Introduction to Computational Linguistics
Session 6: Automatic Speech Recognition

Stephen Bodnar

Universität Tübingen

December 6, 2023

Introduction

- So far, we've mainly covered text technology
- Computational Linguistics can also involve speech processing
- Research at intersection of speech and technology is huge field
  - Professional associations
    E.g. ISCA, (https://www.isca-speech.org/iscaweb/)
  - Large international conferences,
    E.g., Interspeech Conference (https://www.interspeech2023.org/)
  - Academic Journals
    E..g., Speech Communication
    (https://www.sciencedirect.com/journal/speech-communication)
- Speech-interactive interfaces in many places
  - Digital Assistants
  - Document dictation
  - Voice search
  - Spoken dialogs
  - Language training

Plan

- A simplified linguistic description of speech
- Representing speech with computers
- Different speech processing applications
- An overview of Automatic Speech Recognition

Content for today adapted from

- [Jurafsky and Martin, 2009]
- [Lecorvé, 2023]
- [Bell, 2023]

A brief introduction to speech

- Area that studies speech sounds called phonetics:

  study of linguistic sounds, how they are produced by the articulators of the human vocal tract, how they are realized acoustically, and how this acoustic realization can be digitized and processed. (Jurafsky and Martin, 2009, p. 215)

Units of Speech - Phones

Can think of speech as consisting of different units, at different levels:

- One intuitive, everyday unit is the word (e.g., cat, operationalisation)
- Words composed of smaller, common units of speech called *phones* (also known as segments)
- Phones can be represented with written symbols, e.g. International Phonetic Alphabet.
  - Mapping between symbols and sounds transparent (compare with English, which is much more opaque)
  - E.g. /kæt/
  - E.g. /ɒpəˈɹeɪʃ(ə)nəlʌɪzeɪʃ(ə)n/

Units in Speech - Prosody

Can also describe speech at a level above phones:

- syllable stress (e.g., **pre**sent vs. pre**sent**)
- intonation (e.g., questions vs. statements)

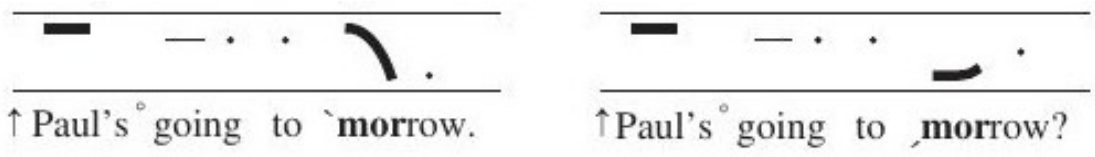


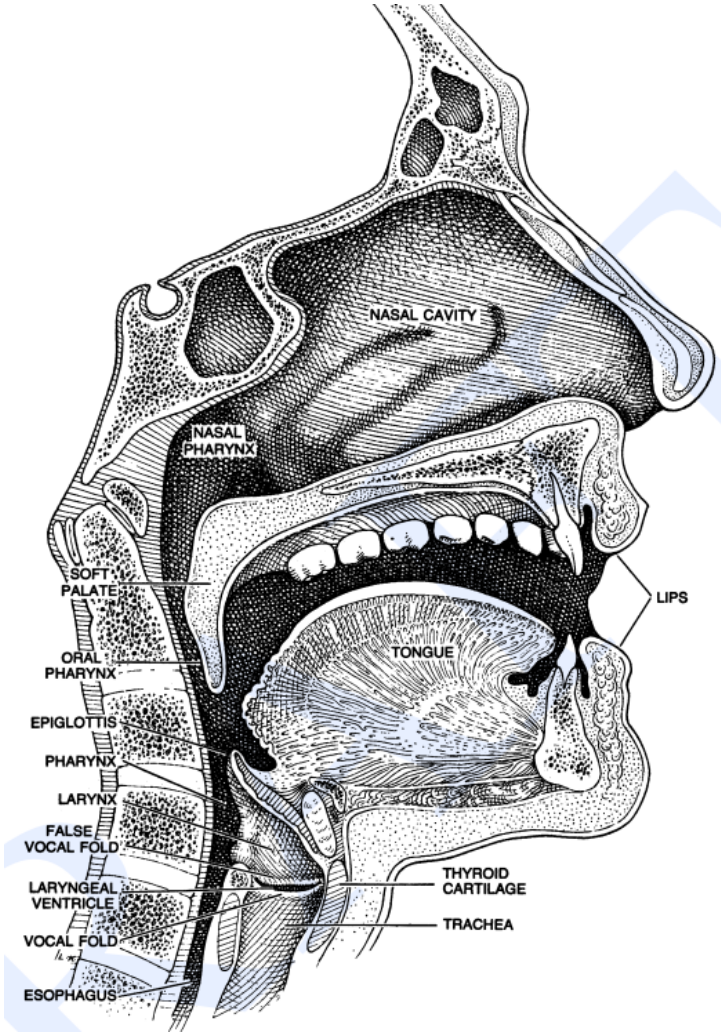
Image source: Collins, Beverley; Mees, Inger M. 2003. The Phonetics of English and Dutch. 5th ed. Leiden: Brill., Downloaded from https://forum.language-learners.org/viewtopic.php?t=14685

- voice quality (e.g., creakiness [Link])
- Known as prosodic or suprasegmental aspects of speech

- For ASR, segmental aspects of speech traditionally most important, so we focus on phones here

Production of phones

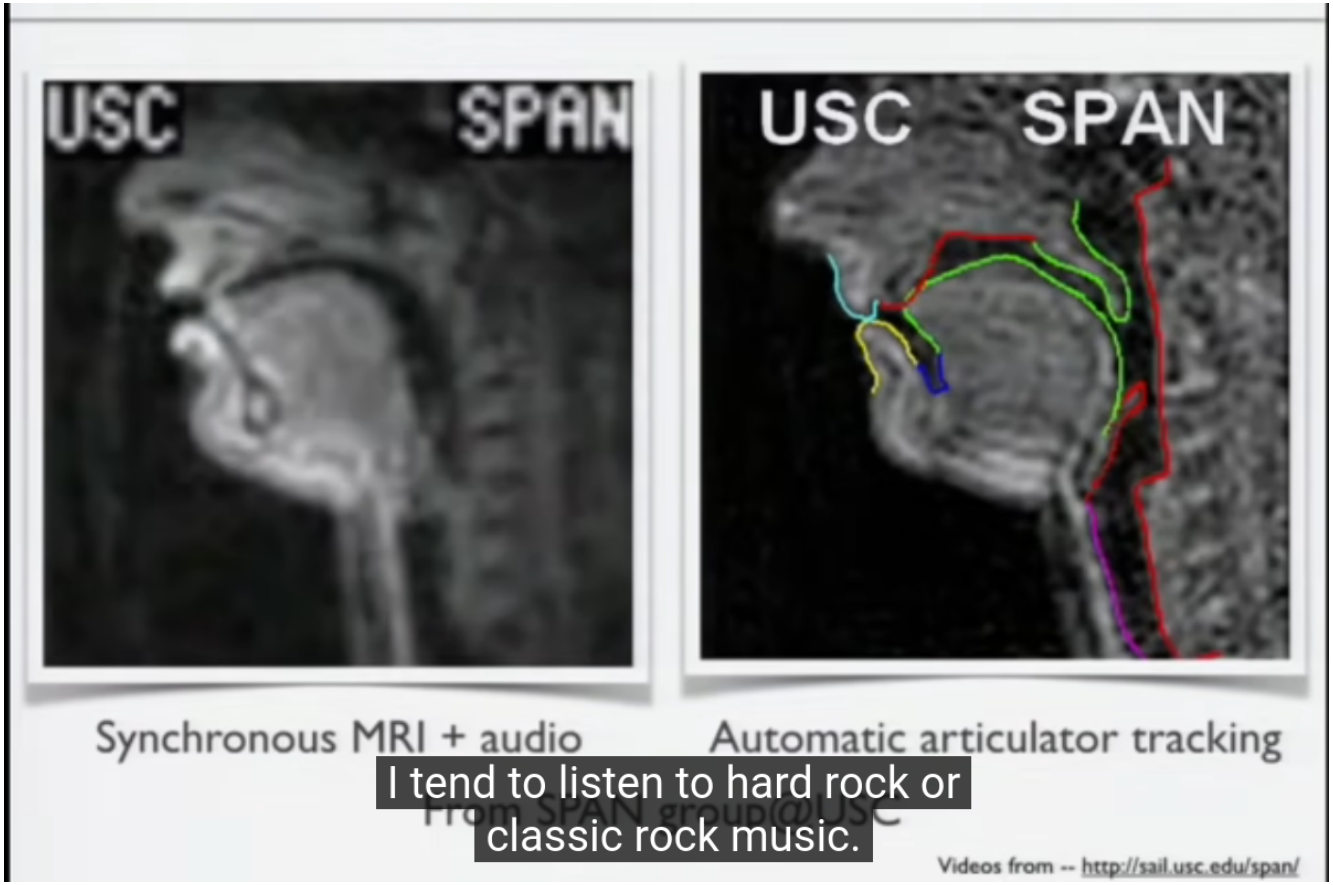
- Speech produced with vocal tract (Image from Jurafsky & Martin, 2009, p. 219)



Production of phones

- The different components of the tract allow you to make the different sounds we know as vowels and consonants
- Different vocal tract configurations produce different phones
- Special terminology. Some examples:
  - Voiced vs unvoiced sounds: With or without vocal chord vibration (e.g., [s] vs. [z], [d] vs. [t])
  - Stop: consonant where airflow blocked for short time (e.g., [b], [k])
  - Fricative: airflow open but restricted (e.g., [f], [z])

Production of phones - Example



Live MRI of vocal tract in action (Microsoft Research) [Link]

Phone variation and Phonemes

- Lots of variation between in how phones are realised, due to in-word position, style of speech (reading vs. spontaneous conversation), etc.
- Phonemes: the abstract class representing categories of meaningful sounds in a language
  - E.g., /t/

Phone variation - Allophones

- Allophone: all the different surface realisations of a phoneme, in different contexts (Table from Jurafsky & Martin, 2009, p.226)

IPA	ARPABet	Description	Environment	Example
t^h	[t]	aspirated	in initial position	*toucan*
t		unaspirated	after [s] or in reduced syllables	*starfish*
ʔ	[q]	glottal stop	word-finally or after vowel before [n]	*kitten*
ʔt	[qt]	glottal stop t	sometimes word-finally	*cat*
ɾ	[dx]	tap	between vowels	*butter*
$t^{\neg}$	[tcl]	unreleased t	before consonants or word-finally	*fruitcake*
t̪		dental t	before dental consonants ([θ])	*eighth*
		deleted t	sometimes word-finally	*past*

Figure 7.9 Some allophones of /t/ in General American English.

Summary: Phones are the messy, real-world realisations in speech, and phonemes are the abstract labels we use to categorize sounds together.

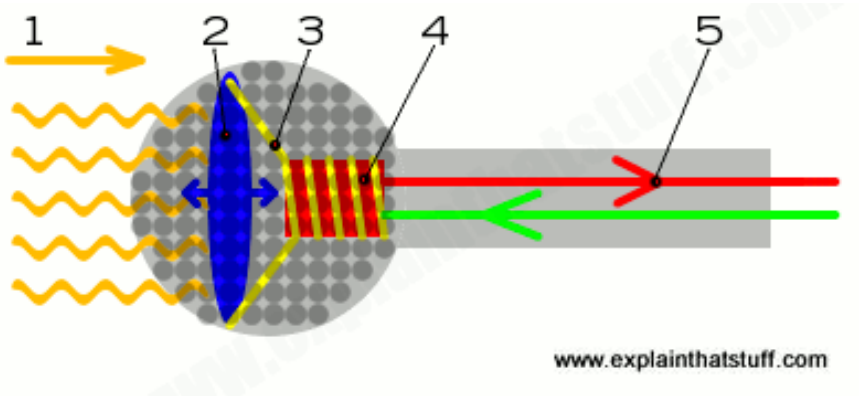
Plan

- A simplified linguistic description of speech
- => Representing speech with computers
- Different speech processing applications
- An overview of Automatic Speech Recognition

Speech and computers

Speech as a signal

- Speech as a sound wave propogating through medium
- Speech as an analog signal, coming through microphone
- Sound waves can be transduced: conversion of a sound into an electric signal by a microphone, ear

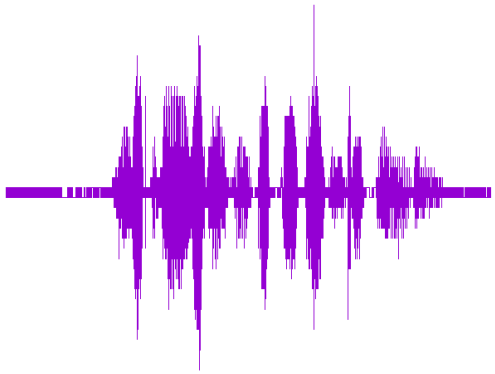


Digitizing Speech

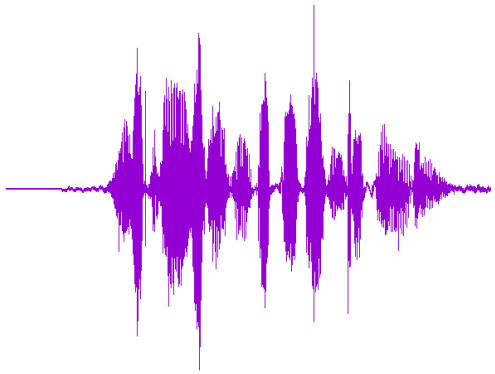
- Sampling: periodically measure the electrical current when sound is transduced with a microphone (e.g., 44100 times / second); Store electric signal samples as sequence bits and bytes
- Quantization: representing real valued numbers as integers (like a digital camera, potentially but not necessarily lossy)

Encoding Sound

- Early computers had 8bit sounds, with 2^8, or 256 amplitude steps for entire range



- In contrast, cd audio uses 16bit numbers, $2^{16} = 65536$ steps



Computer plays back speech by processing the sequence and using electric current to move the speaker appropriately

Plan

- A simplified linguistic description of speech
- Representing speech with computers
- => Different speech processing applications
- An overview of Automatic Speech Recognition

Speech Processing Applications

Instead of playing the audio, we can feed the audio into other kinds of computer programs to do interesting things:

- Automatic Transcription (we'll look at this today)
- Spoken Dialog Systems
- Speaker Recognition
- Diarization
- Emotion Recognition in Speech
- Pronunciation / spoken intelligibility scoring
- Language Identification

Spoken Dialogue Systems

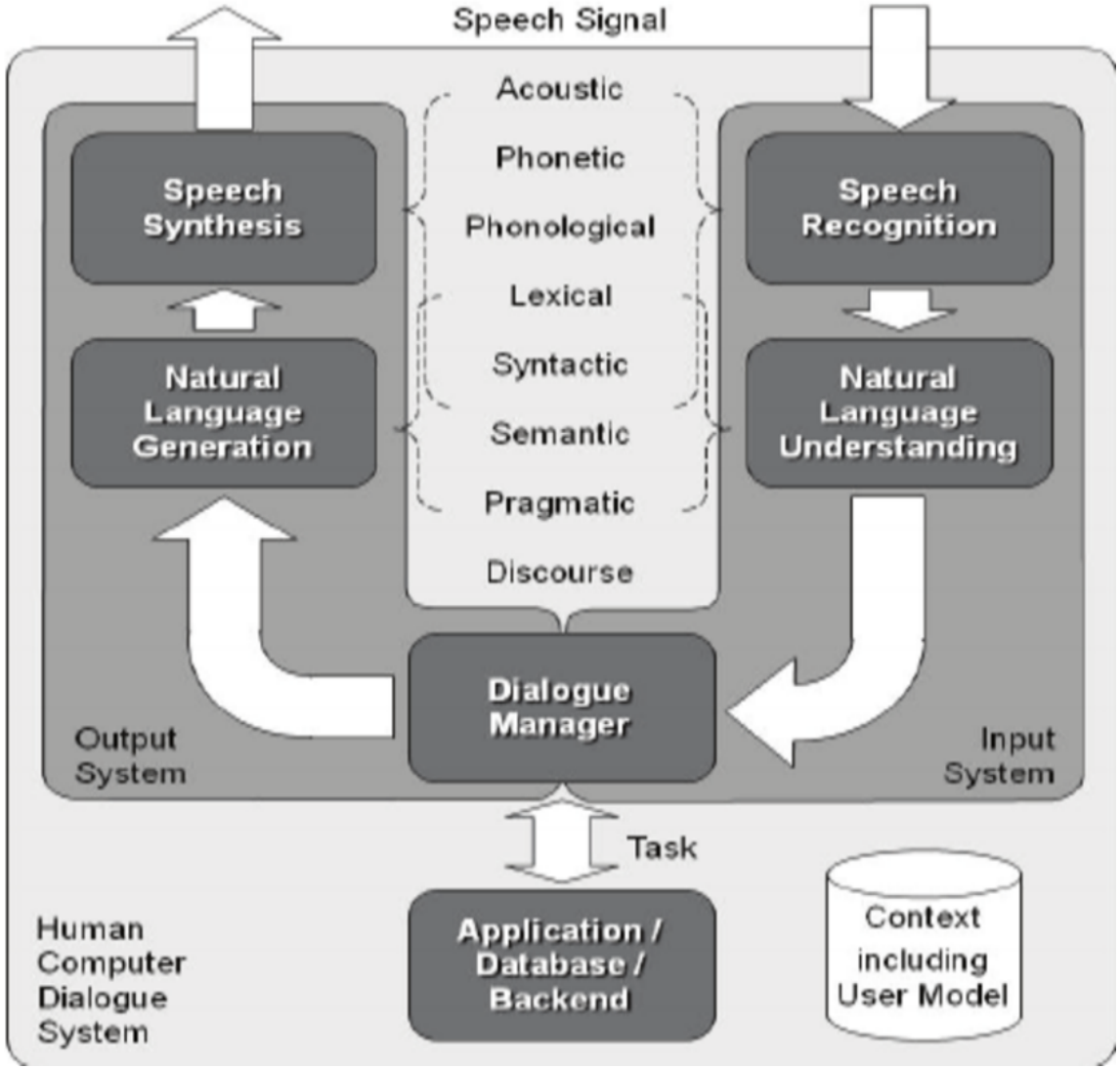


Diagram from [Le Bigot et al., 2008]

Speaker Recognition

Speaker Identification:

- Questions like:
  - Which of the N speakers I know about produced this utterance?
  - Was this recording produced by Speaker X or not?
- Can use for Speaker Verification (voice as password)
- Forensic Applications:
  - "process of determining if a specific individual (suspected speaker) is the source of a questioned voice recording (trace). " [Drygajlo, 2009]
  - Tools: e.g. https://www.atis-systems.com/en/automatic-forensic-speaker-identification/
  - Textbook: https://www.routledge.com/Forensic-Speaker-Identification/Rose/p/book/9780415271820
- Also used in Speaker Diarization ...

Speaker Diarization

- "determining 'who spoke when' in a long multi-speaker audio recording, marking the start and end of each speaker's turns in the interaction." (Jurafsky & Martin, 2023, p. 23)

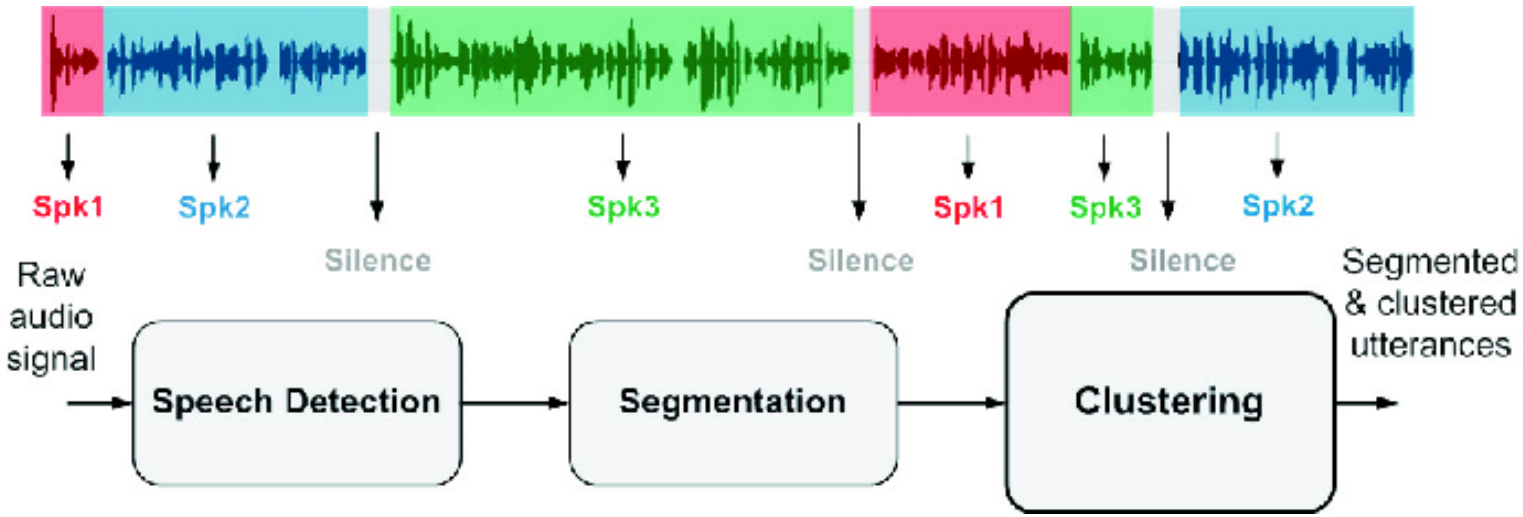


Image from https://ieeexplore.ieee.org/document/5989833

Emotion Recognition in Speech

- "automatically identifying the emotional or physical state of a human being from his or her voice." [Ververidis and Kotropoulos, 2006]
- In addition to phones, also needs prosodic information (e.g., pitch, etc., see article)

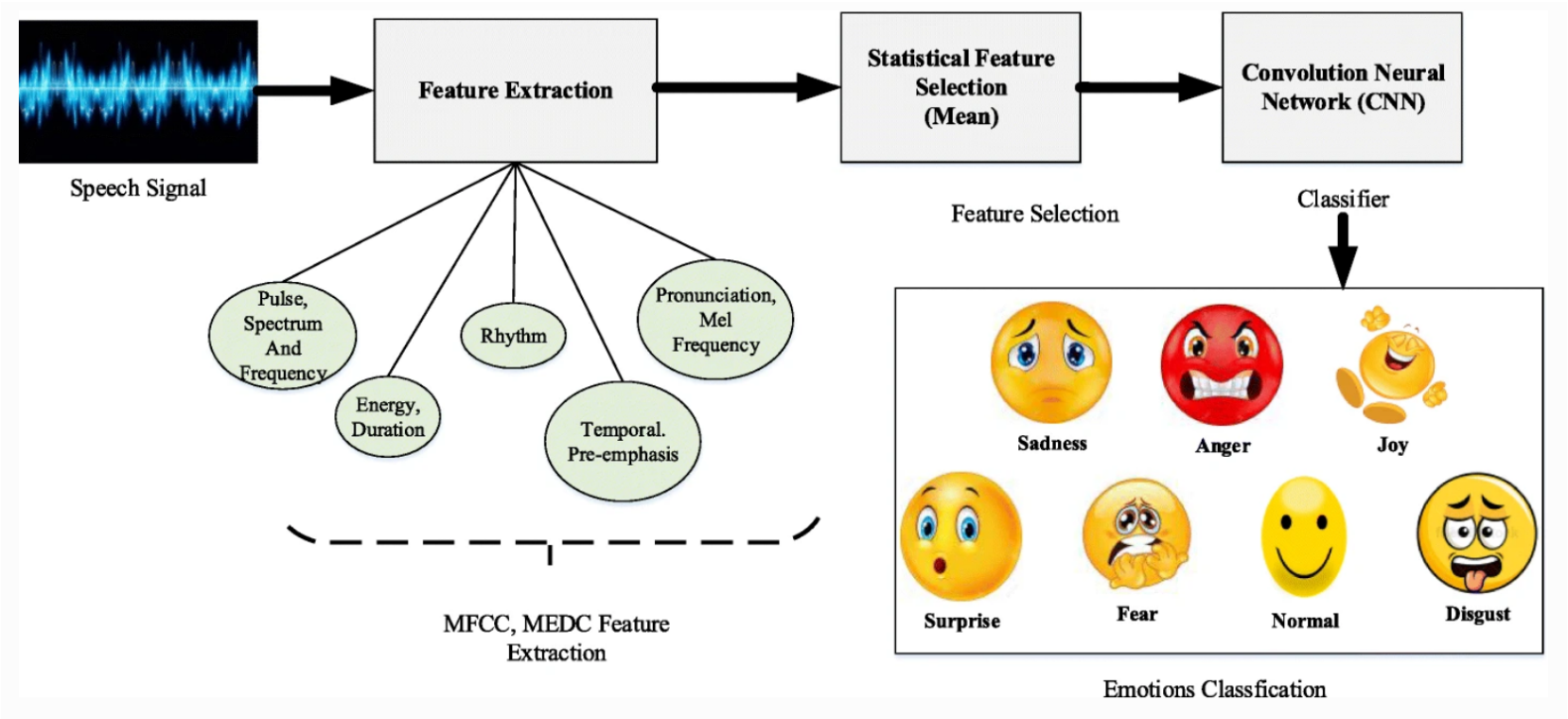


Image from https://link.springer.com/article/10.1007/s11042-020-10329-2

Pronunciation / Spoken Intelligibility Scoring

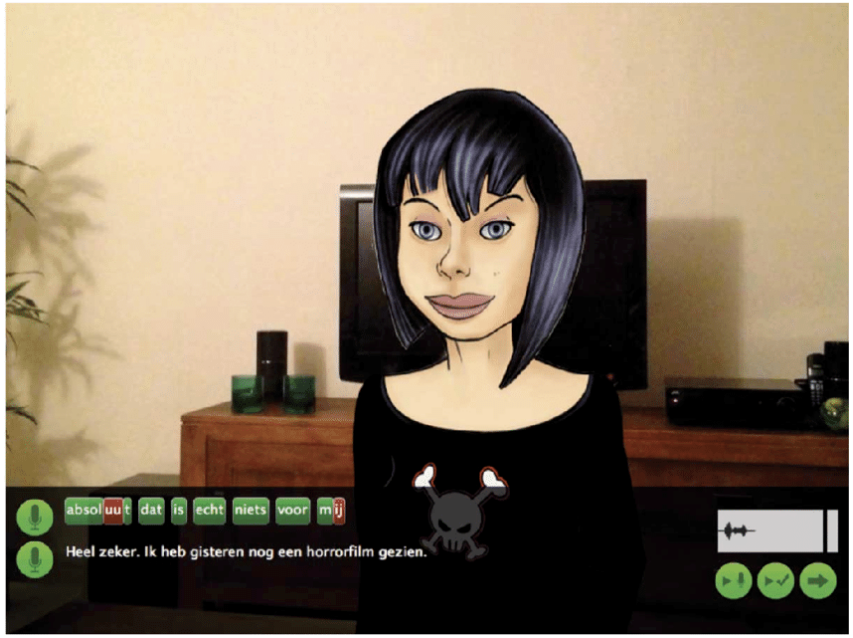


Image from [O'Brien et al., 2018]

Plan

- A simplified linguistic description of speech
- Representing speech with computers
- Different speech processing applications
- => An overview of Automatic Speech Recognition

A look at Automatic Speech Recognition

- Input is speech signal
- Output is string of words user said
- Need to map sounds to phoneme sequences
- From ambiguous phoneme sequences, need to choose best word sequence
- Types of ASR
  - continuous vs discrete word
  - large vocabulary / small vocabulary

Challenges in ASR (why it's a hard problem)

- Ambiguity,
  E.g. "recognize speech" vs. "wreck a nice beach"
- Variety in speech
  - accent
  - gender
  - age
  - L2
- Types of speech
  - E.g., read speech, spontaneous speech, emotional speech
- Environmental factors
  - background noise
  - microphone quality
- Single speaker or multiple speakers (e.g., meeting), possibly overlapping turns

Approaches in ASR

Many changes, especially in last few years

- Dynamic Time Warping
- Statistical approaches (classical approach)
- Deep Learning approaches
- End-to-end models

We will look at 'classical' statistical speech recognition today

Statistical Approach to ASR

Noisy Channel Model

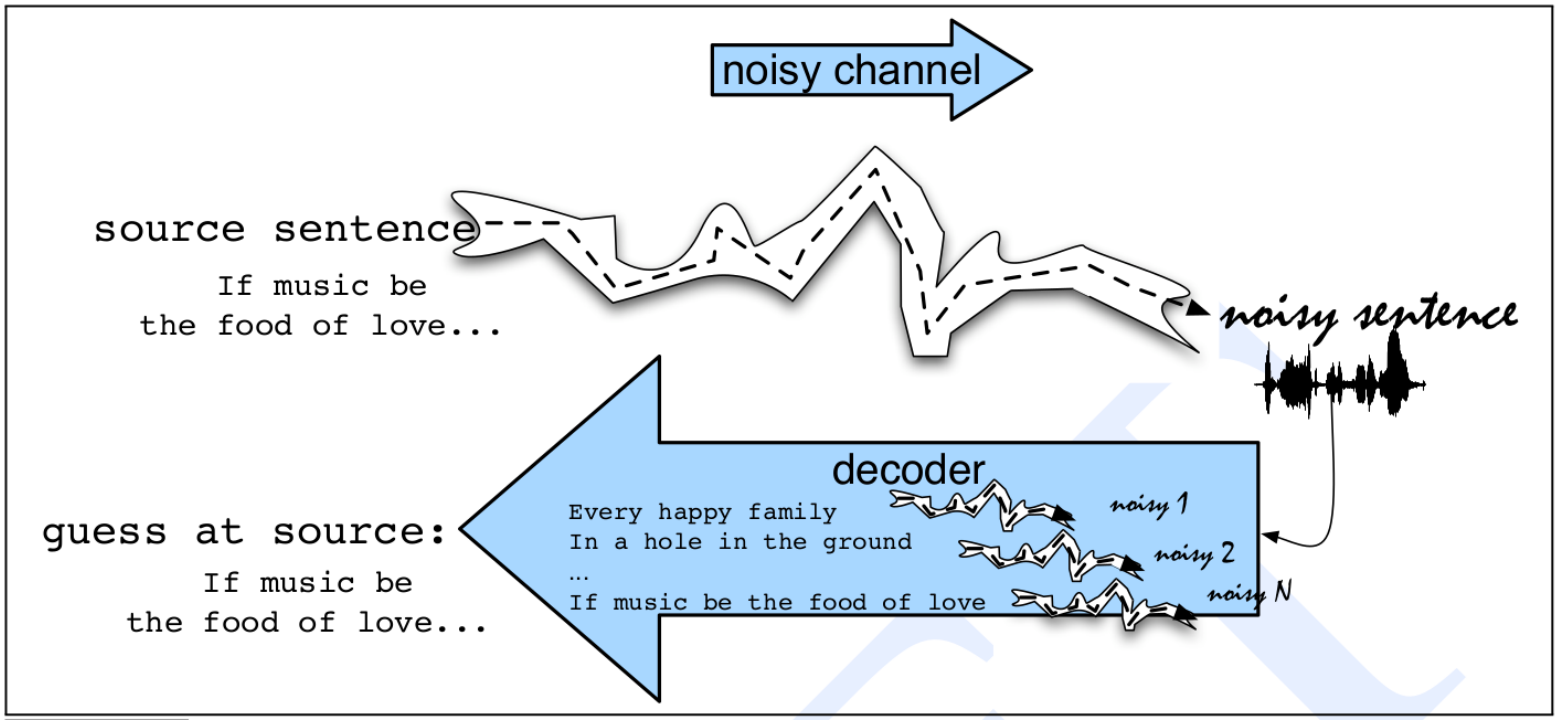
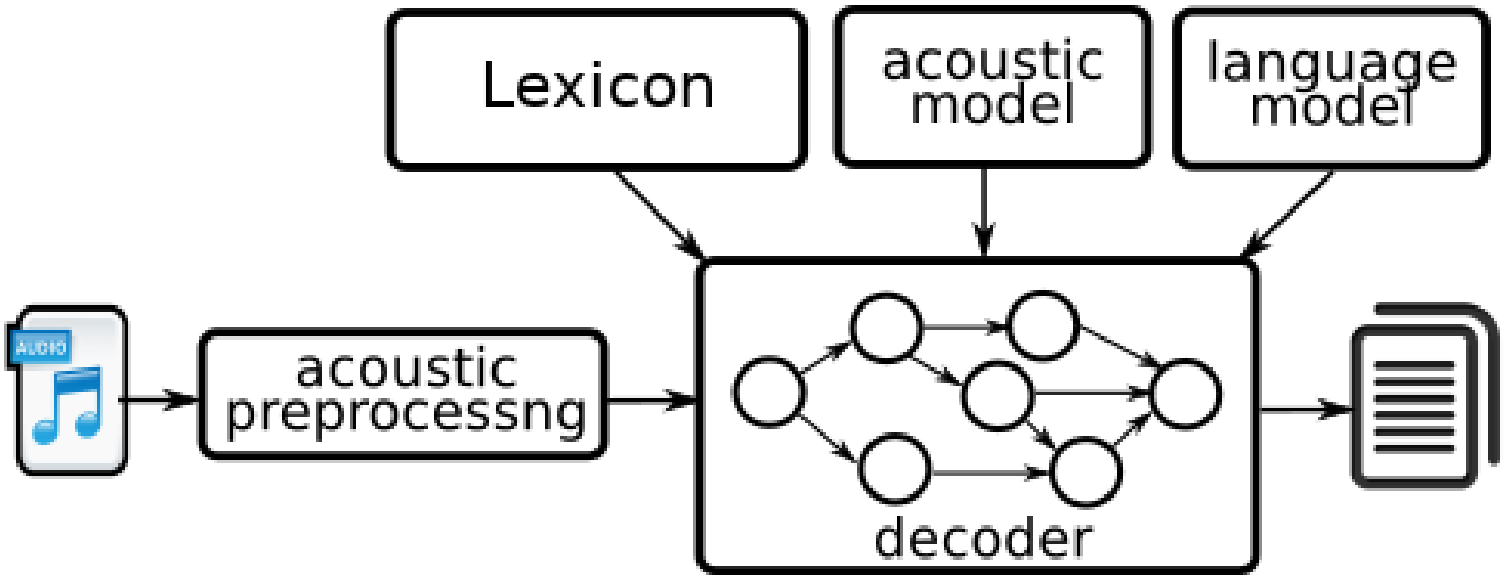
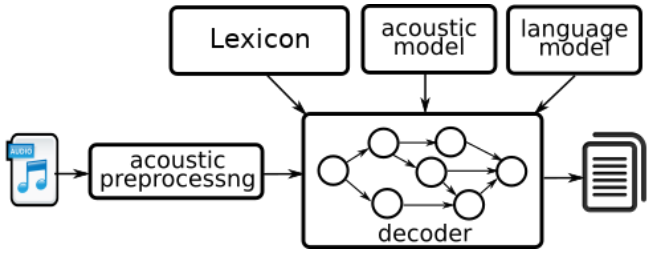


Image from (Jurafsky Martin, 2009, p. 290)

ASR System Architecture

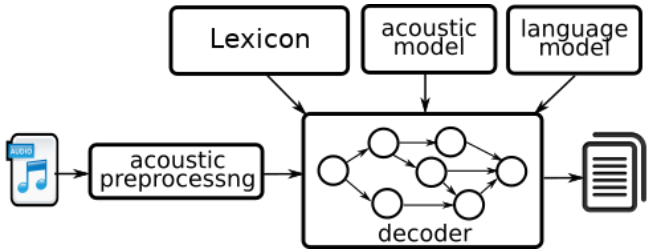


ASR System architecture - Models of linguistic knowledge



- Acoustic models: Given audio, which phonemes are likely?
- Lexicon: maps orthographic words (cow) to phoneme sequences (K AW) E.g. CMUdict [Link]
- Language model: Models the probabilities of different sequences of words
  - Helps to rank the different acoustic hypotheses

ASR System architecture - Other components



- Acoustic preprocessing: transform audio recording into sequence of feature vectors
- Lattice: a directed acyclic graph, used to represent the search space
- Decoder: the part of the program that considers the input and explores different pathways through the search space

Steps in ASR

Statistical ASR uses a supervised machine learning approach:

1. Building a data set
2. Feature extraction
3. Training
4. Testing
5. Evaluation

Building a data set

- Gather speech representative of audio in your application
- Annotate speech with transcriptions:
  - Word-level
  - Phone-level (expensive!)
- Ideally time-aligned

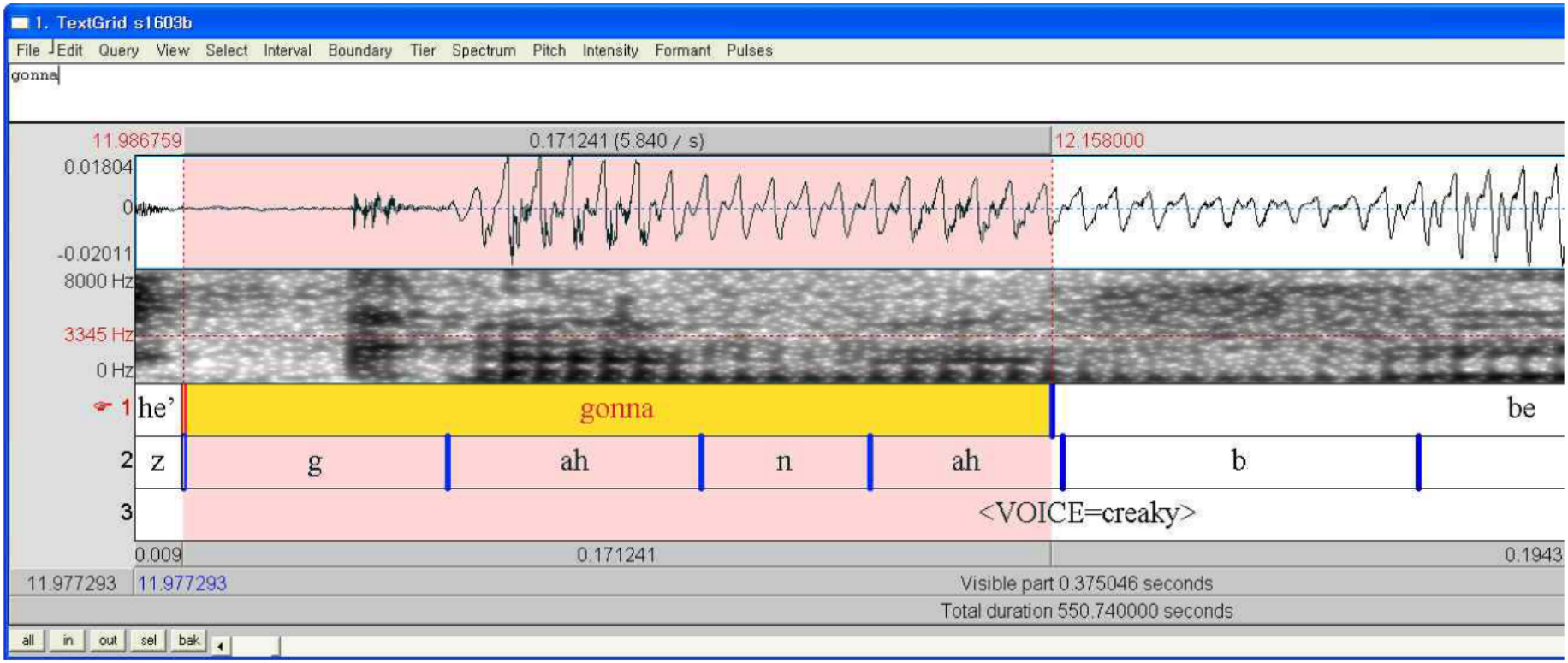


Image from https://koreascience.kr/article/JAKO201215734995939.pdf

Feature extraction

- Convert byte sequences representing audio into feature vectors that can be used for machine learning
- Most common representation are so-called MFCCs (mel frequency cepstral coefficients)
- Place window over small amount of speech and calculate statistics; Slide window forward a little and repeat.
- In the end, we get a sequence of feature vectors that describe the speech signal over time

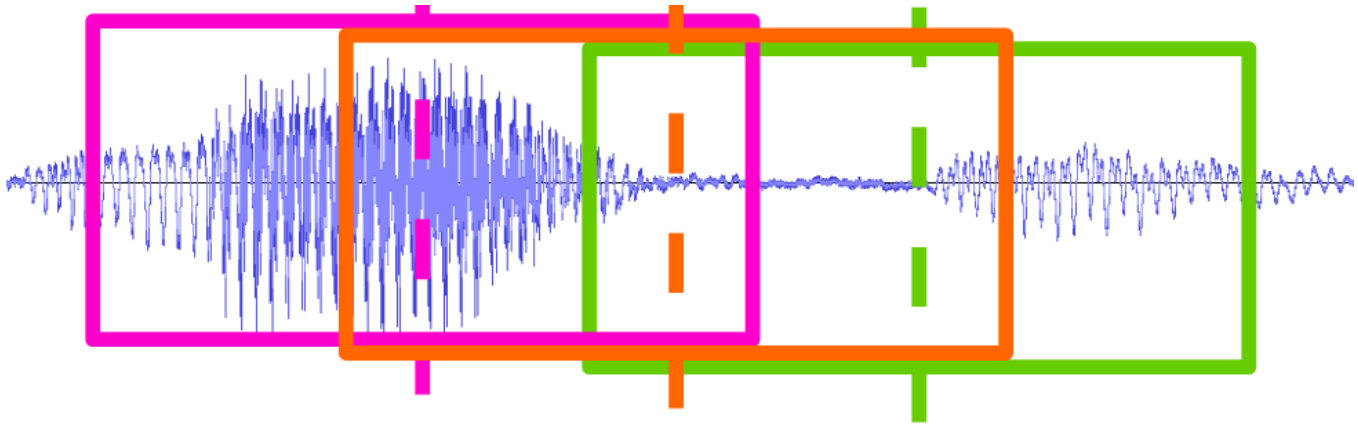


Image from Slides on ASR by Gwénolé Lecorvé

Feature extraction - Process

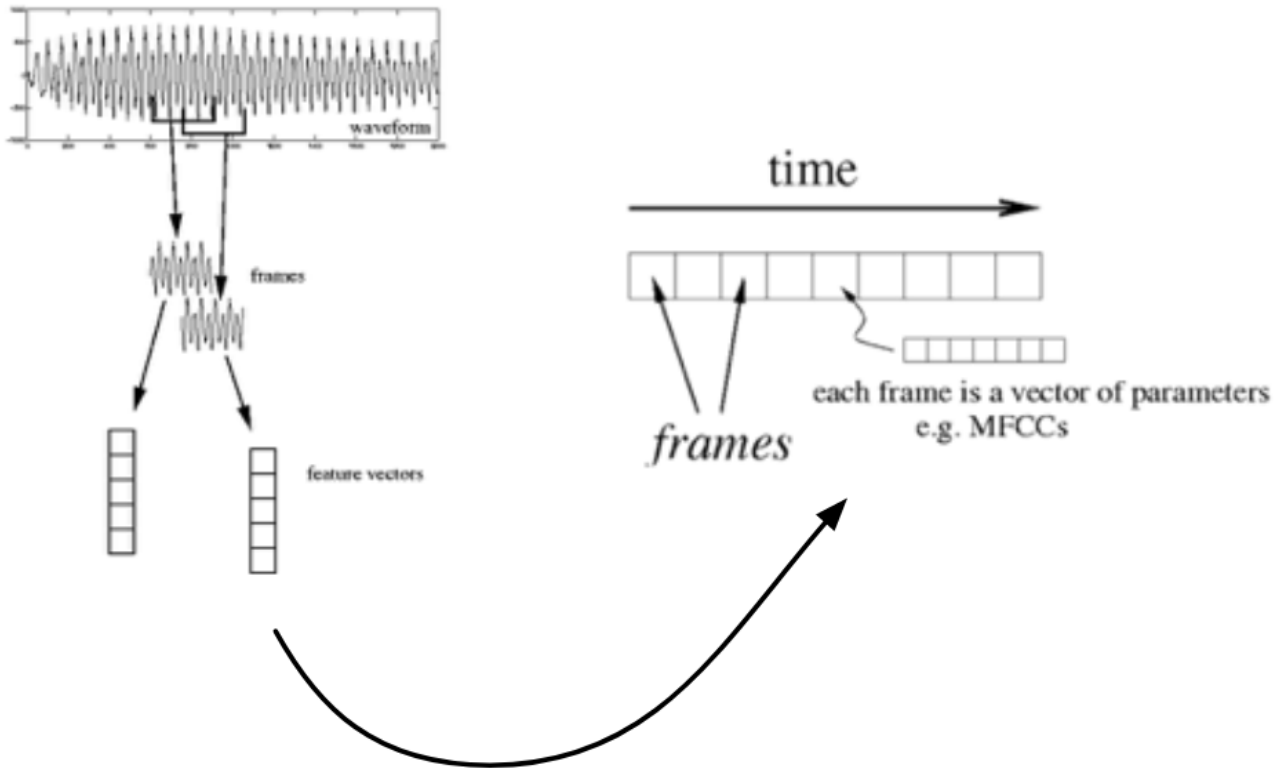


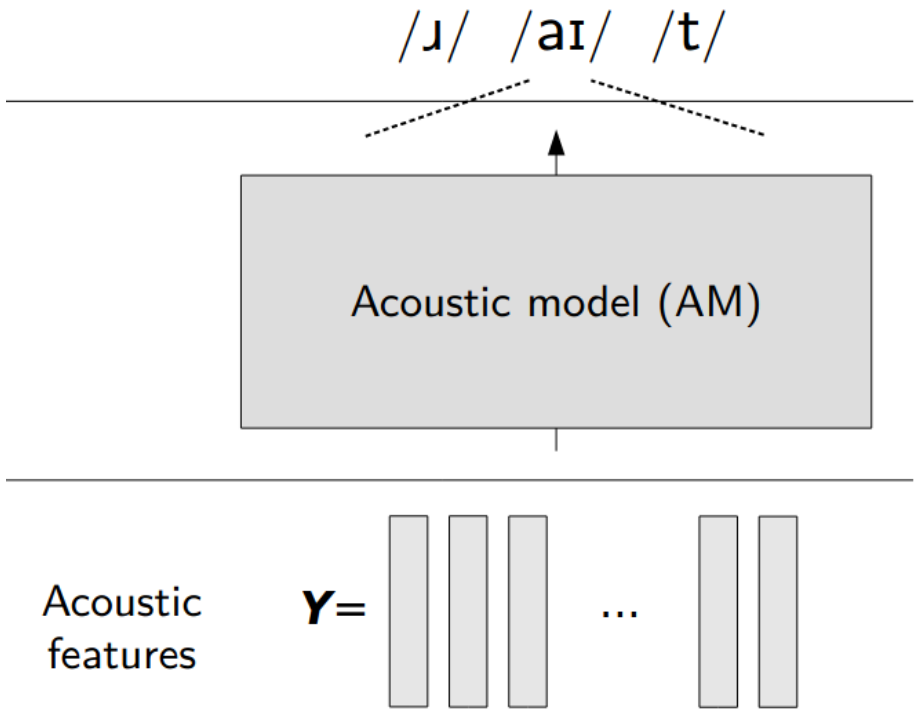
Image from Slides on ASR by Peter Bell

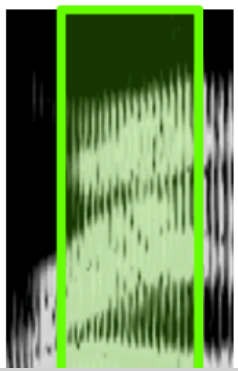
Training

- Stage where we build the statistical models the Decoder will use to explore and rank different paths through the lattice
- We iterate over entries in the corpus with goal of building two models: acoustic model and language model (n-gram)
  - acoustic model, and
  - language model (n-gram)

Acoustic Model

- Accoustic model is more complicated, involves computing probabilities on real-valued data (not discrete set of tokens)
- Variety of mechanisms possible: GMMs, neural networks, Support Vector Machines, Conditional Random Fields
- For our purposes, can treat acoustic model as black box that accepts a feature vector and outputs a score that reflects likelihood of different phonemes





Language Model

From our lecture on Writer's Aids

- *That factory's speciality is wooden boets; they are nice but they cost a lot.*
- What suggestions should our spell checker offer for *boets*?

Similar problem in speech

- E.g., "Recognize speech" vs. "Wreck a nice beach"
- Phonemes allow both possibilities - How to decide between these candidates ?
- We need a resource that we can ask questions like "Given this sentence, which of these N words fits best?"

Language Model and Candidate Ranking

- Problem can be solved to some degree with a statistical Language Model
  - Models the likelihood of word sequences
    - Likelihood("That factory's speciality is wooden **boots**") = relatively low
    - Likelihood("That factory's speciality is wooden **boats**") = relatively high
    - Likelihood("Wreck a nice beach") = relatively low (probably)
    - Likelihood("Recognize speech") = relatively high

Probability - An intuitive description

- Function *P(event)* produces a score intended to reflect how likely an event is
- Ranges between 0 (certain to not occur) and 1.0 (certain to occur)
- Consider coin toss:
  - Heads and Tails have equal chance
  - P(Heads) = 0.5; P(Tails) = 0.5
- Now consider biased coin toss:
  - How to check for biased coin?
  - Can flip coin a hundred times and record outcomes
  - Might yield heads 75 / 100 flips
  - P(Heads) = 0.75; P(Tails) = 0.25

Applying Probability in Language Models

That factory's speciality is wooden boets; they are nice but they cost a lot.

- Simple approach: compare P(boots) and P(boats), and choose most likely word
- Estimate probabilities by observing each word in set of documents
- Keep track of
  - Number of words you look at
  - Number of times you see 'boots' and 'boats'
- P(boats) = number of occurences of boats / number of words
- Choose word with highest probability

Including More Context

That factory's speciality is wooden boets; they are nice but they cost a lot.

- Problem arises if your data contains more occurences of 'boots'
- As humans, we see 'wooden' and intuitively know 'boats' is more likely
- Can calculate P(wooden boats) and P(wooden boots) in same way
- Keep track of
  - Number of word pairs we look at
  - Number of times we see each unique pair of words (e.g., 'wooden boat', but also 'walk to')
- P(wooden boats) = number of occurences of string 'wooden boats' / number of word pairs
- Compare probabilities of pairs of words; choose pair with highest probability

Generalising from Pairs to N-grams

N-grams:

- sequences of observations
- context window size n: bigrams, trigrams, ...
- common applications in ASR, spell checking

"That store sells electric rubber bo...(boots? boats?)"

word n-grams:

- word unigrams: *That*, *store*, *sells*, ...
- word bigrams: *That store*, *store sells*, *sells electric*, *electric rubber*...
- word trigrams: *sells electric rubber*, *electric rubber boots*, ...

N-grams - Much More

- Covered high-level introduction
- Have not covered mathematical definition
- Nice video on YouTube introducing N-grams and some of the mathematics behind them: https://www.youtube.com/watch?v=UKAQF8wHxkE

Testing ASR Performance

- Test the performance of the system
- Relies on Decoder component:

  "It's the job of a decoder to simultaneously segment the utterance into words and identify each of these words." (Jurafsky & Martin, 2009, p. 318)
- System builds a search lattice, all the different possibilities to consider, derived from combining the lexicon and acoustic models
- Lattice is huge DAG with 10s of millions of states! [Link]
- Decoder searches through the lattice, to find best matching pathways.
- Multiple searches in parallel to maintain a certain top N number of plausible transcriptions

Decoding Approach

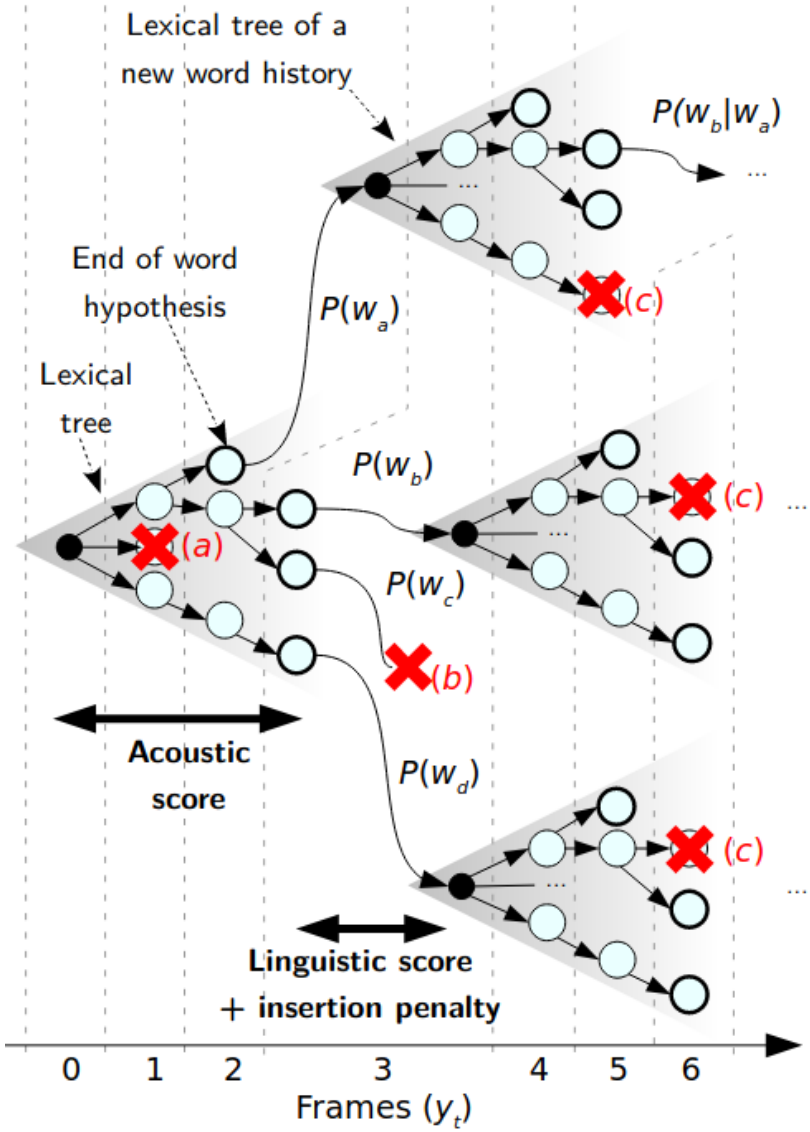


Image from Slides on ASR by Gwénolé Lecorvé

Decoding Approach - Steps

Called a Beam search

- Start at feature vector 0
- Consult with all acoustic models; retain acoustic scores
- Continue to process feature vectors until end of word event is detected
- For hypotheses that arrive at an end of word, apply language model score to running accoustic model score
- Remove unlikely hypotheses (outside the beam)
- Move to next frame
- Stop when there are no more frames, and return N-best hypotheses

Very high-level description. In reality, many more details.

Evaluation

- Standard measure is the Word Error Rate (WER)
- Can think of it as percentage of words system got wrong.
- Builds on Minimum Edit Distance (session 3) (in words not characters)
- Compare reference string and string output by ASR (hypothesis)
- Count word insertions, word deletions, word substitutions necessary to transform hypothesis into reference string
- WER is then 100 * (I + D + S) / num_words in corpus
- Accuracy is 100 - WER

Error analysis

Troubleshoot system

- Can triage errors by looking at which occur most often
- Can use confusion matrices: statistics about which phonemes or words were commonly confused
- Other useful metrics:
  - commonly inserted or deleted words
  - error rates by speaker, gender, L2 etc.

Error analysis - Confusion matrix

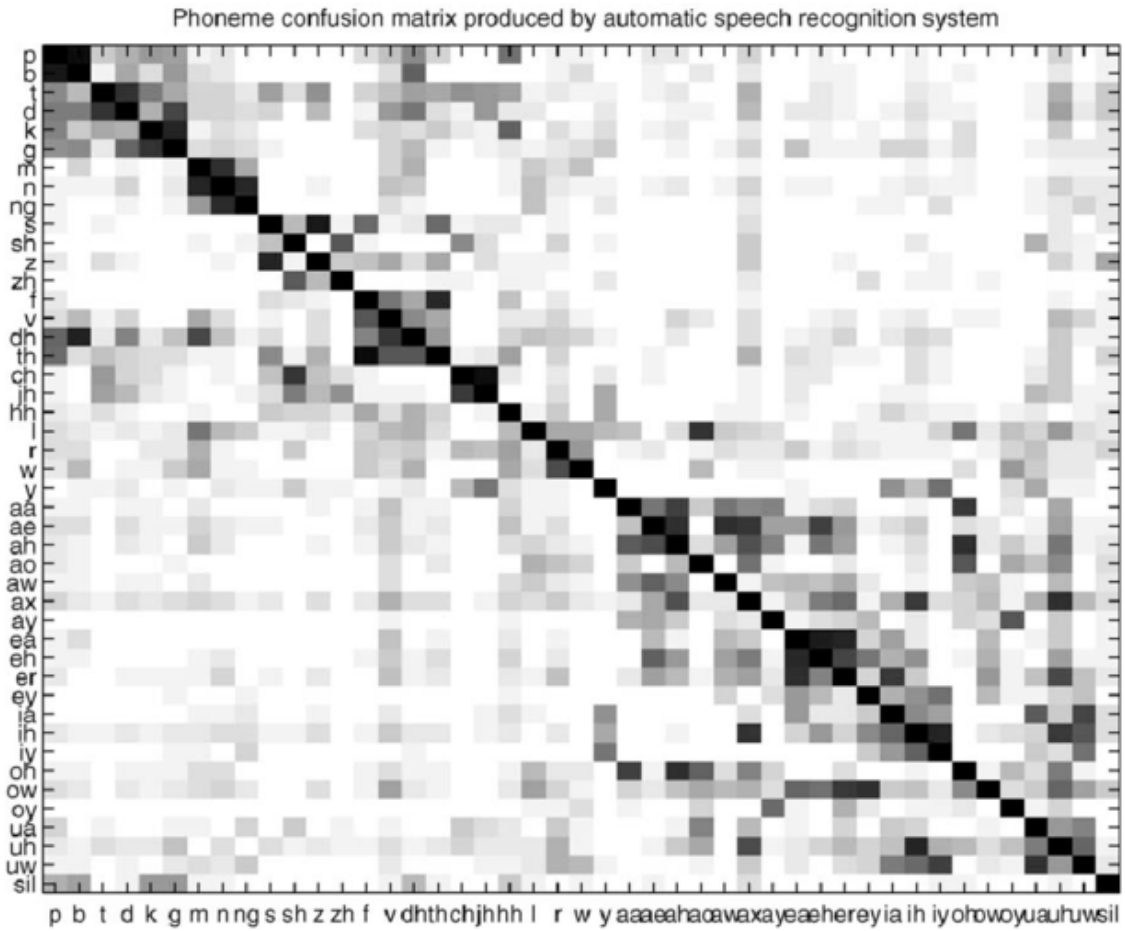


Image from Cox & Vinagre 2009 (https://www.sciencedirect.com/science/article/abs/pii/S0167639303001213?via%3Dihub)

Wrapping Up

Today's topics:

- A simplified linguistic description of speech
- Representing speech with computers
- Different speech processing applications
- An overview of Automatic Speech Recognition

Next time: Review session and Q & A.

References and Acknowledgments

Content for today's slides adapted from

- [Jurafsky and Martin, 2009]
- [Lecorvé, 2023]
- [Bell, 2023]

Reference List I

Peter Bell. Automatic speech recognition: Introduction, 2023. URL https://www.inf.ed.ac.uk/teaching/courses/asr/2022-23/asr01-intro.pdf.

Andrzej Drygajlo. *Voice, Forensic Evidence of*, pages 1388–1396. Springer US, Boston, MA, 2009. ISBN 978-0-387-73003-5. doi: . URL https://doi.org/10.1007/978-0-387-73003-5_104.

Dan Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009. ISBN 9780131873216 0131873210. URL http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.

Ludovic Le Bigot, Philippe Bretier, and Patrice Terrier. *Detecting and Exploiting User Familiarity in Natural Language Human-computer Dialogue*. 10 2008. ISBN 978-953-7619-14-5. doi: .

Reference List II

Gwénolé Lecorvé. Automatic speech recognition, 2023. URL http://people.irisa.fr/Gwenole.Lecorve/lectures/ASR.pdf.

Mary Grantham O'Brien, Tracey M. Derwing, Catia Cucchiarini, Debra M. Hardison, Hansjörg Mixdorff, Ron I. Thomson, Helmer Strik, John M. Levis, Murray J. Munro, Jennifer A. Foote, and Greta Muller Levis. Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, 4(2):182–207, 2018. ISSN 2215-1931. doi: . URL https://www.jbe-platform.com/content/journals/10.1075/jslp.17001.obr.

Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006. ISSN 0167-6393. doi: . URL https://www.sciencedirect.com/science/article/pii/S0167639306000422.