Speech 101: Introduction to speech processing

Herman Kamper

This is a self-paced course. The goal is to get you to modern speech recognition and synthesis as quickly as possible. As a concrete aim, I want to get you to a point where you can read the HuBERT paper (Hsu et al. 2021) and understand how speech recognition is performed on top of HuBERT with the CTC loss (Jurafsky and Martin 2025, Sec. 16.4). Speech models are typically combined with a text-based language model (Jurafsky and Martin 2025, Sec. 16.3). So along the way another aim would be that you have enough background to read some of the GPT paper (Radford et al. 2018), which introduces arguably the most important language modelling approach of our time.

The strategy for this course will be to go almost from scratch. It is completely okay if you do not understanding everything immediately: there is quite a lot and it normally takes time for things to sink in.

For most of my own videos, you can find links to the slides or accompanying notes in the video comments.

Please let me know if you find any errors or typos in this document.

Basic signal processing and speech science

- Introduction to speech features

  These videos describe feature extraction were an audio waveform is converted to a sequence of higher-level acoustic vectors. It is worth noting that, although you will still find such acoustic features in practice, many end-to-end speech models these days take in raw speech samples directly. I.e. if the audio signal is sampled at 16 kHZ, then the input sequence $x_{1:T}$ to the speech model would consist of $16\,000$ floating values for every second of audio input.

- Dynamic time warping

  This is an old-school technique that is rarely used to do speech recognition directly these days. But it pops up quite often in low- and zero-resource speech applications.

- Speech 101 by Roger Moore [slides]:
  - Part 1: Sound, speaking, hearing (48 min)

  - Part 2: Sounds and symbols, articulatory phonetics, acoustic phonetics (44 min)
  - Part 3: Phonology, prosody, behaviour (1 h)

Machine learning

You need to get to a point where you understand neural networks, including architectures like recurrent neural networks (RNNs), convolutional neural networks (CNNs) and transformers. If you watch/read the following in the given order, then you should be there:

- Introduction to machine learning
- Linear regression:
  - Simple linear regression (14 min)
  - Vector and matrix derivatives (13 min)
  - Multiple linear regression - Model and loss (16 min)
  - Polynomial regression and basis functions (15 min)
  - Overfitting (10 min)
  - Regularisation (15 min)
  - Evaluation and interpretation (11 min)
- Training, validating and testing (18 min)
- Maximum likelihood estimation (20 min)
- Classification:
  - Task (9 min)
  - K-nearest neighbours (15 min)
- Logistic regression:
  - Model and loss (14 min)
  - Gradient descent - Fundamentals (11 min)
  - Optimisation (7 min)
  - Basis functions and regularisation (6 min)
  - Multiclass - One-vs-rest classification (5 min)
  - Multiclass - Softmax regression (15 min)
- Preprocessing
- Introduction to unsupervised learning (19 min)
- K-means clustering:
  - Algorithm (16 min)
  - Details (14 min)
- Introduction to neural networks:
  - Neural network preliminaries: Vector and matrix derivatives (5 min)

  - Neural network preliminaries: The chain rule for vector derivatives (7 min)
  - Neural network preliminaries: Gradient descent (4 min)
  - Neural network preliminaries: Logistic regression, softmax regression and basis functions (6 min)
  - From logistic regression with basis functions to neural networks (19 min)
  - Why is it called a neural network? (4 min)
  - Backpropagation (without forks) (31 min)
  - Backprop for a multilayer feedforward neural network (4 min)
  - Computational graphs and automatic differentiation for neural networks (7 min)
  - What is the difference between negative log likelihood and cross entropy? (in neural networks) (9 min)
  - Neural networks in practice (7 min)

- Before moving on to the videos below, it might actually be good to first cover the content in the language modelling section below and then come back here and continue.

- Recurrent neural networks:

  - D2L Sec. 9.4
  - From feedforward to recurrent neural networks (15 min)
  - RNN language model loss function (9 min)
  - RNN definition and computational graph (3 min)
  - Vanishing and exploding gradients in RNNs (13 min)
  - Solutions to exploding and vanishing gradients (in RNNs) (10 min)
  - Extensions of RNNs (8 min)

- Convolutional neural networks:

  - D2L Sec. 7.2

  - VGG convolutional neural network practical

    You only have to look at Part 1. This is part of the documentation for MatConvNet package (which I do not use), but I found this tutorial very helpful for getting a handle on convolutional neural networks.

    Most end-to-end speech recognition models use RNNs or transformer layers before the final classification layer, but often the early layers of the speech recognition system would be convolutional.

- Encoder-decoder models, machine translation and attention:

  - A basic encoder-decoder model for machine translation (13 min)
  - Training and loss for encoder-decoder models (10 min)
  - Encoder-decoder models in general (18 min)
  - Greedy decoding (5 min)
  - Beam search (18 min)
  - Basic attention (22 min)

- Self-attention and transformers

Language modelling

Natural language processing (NLP) typically refers to text processing (rather than speech processing). But speech recognition systems output text. So many NLP techniques are also relevant in speech recognition systems. Most notably, text-based language models (trained only on text) are often integrated into a speech recognition system. The content below introduces the basics of NLP so that you are able to understand state-of-the-art text-based language models.

- What is natural language processing?
- A first NLP example
- Text normalisation and tokenisation
- Byte-pair encoding (BPE)
- Edit distance

  To evaluate speech recognition systems, the output from the model is often compared to the transcription produced by a human. Normally the number of words in the two sequences don't match up exactly, so the edit distance algorithm is used to produce an alignment between the predicted and ground truth sequences. This alignment is then used to calculate a word error rate (WER); see the next section for more details on this metric.

- Language modelling with N-grams:
  - The language modelling problem
  - N-gram language models
  - Why use log in language models?
- Neural networks examples: Natural language processing

Large language models

- Intro to large language models by Andrej Karpathy (1 h)
- Large language model training and inference (14 min)
- The difference between GPT and ChatGPT (13 min)
- Reinforcement learning from human feedback (15 min)

Speech recognition

- Encoder-decoder models for speech recognition:
  - Sec. 16.3 of (Jurafsky and Martin 2025, Chap. 16)
- Sequence modelling with CTC
- CTC-based speech recognition:
  - Sec. 16.4 of (Jurafsky and Martin 2025, Chap. 16)

- Evaluating speech recognition systems with word error rate:
  - Edit distance video above
  - Sec. 16.5 of (Jurafsky and Martin 2025, Chap. 16)
- HuBERT
  - Read together
  - HuBERT forms the basis of many speech recognition models, but are also used as the basic feature extraction for unsupervised learning models and even speech synthesis systems.
- wav2vec 2.0

  This model is often used as the basis for speech recognition, e.g. on Hugging Face. A multilingual variant called XLSR (Conneau et al., 2020) is often used to initialise speech recognition models when targetting a new language.

Speech synthesis

The terms "speech synthesis" and "text-to-speech" are sometimes used interchangably.

- Tacotron
  - Caused the big shift towards neural speech synthesis
- Very attentative Tacotron (?)

Future

- Explain what forced alignments are

Further reading

- Hugging Face NLP tutorial
- Hugging Face audio tutorial