Automatic Speech Recognition

Slides now available at

www.informatics.manchester.ac.uk/~harold/LELA300431/

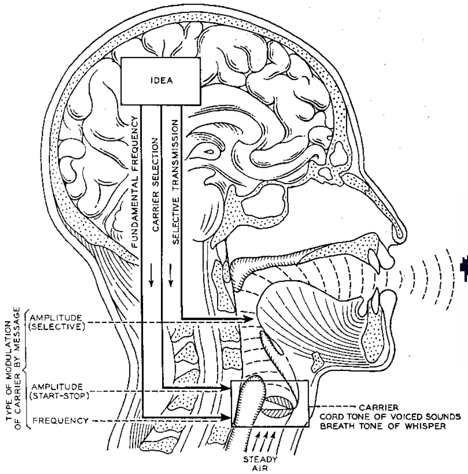
Automatic speech recognition

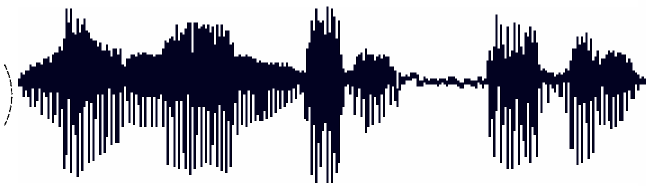
- What is the task?
- What are the main difficulties?
- How is it approached?
- How good is it?
- How much better could it be?

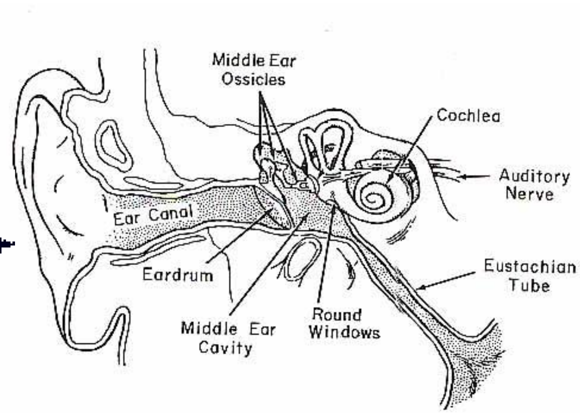
What is the task?

- Getting a computer to understand spoken language
- By “understand” we might mean
  - React appropriately
  - Convert the input speech into another medium, e.g. text
- Several variables impinge on this (see later)

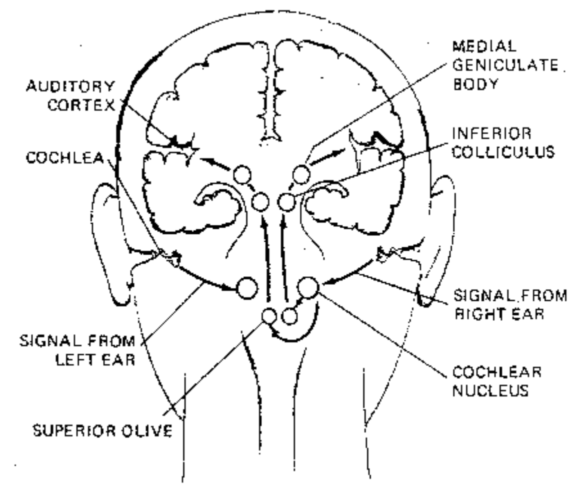
How do humans do it?



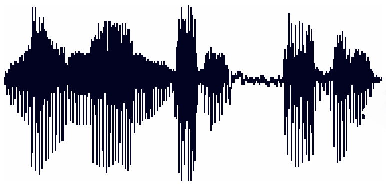




- Articulation produces
- sound waves which
- the ear conveys to the brain
- for processing

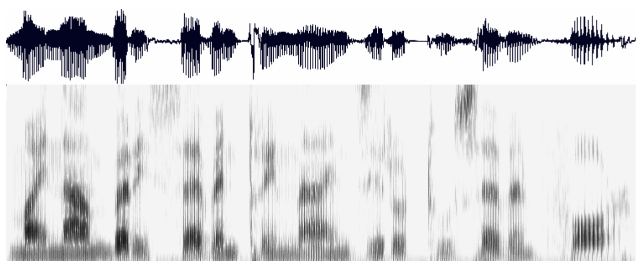


How might computers do it?

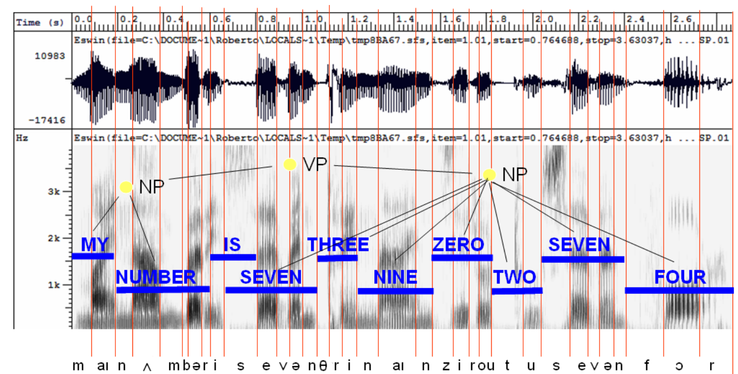


Acoustic waveform

Acoustic signal



- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation



Speech recognition

What's hard about that?

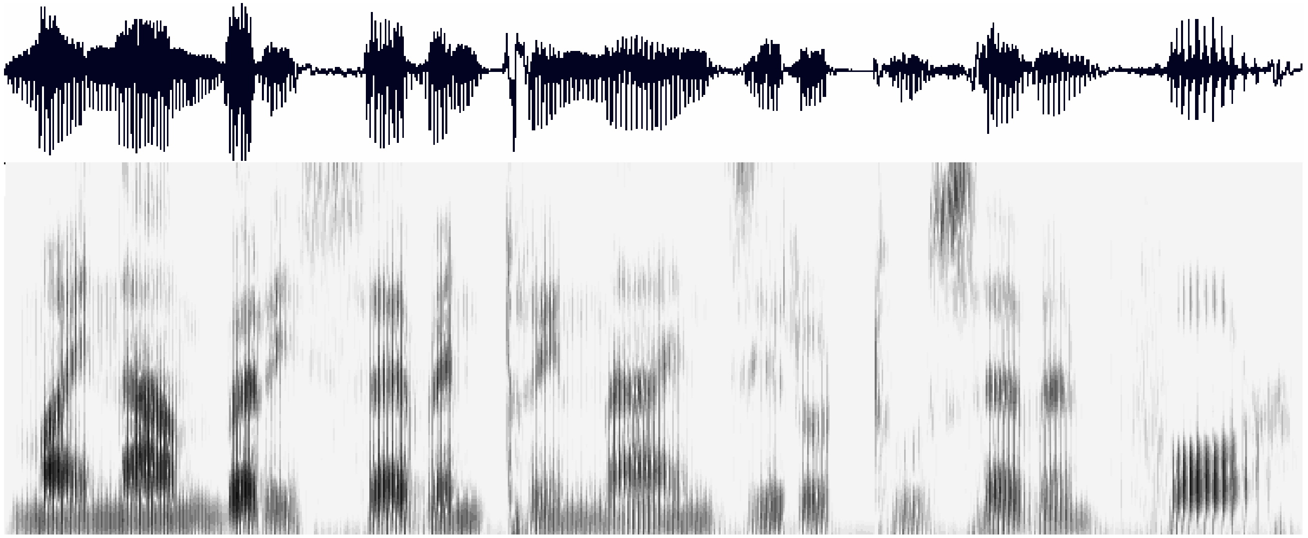
- Digitization
  - Converting analogue signal into digital representation
- Signal processing
  - Separating speech from background noise
- Phonetics
  - Variability in human speech
- Phonology
  - Recognizing individual sound distinctions (similar phonemes)
- Lexicology and syntax
  - Disambiguating homophones
  - Features of continuous speech
- Syntax and pragmatics
  - Interpreting prosodic features
- Pragmatics
  - Filtering of performance errors (disfluencies)

Digitization

- Analogue to digital conversion
- Sampling and quantizing
- Use filters to measure energy levels for various points on the frequency spectrum
- Knowing the relative importance of different frequency bands (for speech) makes this process more efficient
- E.g. high frequency sounds are less informative, so can be sampled using a broader bandwidth (log scale)

Separating speech from background noise

- Noise cancelling microphones
  - Two mics, one facing speaker, the other facing away
  - Ambient noise is roughly same for both mics
- Knowing which bits of the signal relate to speech
  - Spectrograph analysis



Variability in individuals' speech

- Variation among speakers due to
  - Vocal range (f0, and pitch <u>range</u> – see later)
  - Voice quality (growl, whisper, physiological elements such as nasality, adenoidality, etc)
  - ACCENT !!! (especially vowel systems, but also consonants, allophones, etc.)
- Variation within speakers due to
  - Health, emotional state
  - Ambient conditions
- Speech style: formal read vs spontaneous

Speaker-(in)dependent systems

- Speaker-dependent systems
  - Require “training” to “teach” the system your individual idiosyncracies
    - The more the merrier, but typically nowadays 5 or 10 minutes is enough
    - User asked to pronounce some key words which allow computer to infer details of the user’s accent and voice
    - Fortunately, languages are generally systematic
  - More robust
  - But less convenient
  - And obviously less portable
- Speaker-independent systems
  - Language coverage is reduced to compensate need to be flexible in phoneme identification
  - Clever compromise is to learn on the fly

Identifying phonemes

- Differences between some phonemes are sometimes very small
  - May be reflected in speech signal (eg vowels have more or less distinctive f1 and f2)
  - Often show up in coarticulation effects (transition to next sound)
    - e.g. aspiration of voiceless stops in English
  - Allophonic variation

Disambiguating homophones

- Mostly differences are recognised by humans by context and need to make sense

It’s hard to wreck a nice beach

What dime’s a neck’s drain to stop port?

- Systems can only recognize words that are in their lexicon, so limiting the lexicon is an obvious ploy
- Some ASR systems include a grammar which can help disambiguation

(Dis)continuous speech

- Discontinuous speech much easier to recognize
  - Single words tend to be pronounced more clearly
- Continuous speech involves contextual coarticulation effects
  - Weak forms
  - Assimilation
  - Contractions

Interpreting prosodic features

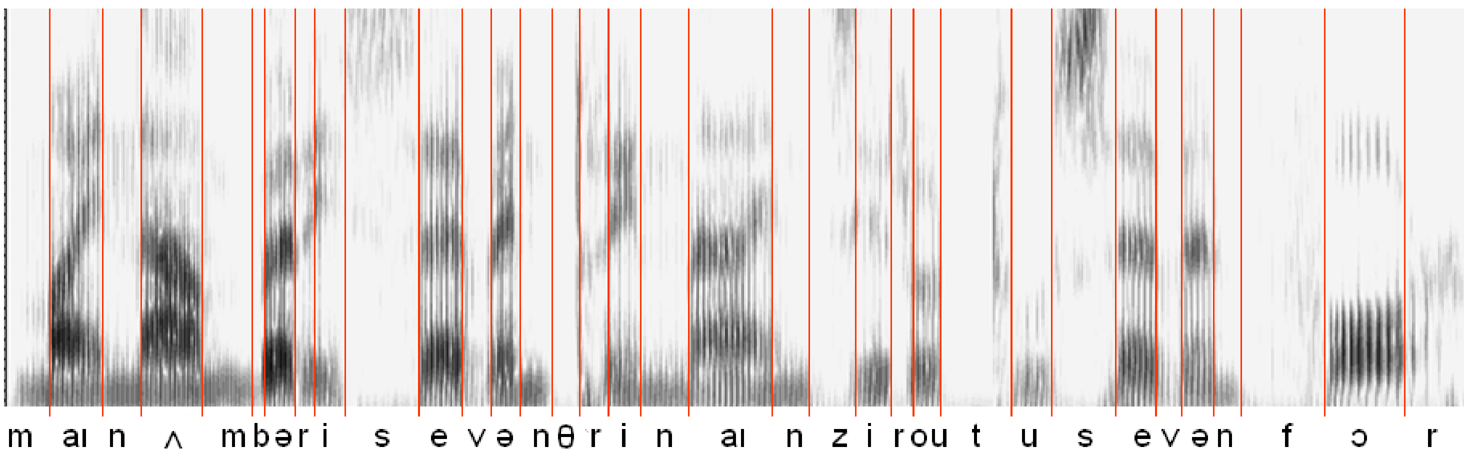
- Pitch, length and loudness are used to indicate “stress”
- All of these are relative
  - On a speaker-by-speaker basis
  - And in relation to context
- Pitch and length are phonemic in some languages

Pitch

- Pitch contour <u>can</u> be extracted from speech signal
  - But pitch differences are relative
  - One man's high is another (wo)man's low
  - Pitch range is variable
- Pitch contributes to intonation
  - But has other functions in tone languages
- Intonation <u>can</u> convey meaning

Length

- Length is easy to measure but difficult to interpret
- Again, length is relative
- It is phonemic in many languages
- Speech rate is not constant – slows down at the end of a sentence



Loudness

- Loudness is easy to measure but difficult to interpret
- Again, loudness is relative

Performance errors

- Performance “errors” include
  - Non-speech sounds
  - Hesitations
  - False starts, repetitions
- Filtering implies handling at syntactic level or above
- Some disfluencies are deliberate and have pragmatic effect – this is not something we can handle in the near future

Approaches to ASR

- Template matching
- Knowledge-based (or rule-based) approach
- Statistical approach:
  - Noisy channel model + machine learning

Template-based approach

- Store examples of units (words, phonemes), then find the example that most closely fits the input
- Extract features from speech signal, then it’s “just” a complex similarity matching problem, using solutions developed for all sorts of applications
- OK for discrete utterances, and a single user

Template-based approach

- Hard to distinguish very similar templates
- And quickly degrades when input differs from templates
- Therefore needs techniques to mitigate this degradation:
  - More subtle matching techniques
  - Multiple templates which are aggregated
- Taken together, these suggested …

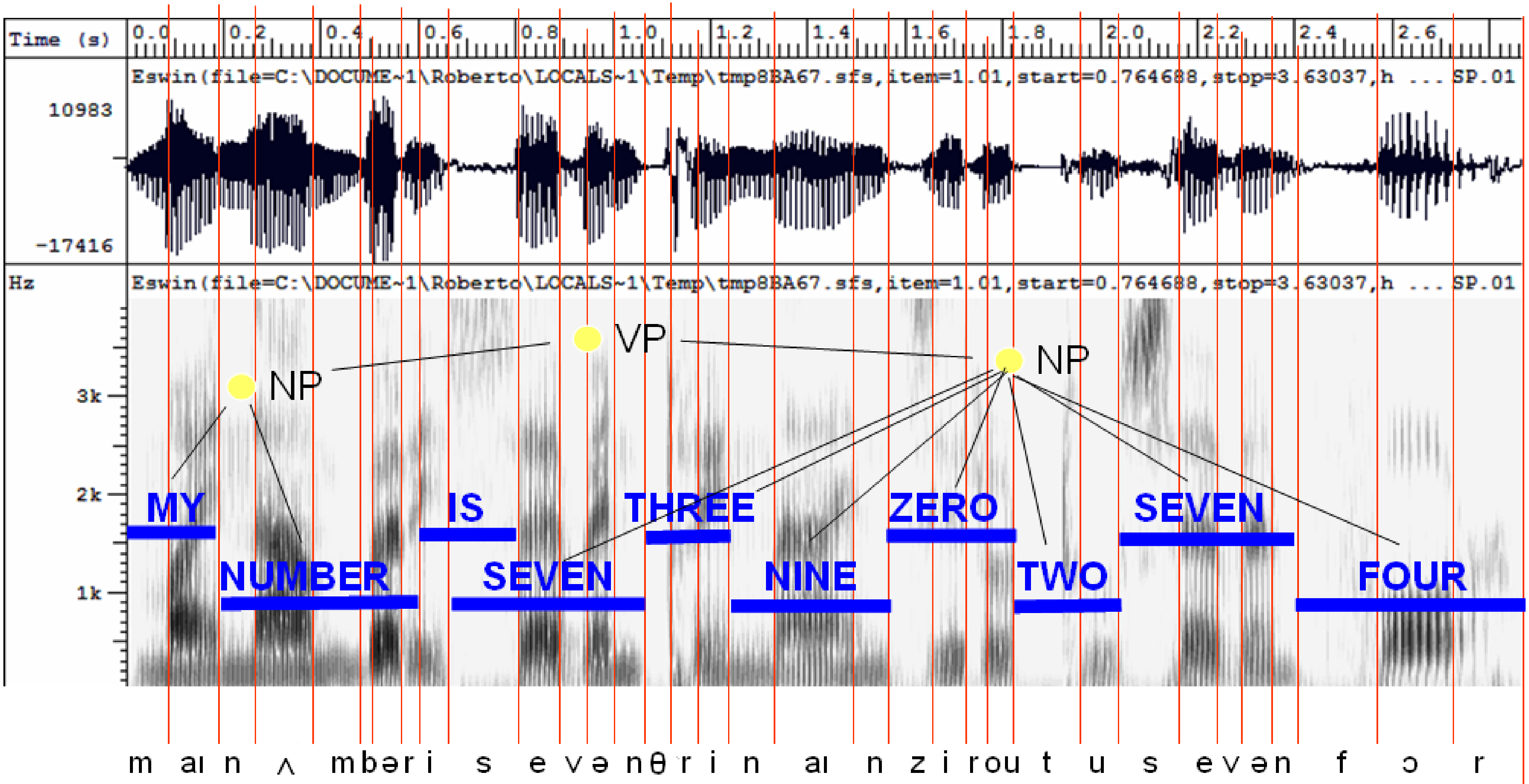
Rule-based approach

- Use knowledge of phonetics and linguistics to guide search process
- Templates are replaced by rules expressing everything (anything) that might help to decode:
  - Phonetics, phonology, phonotactics
  - Syntax
  - Pragmatics

Rule-based approach

- Typical approach is based on “blackboard” architecture:
  - At each decision point, lay out the possibilities
  - Apply rules to determine which sequences are permitted
- Poor performance due to
  - Difficulty to express rules
  - Difficulty to make rules interact
  - Difficulty to know how to improve the system

s	k	iː	ʃ
ʃ	h	iə	tʃ
	p	ɪ	h
	t		s



- Identify individual phonemes
- Identify words
- Identify sentence structure and/or meaning
- Interpret prosodic features (pitch, loudness, length)

Statistics-based approach

- Can be seen as extension of template-based approach, using more powerful mathematical and statistical tools
- Sometimes seen as “anti-linguistic” approach
  - Fred Jelinek (IBM, 1988): “Every time I fire a linguist my system improves”

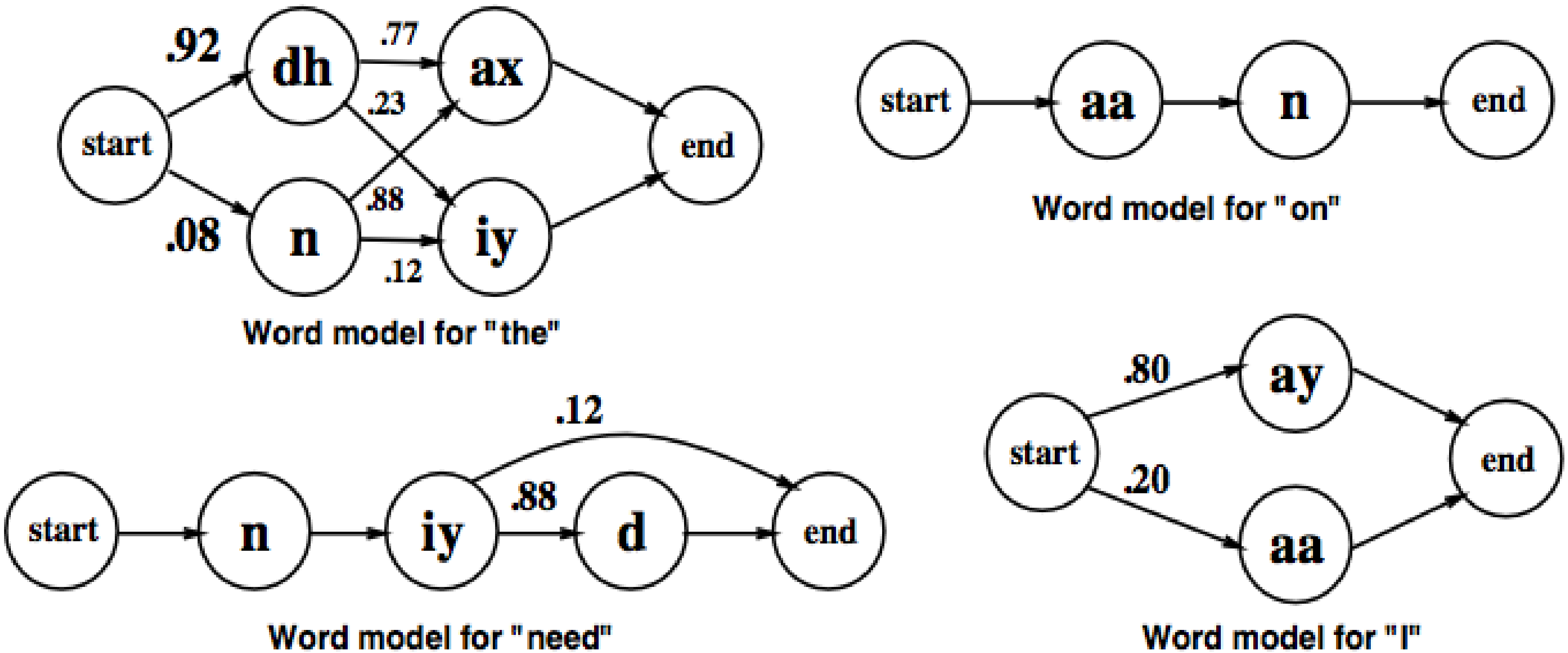
Statistics-based approach

- Collect a large corpus of transcribed speech recordings
- Train the computer to learn the correspondences (“machine learning”)
- At run time, apply statistical processes to search through the space of all possible solutions, and pick the statistically most likely one

Machine learning

- Acoustic and Lexical Models
  - Analyse training data in terms of relevant features
  - Learn from large amount of data different possibilities
    - different phone sequences for a given word
    - different combinations of elements of the speech signal for a given phone/phoneme
  - Combine these into a Hidden Markov Model expressing the probabilities

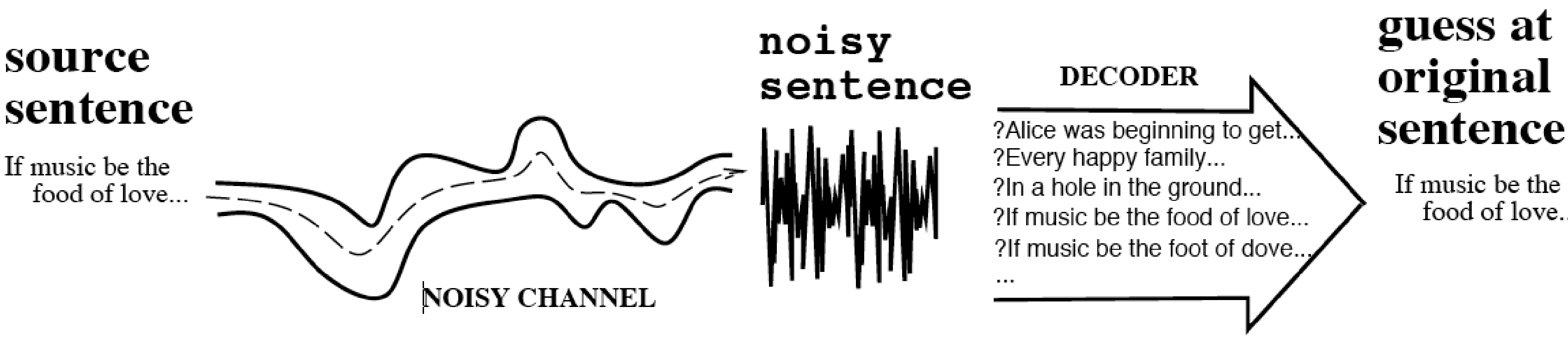
HMMs for some words



Language model

- Models likelihood of word given previous word(s)
- n-gram models:
  - Build the model by calculating bigram or trigram probabilities from text training corpus
  - Smoothing issues

The Noisy Channel Model



- Search through space of all possible sentences
- Pick the one that is most probable given the waveform

The Noisy Channel Model

- Use the acoustic model to give a set of likely phone sequences
- Use the lexical and language models to judge which of these are likely to result in probable word sequences
- The trick is having sophisticated algorithms to juggle the statistics
- A bit like the rule-based approach except that it is all learned automatically from data

Evaluation

- Funders have been very keen on competitive quantitative evaluation
- Subjective evaluations are informative, but not cost-effective
- For transcription tasks, word-error rate is popular (though can be misleading: all words are not equally important)
- For task-based dialogues, other measures of understanding are needed

Comparing ASR systems

- Factors include
  - Speaking mode: isolated words vs continuous speech
  - Speaking style: read vs spontaneous
  - “Enrollment”: speaker (in)dependent
  - Vocabulary size (small <20 … large > 20,000)
  - Equipment: good quality noise-cancelling mic … telephone
  - Size of training set (if appropriate) or rule set
  - Recognition method

Remaining problems

- **Robustness** – graceful degradation, not catastrophic failure
- **Portability** – independence of computing platform
- **Adaptability** – to changing conditions (different mic, background noise, new speaker, new task domain, new language even)
- **Language Modelling** – is there a role for linguistics in improving the language models?
- **Confidence Measures** – better methods to evaluate the absolute correctness of hypotheses.
- **Out-of-Vocabulary (OOV) Words** – Systems must have some method of detecting OOV words, and dealing with them in a sensible way.
- **Spontaneous Speech** – disfluencies (filled pauses, false starts, hesitations, ungrammatical constructions etc) remain a problem.
- **Prosody** –Stress, intonation, and rhythm convey important information for word recognition and the user's intentions (e.g., sarcasm, anger)
- **Accent, dialect and mixed language** – non-native speech is a huge problem, especially where code-switching is commonplace