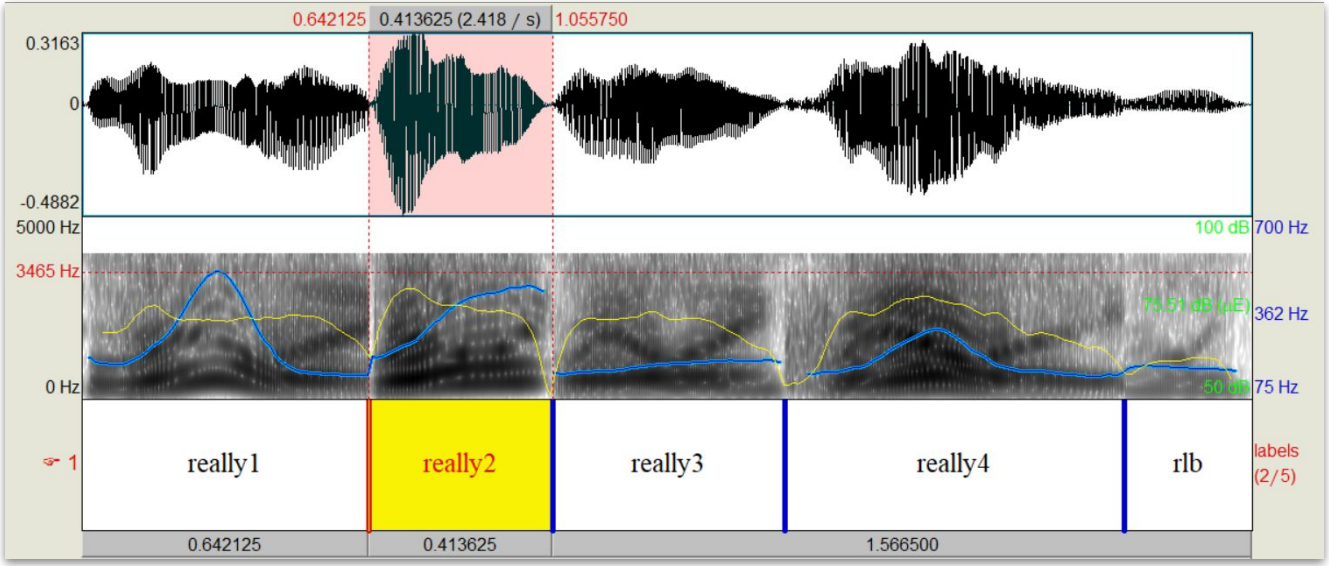
We're going to focus on concatenative synthesis in this class as a starting point.



Towards spoken language understanding

We need to remember that speech contains more than just the text, e.g., speech often varies in terms of prosody:

- e.g. Pitch, Loudness, Timing, Voice quality



Challenge: Mapping from letters to sounds is not always straightforward!

English is particularly bad for this: Lots of irregularity!

Take a look at the UK towns that are mispronounced the most...

1. Marylebone (London)

Wrong: Ma-ree-lee-bone

Right: Mar-lee-bone

2. Teignmouth (Devon)

Wrong: Tane-mouth

Right: Tin-muth

3. Bicester (Oxfordshire)

Wrong: Bi-ses-ter

Right: Bis-ter

4. Hunstanton (Norfolk)

Wrong: Hun-stan-ton

Right: Hun-ston

5. Cholmondeley (Cheshire)

Wrong: Chol-mon-de-lee

Right: Chum-lee

https://www.countryliving.com/uk/homes-interiors/property/a30805564/mispronounced-uk-towns/

Challenge: Transcription needs more than just the words

```
i said to him when you left do you remember i told
you i said to him don't forget dave if you ever get
in trouble give us a call you never know your luck
```

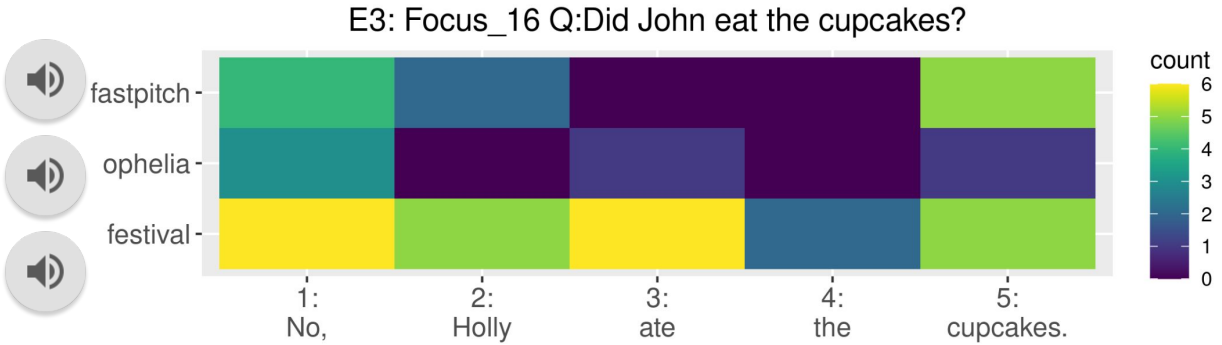
- I said to him: "When you left" (do you remember? I told you) I said to him: "Don't forget Dave if you ever get in trouble give us a call." You never know your luck."
- I said to him when you left: "Do you remember? I told you" I said to him: "Don't forget Dave if you ever get in trouble give us a call." You never know your luck."
- I said to him when you left, (do you remember? I told you) I said to him: "Don't forget Dave if you ever get in trouble give us a call. You never know your luck."

Challenge: Transcription needs more than just the words

```
i said to him when you left do you remember i told
you i said to him don't forget dave if you ever get
in trouble give us a call you never know your luck
```

- I said to him: “When you left” (do you remember? I told you) I said to him: “Don’t forget Dave if you ever get in trouble give us a call.” You never know your luck.”
- I said to him when you left: “Do you remember? I told you” I said to him: “Don’t forget Dave if you ever get in trouble give us a call.” You never know your luck.”
- I said to him when you left, (do you remember? I told you) I said to him: “Don’t forget Dave if you ever get in trouble give us a call. You never know your luck.”

Challenge: Speech is context dependent



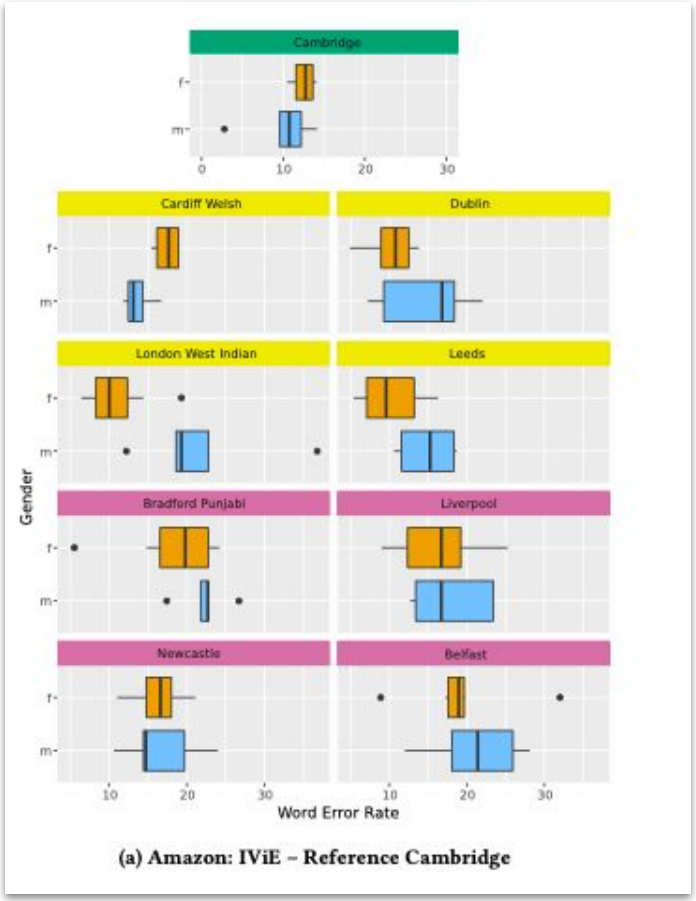
Annotator error markings on words

Current TTS voices sound humanlike, but context/speaker appropriateness is still a ways off. What are the problems? How can we make it better?

Gutierrez et al. (2021) "Location, Location: Enhancing the Evaluation of Text-to-Speech Synthesis Using the Rapid Prosody Transcription Paradigm." in SSW 2021

Challenge: Speaker/contextual variation, algorithmic bias

- We want ASR systems to provide accurate transcriptions for different environment condition and different speakers
- But current deployed systems still show significant performance differences between different accents of English
- There's still a lot of things to understand about how speaker variation can be modelled in ASR.



(a) Amazon: IViE – Reference Cambridge

N. Markl *Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition* In FAccT '22. ACM,https://doi.org/10.1145/3531146.3533117

What does speech processing involve?

- Understanding of what speech is
  - What is it physically? How do we perceive it?
- Development of the technology to represent speech
  - Foundations in human speech production/perception
  - Mathematics to describe representations/methods precisely
  - Algorithms for representing speech using computers (digital = discrete signals)
- Specification of speech processing tasks
  - What do ASR and TTS system actually do? What about other tasks?
  - Consider limitations in the tools available (costs and benefits)
- Methods for understanding whether we achieved our goals
  - How do we measure ASR performance? TTS? Spoken language understanding?
  - How to evaluate speech technologies in context

We need linguistics, cognitive science, computer science, electrical engineering...

Course Components: Modules

Phonetics (PHON):

- 1. Phonetics and Representations of Speech
- 2. Acoustics of Consonants and Vowels

Signal Processing (SIGNALS)

- 3. Digital Speech Signals
- 4. the Source-Filter Model

Modules (cont)

Text-to-Speech (TTS): *focusing on concatenative synthesis*

- 5. Speech synthesis - phonemes and the front end
- 6. Speech Synthesis - waveform generation and connection speech

Automatic Speech Recognition (ASR): *Focusing on HMM based ASR*

- 7. Speech Recognition - Pattern matching
- 8. Speech Recognition - Feature engineering
- 9. Speech Recognition - the Hidden Markov Model
- 10. Speech Recognition - Connected speech & HMM training

Learning Outcomes for this course

- Understand **human speech production and perception**, including the use of tools for visualising and manipulating speech
- Give an overview of the components of **automatic speech recognition** and **speech synthesis** systems and describe a simple version of each component
- **Understand what the difficult problems are** in automatic speech recognition and speech synthesis
- **Perform experiments** with speech technology systems and **relate theory to practice**
- See how knowledge and skills from different areas come together in an **interdisciplinary field**

What background do you need?

This course involves:

- **Linguistics:** phonetics/phonology, sociolinguistics is increasingly helpful
- **Mathematics:** statistics, probability, (a bit of) linear algebra and calculus
- **Computer science:** algorithms, data structures
- **Engineering:** practical implementations, working with real audio

There will be something new for everyone!

If everything on that list is new, you'll probably find the course very hard!

Does speech processing require a lot of maths?!?

In the short term: no, but in the long term: yes!

- Most speech technology is built on mathematical foundations!
- We generally won't go into a lot of mathematical detail in this course
- We won't expect to you to solve complex maths problems for assessment in this course
- BUT if you want to go further (e.g. in semester 2 ASR, TTS courses and beyond), you will need to get comfortable using mathematical notation
- It may seem hard now, but in the long-term it will make life easier!
- You CAN still make a contribution to this field if you're not into maths but it's still good to understand the foundations!

You may find our Speech and Language Processing prep pointers on speech.zone helpful.

Delivery - in-person, live

Lectures:

- Thursday 9-11am, in-person, recorded

Labs:

- Wednesday 9-11am or 4:10-6pm
- In-person: PPLS AT computer lab, AT 4.02

Go to the lectures and one of the two labs sessions per week

How to approach the material

- Videos - get an introduction to the concepts, watch before the lecture
- Lecture - more on the concepts, flipped class-room style exercises
- Readings - consolidate/deepen your understanding
- Forums - ask questions for topics you are unsure of or want to know more about!
- Lab exercises - get a taste of the practical side of speech processing

Classes will help you synthesise together all the different sources of information and different ways of learning. No single source will be enough on its own.

Labs

- Work through practical exercises at your own pace
- 2 staff members per lab - ask them for help if you get stuck!
- Talk to each other!

The labs we have been allocated are linux computers with the various software you will need already installed.

You can also use the remote desktop option (see Module 0 on speech.zone)

Assessments: Online tests

- 30% of final grade
- Multiple choice questions:
  - Test open for 2 days, but once you start you need to complete within 1 hour here on Learn
- Due dates:
  - Phon/Signals (10%): Mon 12pm 16/10/23 - Wed 12pm 18/10/23
  - TTS (10%): Mon 12pm 30/10/23 - Wed 12pm 1/11/23
  - ASR (10%): Mon 12pm 27/11/23 - Wed 12pm 29/11/23
- Do these tests on LEARN

Marks may be scaled/moderated to fit the common marking scheme

Written Assessments

- Written Assessment 1 (30%):
  - lab report on the Speech Synthesis (Festival) assignment (1500 words)
  - Identify the type, origin and implication of errors in a text-to-speech system
  - due date: **6th November 2023 by 12 noon**
- Written Assessment 2 (40%)
  - lab report on the Automatic Speech Recognition assignment (3000 words)
  - Build a simple spoken digit recognizer
  - Run and write up experiments to test it's robustness given different training data, model settings,...
  - due date: **7th December 2023 by 12 noon**

Different levels of learning materials

- **Essential:** Need to understand, may be examinable
- **Recommended:** Not directly examinable but incorporating concepts from recommended materials (correctly!) will help you get higher marks
- **Extra/Extension:** This will deepen your understanding but not examinable; only for your own interest or fun! (generally more maths stuff)

We have a positive marking policy: you won't get penalized for writing incorrect things (you just won't get credit for it).

What now?

This and next week will give a brief introduction to phonetics

- Watch videos for Module 1 (this week) and Module 2 (next week)
- Do the readings for Module 1
- Look at the lab materials - get started with Praat

If you already know a lot about articulatory and acoustic phonetics, but not so much about maths, now would be a good time to start on Sharon Goldwater's maths tutorials:

Plus other resources linked here: https://speech.zone/courses/prepare-for-study-in-speech-and-language-processing/brush-up-your-mathematics/

Where to get help

In-person

- Come to lab session and talk to tutors and lecturers
- Come to Catherine’s office hours (Monday 3-4pm)
- Book an appointment with Catherine, Simon or Rebekka

Online

- Post questions on the speech.zone forum
- Email the lecturers

For administrative matters contact Catherine as the course organiser

Questions?

Delivery - Asynchronous

Additional course materials on speech.zone

- Videos - by Simon King and Rebekka Puderbaugh
- Readings - available online through the library
- Forums - lots of handy answers from previous years