Natural Language for Communication (con't.) -- Speech Recognition

Chapter 23.5

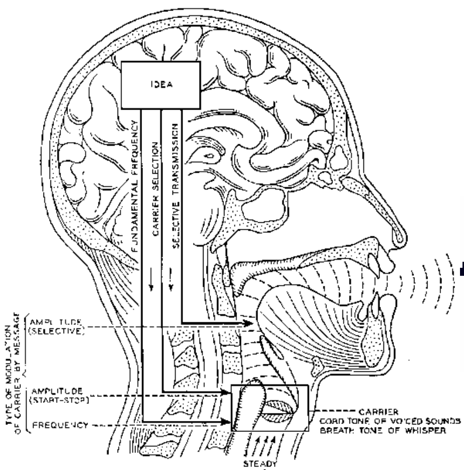
Automatic speech recognition

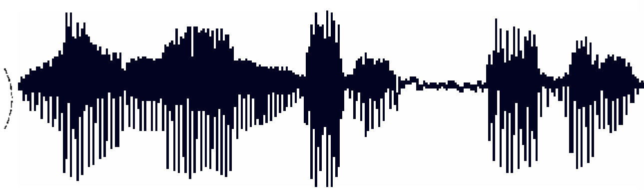
- What is the task?
- What are the main difficulties?
- How is it approached?
- How good is it?
- How much better could it be?

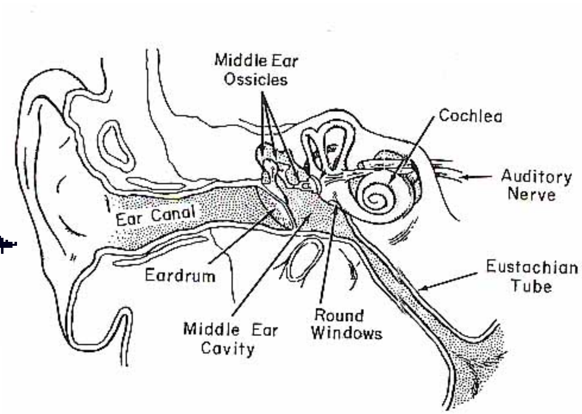
What is the task?

- 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)

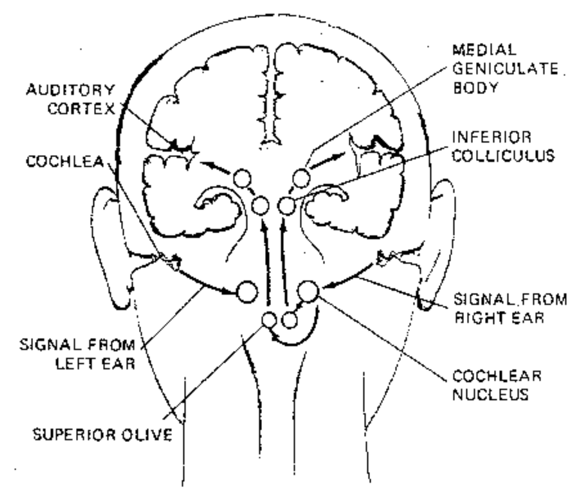
How do humans do it?







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



Human Hearing

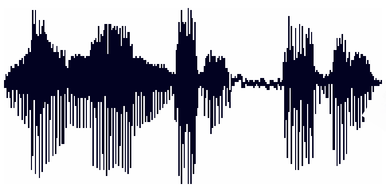
- The human ear can detect frequencies from 20Hz to 20,000Hz but it is most sensitive in the critical frequency range, 1000Hz to 6000Hz, (Ghitza, 1994).

- Recent Research has uncovered the fact that humans do not process individual frequencies.

- Instead, we hear groups of frequencies, such as format patterns, as cohesive units and we are capable of distinguishing them from surrounding sound patterns, (Carrell and Opie, 1992) .

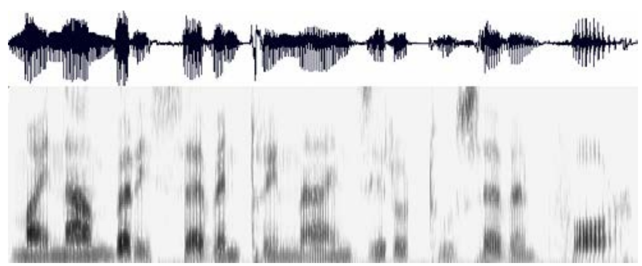
- This capability, called *auditory object formation*, or *auditory image formation*, helps explain how humans can discern the speech of individual people at cocktail parties and separate a voice from noise over a poor telephone channel, (Markowitz, 1995).

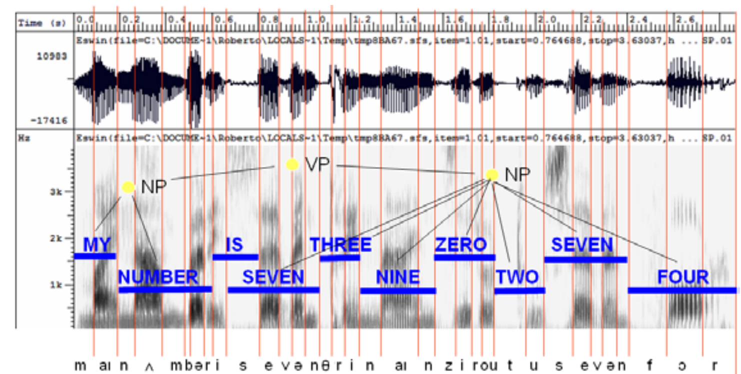
How might computers do it?



Acoustic waveform

Acoustic signal





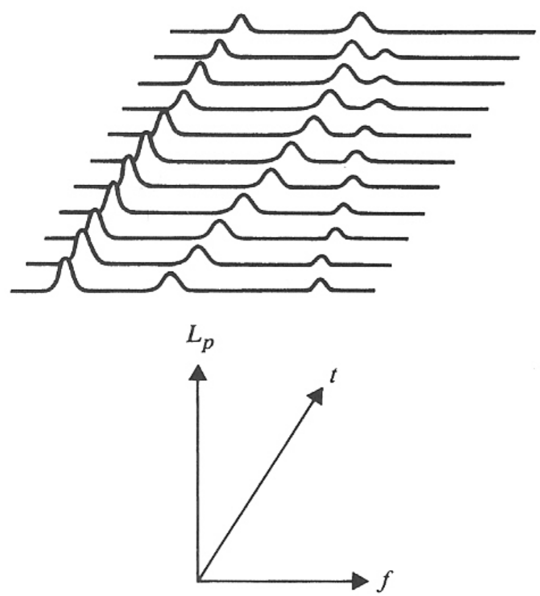
Speech recognition

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

What's hard about that?

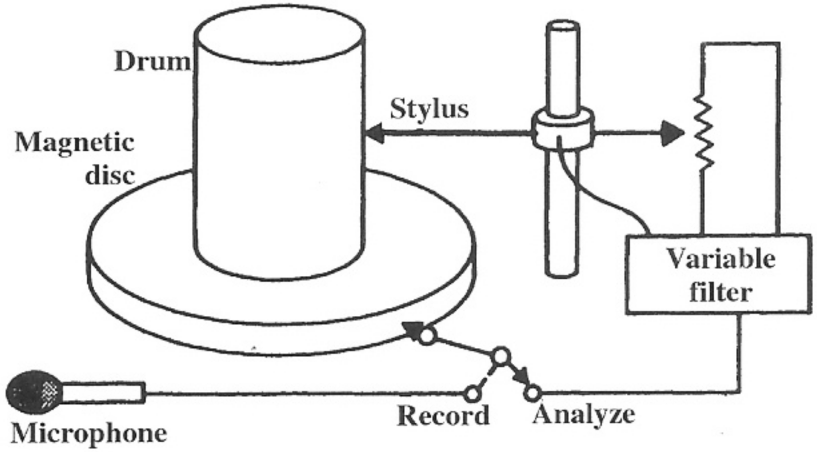
- Digitization
  - Converting analogue signal into digital representation
- Signal processing
  - Separating speech from background noise
- Phonetics
  - Variability in human speech
- Phonology
  - Recognizing individual sound distinctions (similar phonemes)
- Lexicology and syntax
  - Disambiguating homophones
  - Features of continuous speech
- Syntax and pragmatics
  - Interpreting prosodic features (e.g., pitch, stress, volume, tempo)
- Pragmatics
  - Filtering of performance errors (disfluencies, e.g., um, erm, well, huh)

Analysis of Speech

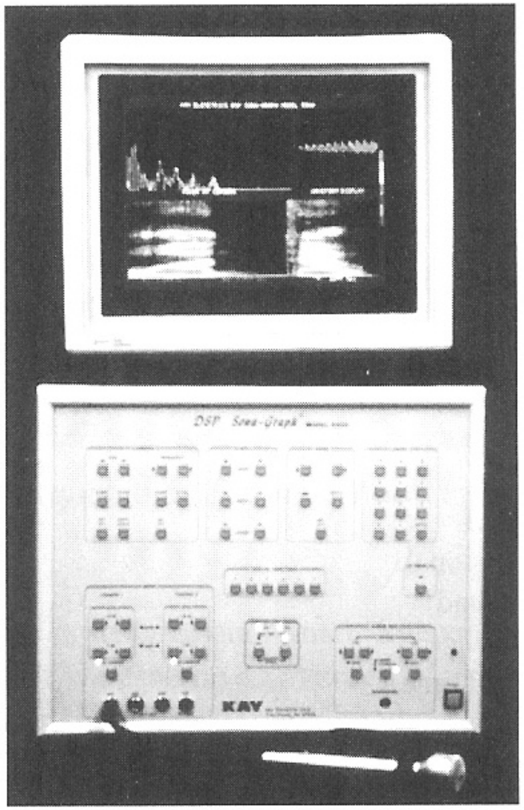


3D Display of sound level vs. frequency and time

Speech Spectograph



(a)

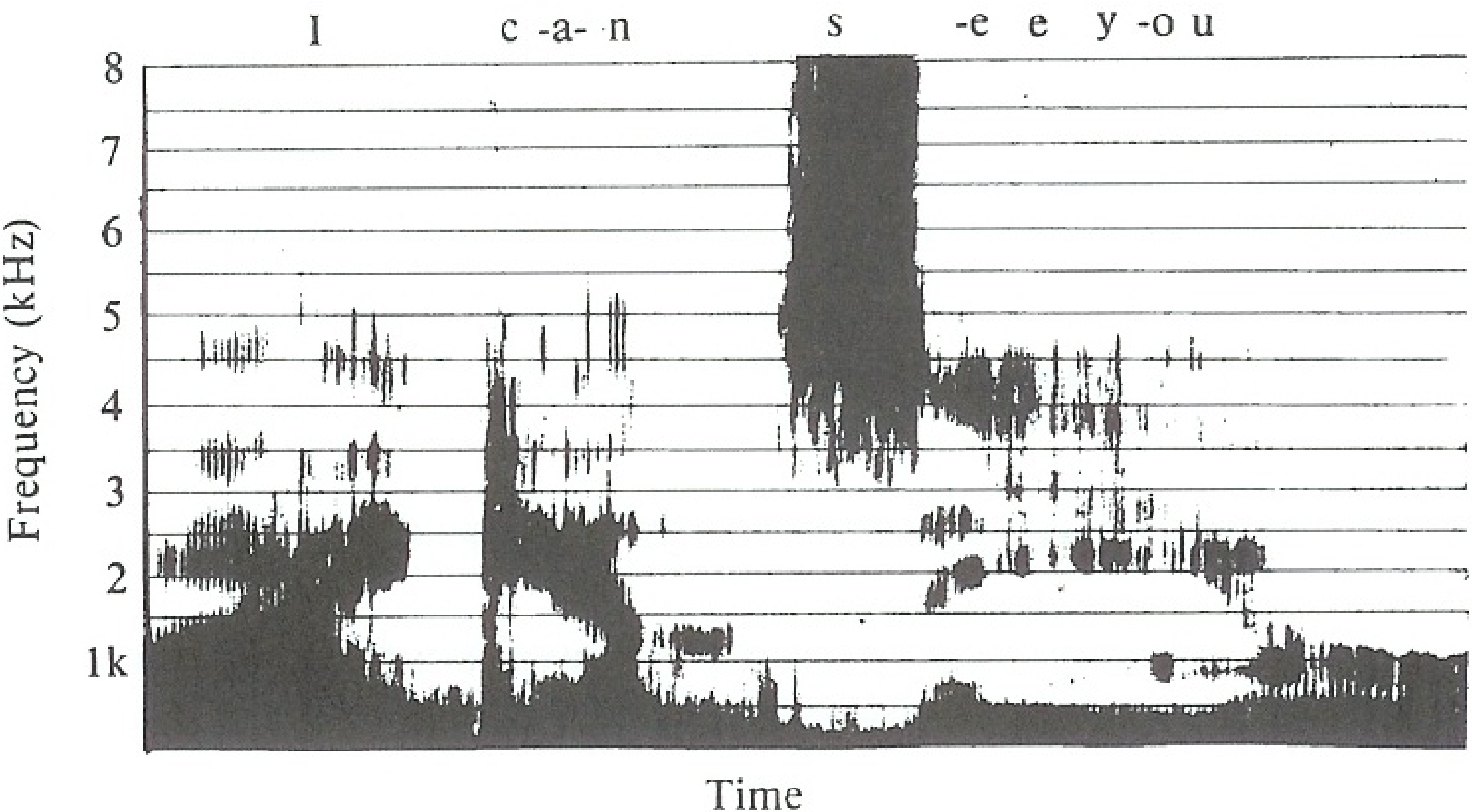


(b)

AS DEVELOPED AT BELL LABORATORIES (1945)

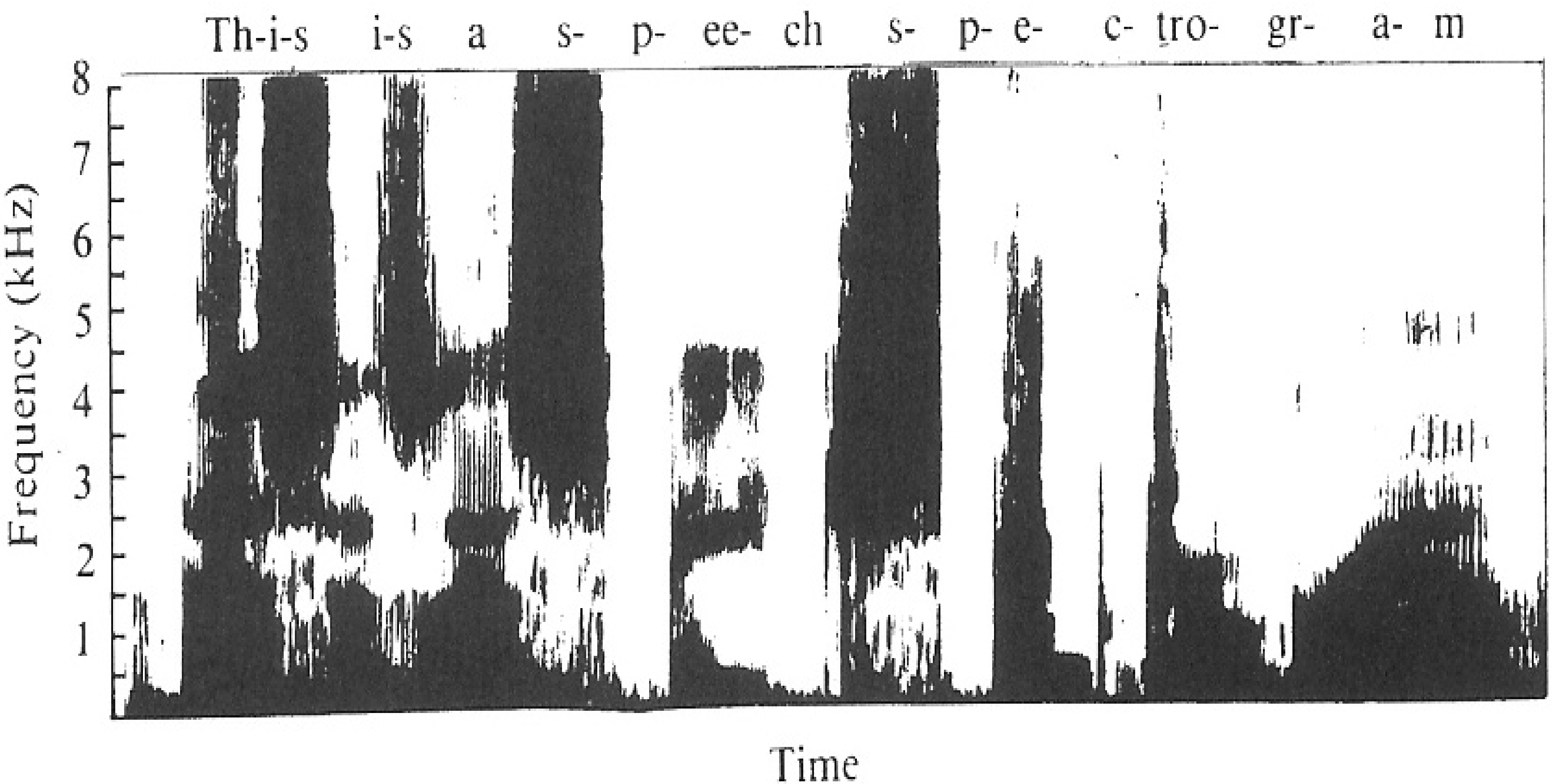
DIGITAL VERSION

Speech Spectogram



SPEECH SPECTROGRAM OF A SENTENCE:

This is a speech spectrogram

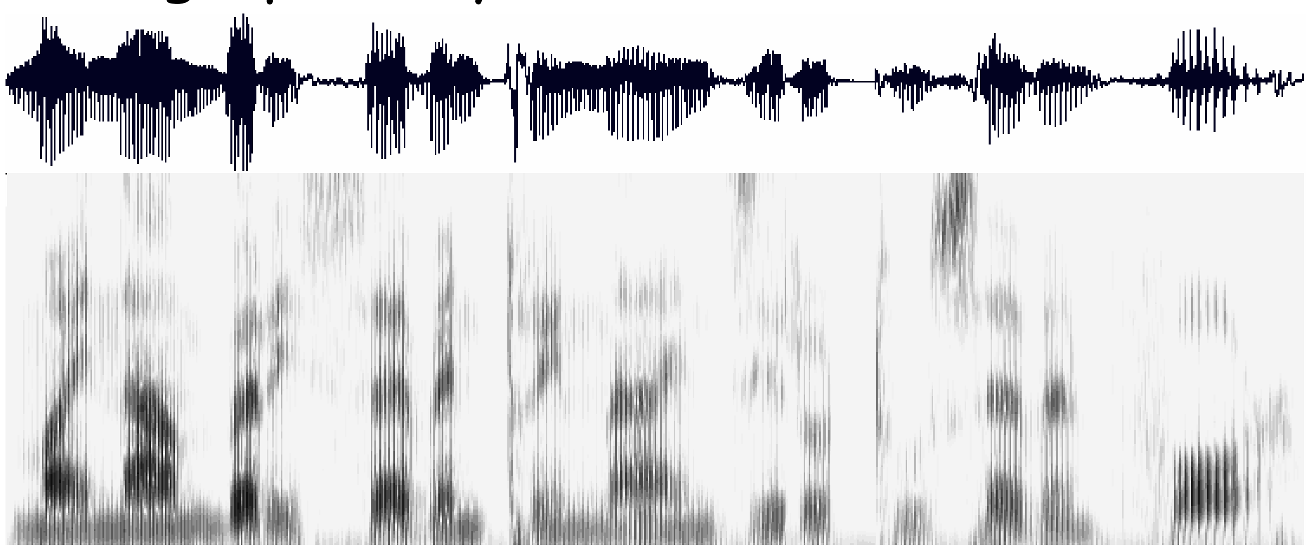


Digitization

- 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
  - 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 (e.g., 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 (allophone is one of a set of sounds used to pronounce a single phoneme)

International Phonetic Alphabet: Purpose and Brief History

- Purpose of the alphabet: to provide a universal notation for the sounds of the world's languages
  - "Universal" = If any language on Earth distinguishes two phonemes, IPA must also distinguish them
  - "Distinguish" = Meaning of a word changes when the phoneme changes, e.g. "cat" vs. "bat."
- Very Brief History:
  - 1876: Alexander Bell publishes a distinctive-feature-based phonetic notation in "Visible Speech: The Science of the Universal Alphabetic." His notation is rejected as being too expensive to print
  - 1886: International Phonetic Association founded in Paris by phoneticians from across Europe
  - 1991: Unicode provides a standard method for including IPA notation in computer documents

ARPAbet Vowels
(for American English)

	b_d	ARPA		b_d	ARPA
1	bead	iy	9	bode	ow
2	bid	ih	10	booed	uw
3	bayed	ey	11	bud	ah
4	bed	eh	12	bird	er
5	bad	ae	13	bide	ay
6	bod(y)	aa	14	bowed	aw
7	bawd	ao	15	Boyd	oy
8	Budd(hist)	uh			

There is a complete ARPAbet phonetic alphabet, for all phones used in American English.

ARPABET List

(Phonetic Labels from TIMIT Speech Corpus)

No.	ARPABET	Examples	No.	ARPABET	Examples	No.	ARPABET	Examples
1	iy	b***ea***t	22	r	***r***ed	43	zh	mea***s***ure
2	ih	b***i***t	23	y	***y***et	44	sh	***sh***oe
3	eh	b***e***t	24	w	***w***et	45	v	***v***ery
4	ae	b***a***t	25	m	***m***om	46	f	***f***ief
5	ix	ros***e***s	26	em	butt***om***	47	dh	***th***ey
6	ax	th***e***	27	n	***n***on	48	th	***th***ief
7	ah	b***u***tt	28	nx	(flapped) n	49	hh	***h***ay
8	uw	b***oo***t	29	en	butto***n***	50	hv	Le***h***eigh
9	uh	b***oo***k	30	ng	si***ng***	51	dcl	(d closure)
10	ao	ab***ou***t	31	eng	Washi***ng***ton	52	bcl	(b closure)
11	aa	c***o***t	32	ch	***ch***urch	53	gcl	(g closure)
12	er	b***ir***d	33	jh	***j***udge	54	tcl	(t closure)
13	axr	din***er***	34	b	***b***ob	55	pcl	(p closure)
14	ey	b***ai***t	35	p	***p***op	56	kcl	(k closure)
15	ay	b***i***te	36	d	***d***ad	57	q	(glottal stop)
16	oy	b***o***y	37	dx	bu***tt***er	58	epi	(epinthetic closure)
17	aw	b***ou***ght	38	t	***t***ot	59	qcl	(d closure)
18	ow	b***oa***t	39	g	***g***ag	60	h#	beg. sil.
19	ux	b***eau***ty	40	k	***k***ick	61	#h	end sil.
20	l	***l***ed	41	z	***z***oo	62	pau	betwe. sil.
21	el	bott***le***	42	s	***s***is			

back...

Disambiguating homophones

(words that sound the same but have different meaning)

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

Ice cream	*Four candles*	*Example*
I scream	*Fork handles*	*Egg Sample*

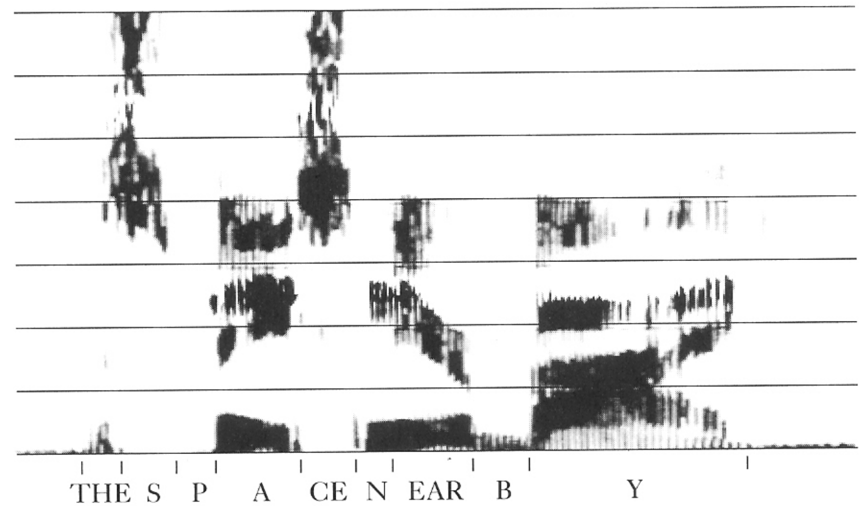
- Systems can only recognize words that are in their lexicon, so limiting the lexicon is an obvious ploy
- Some ASR systems include a grammar which can help disambiguation

(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

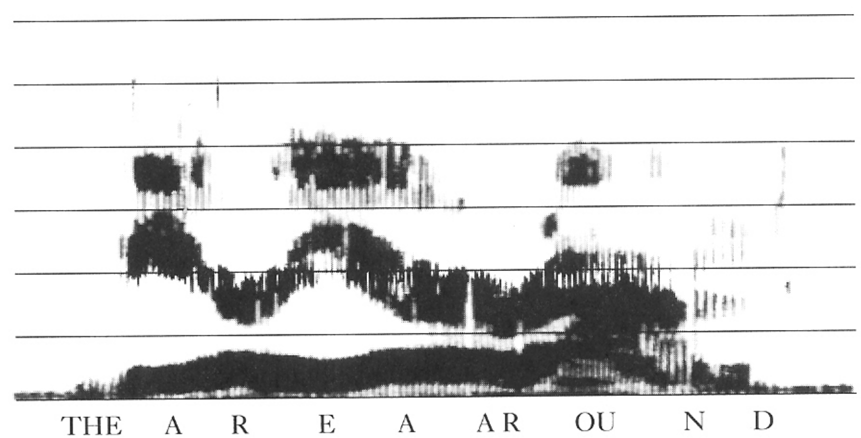
Recognizing Word Boundaries

“THE SPACE NEARBY”
WORD BOUNDARIES CAN BE LOCATED BY THE INITIAL OR FINAL CONSONANTS



(a)

“THE AREA AROUND”
WORD BOUNDARIES ARE DIFFICULT TO LOCATE



(b)

Interpreting prosodic features

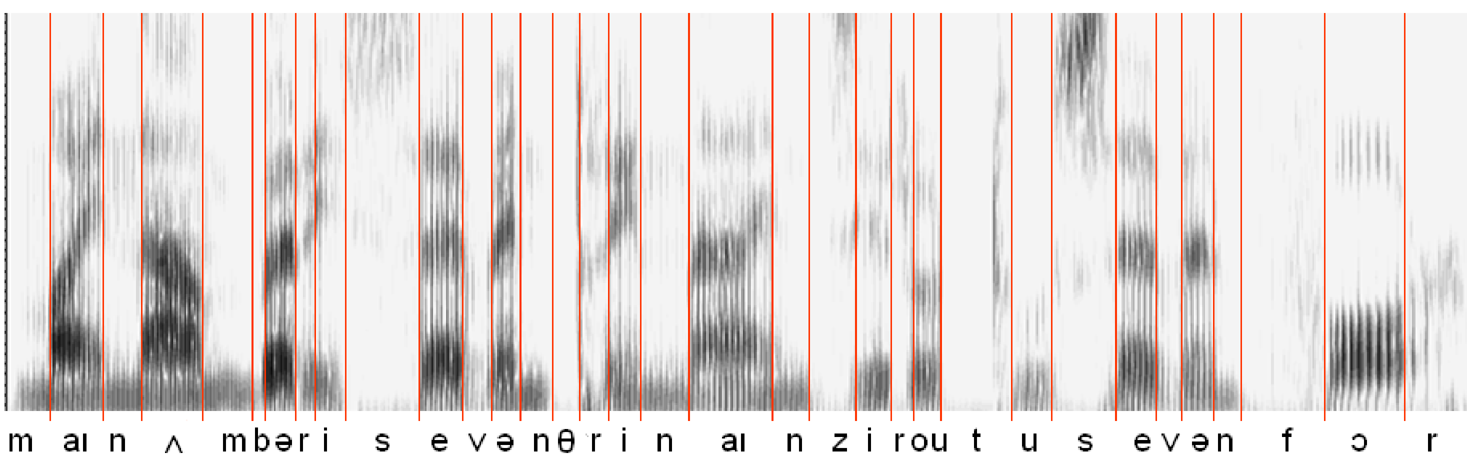
- Pitch, length and loudness are used to indicate “stress”
- All of these are relative
  - On a speaker-by-speaker basis
  - And in relation to context
- Pitch and length are phonemic in some languages

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
- 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

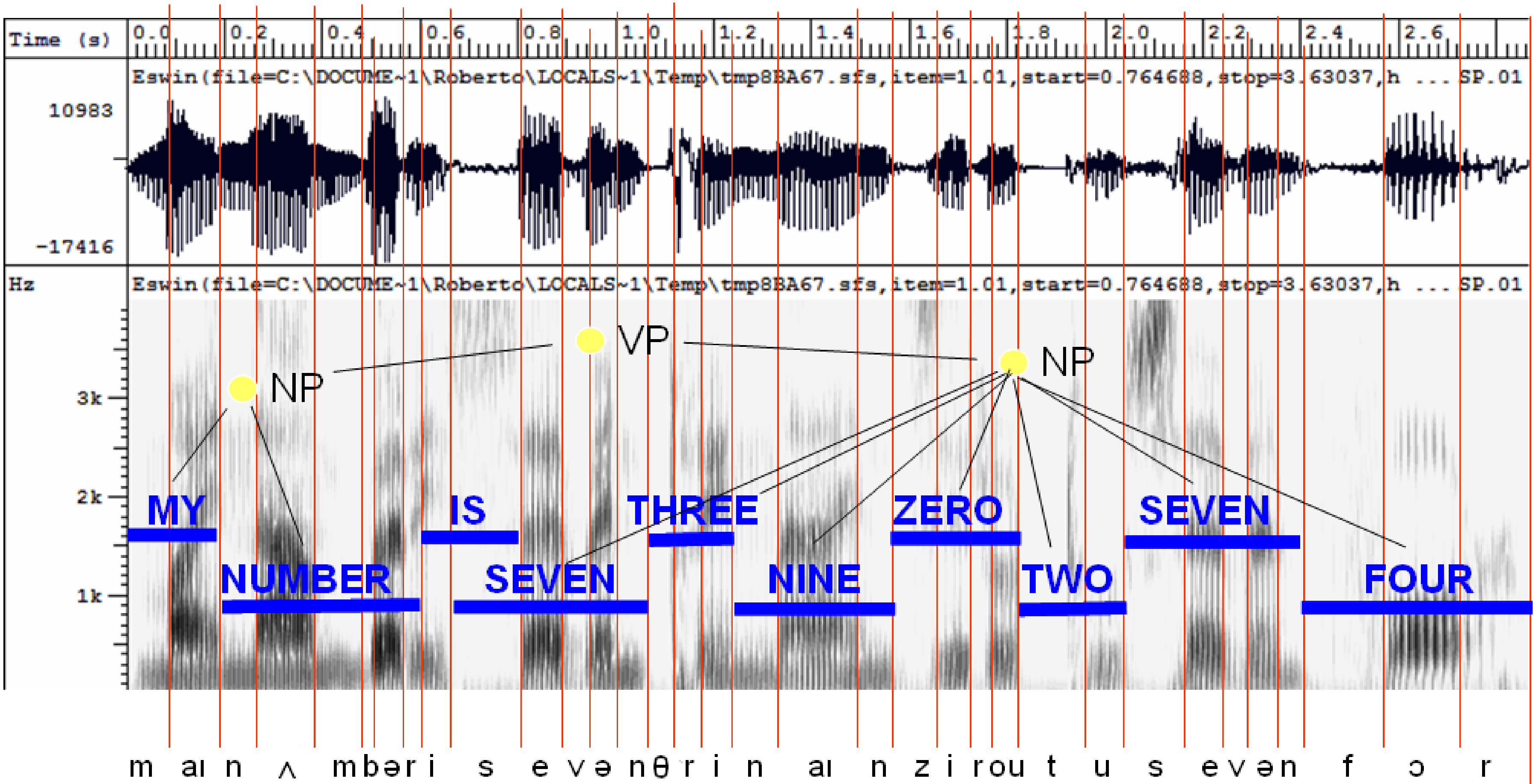
- Hard to distinguish very similar templates
- And quickly degrades when input differs from templates
- Therefore needs techniques to mitigate this degradation:
  - More subtle matching techniques
  - Multiple templates which are aggregated
- Taken together, these suggested …

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



- 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

Overall ASR Architecture

1) Feature Extraction:
   39 “MFCC” ("mel frequency cepstral coefficients“) features

2) Acoustic Model:
   Gaussians for computing $p(o|q)$

3) Lexicon/Pronunciation Model
   - HMM: what phones can follow each other

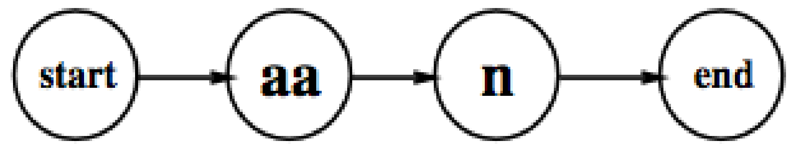
4) Language Model
   - N-grams for computing $p(w_i|w_{i-1})$

5) Decoder
   - Viterbi algorithm: dynamic programming for combining all these to get word sequence from speech!

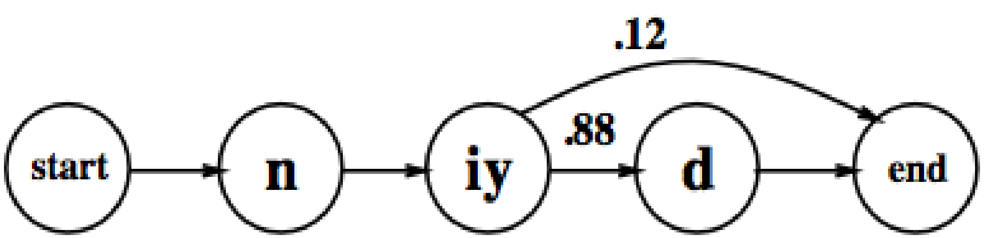
Machine learning

- Acoustic and Lexical Models
  - Analyze 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

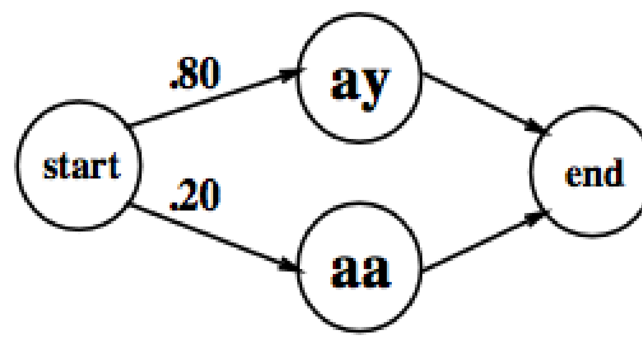
HMMs for some words



Word model for "on"



Word model for "need"

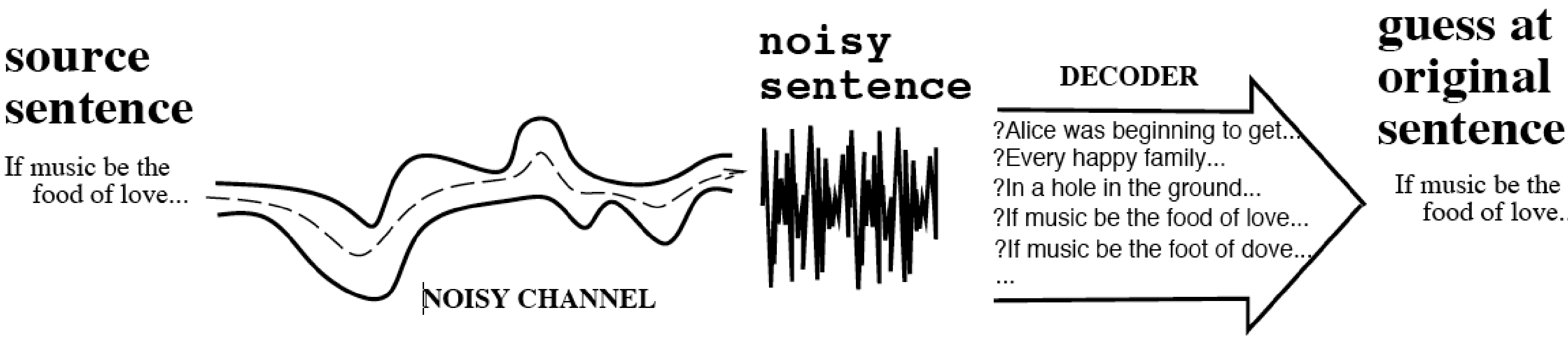


Word model for "I"

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

The Noisy Channel Model (2)

- What is the most likely sentence out of all sentences in the language L given some acoustic input O?
- Treat acoustic input O as sequence of individual observations
  - $O = o_1, o_2, o_3, \ldots, o_t$
- Define a sentence as a sequence of words:
  - $W = w_1, w_2, w_3, \ldots, w_n$

Noisy Channel Model (3)

- Probabilistic implication: Pick the highest prob S: $\hat{W} = \underset{W \in L}{\arg\max} P(W \mid O)$
- We can use Bayes rule to rewrite this:

$$\hat{W} = \underset{W \in L}{\arg\max} \frac{P(O \mid W)P(W)}{P(O)}$$

- Since denominator is the same for each candidate sentence W, we can ignore it for the argmax:

$$\hat{W} = \underset{W \in L}{\arg\max} P(O \mid W)P(W)$$

Noisy channel model

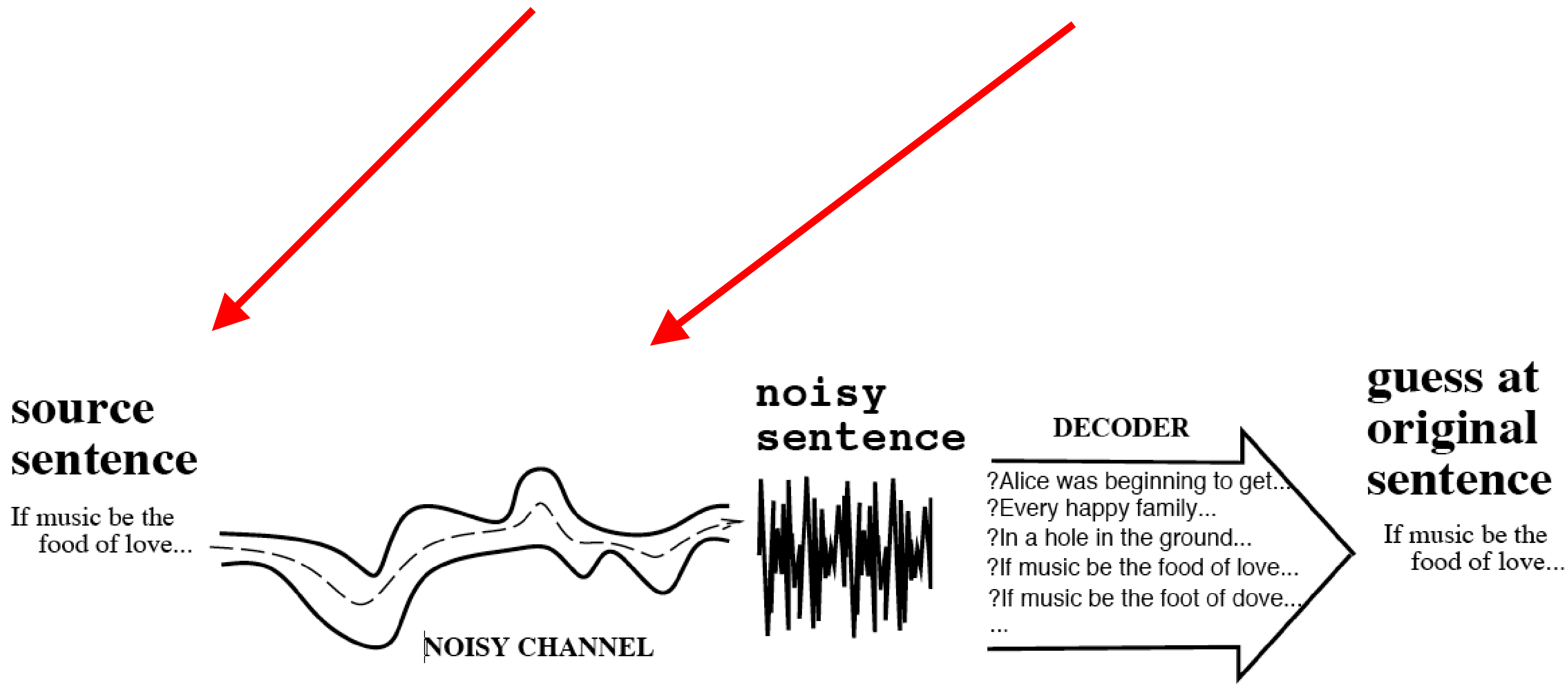
likelihood

prior

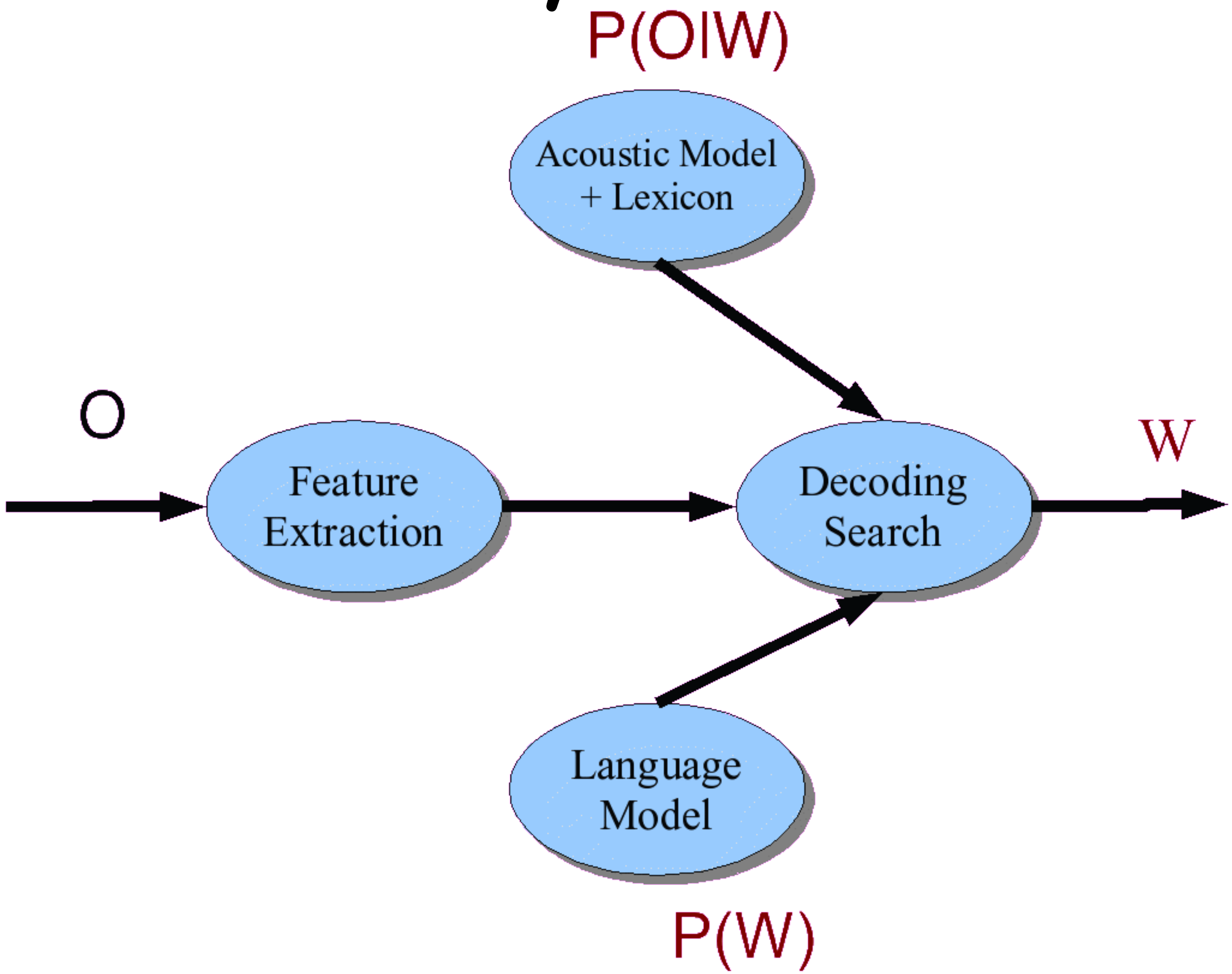
$$\hat{W} = \underset{W \in L}{\operatorname{argmax}} P(O \mid W) P(W)$$

The noisy channel model

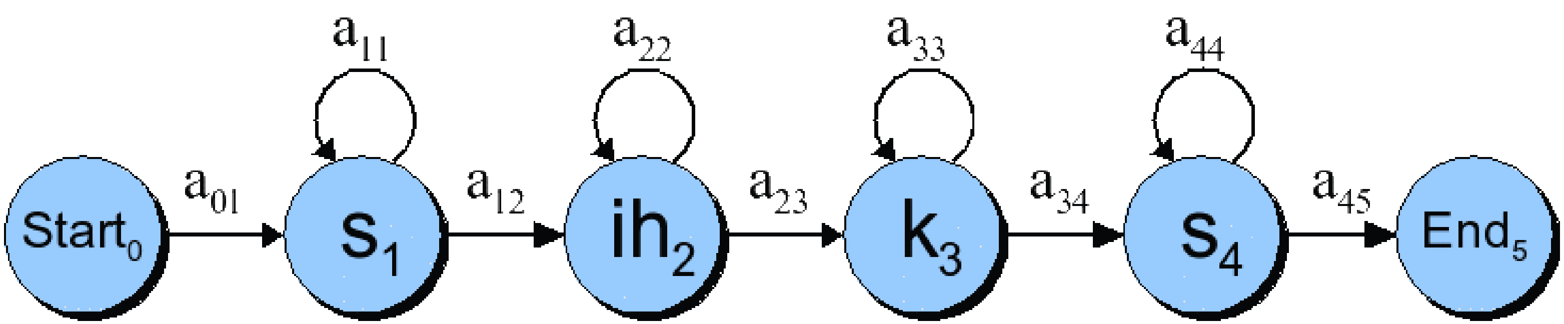
- Ignoring the denominator leaves us with two factors: P(Source) and P(Signal|Source)



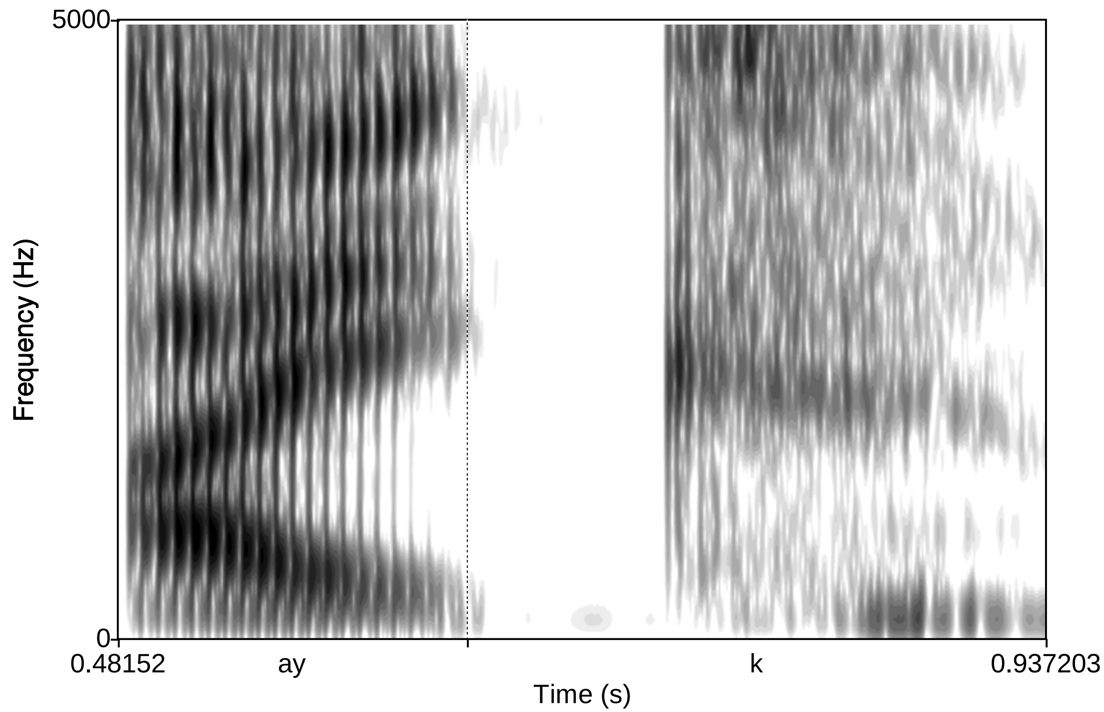
Speech Architecture meets Noisy Channel



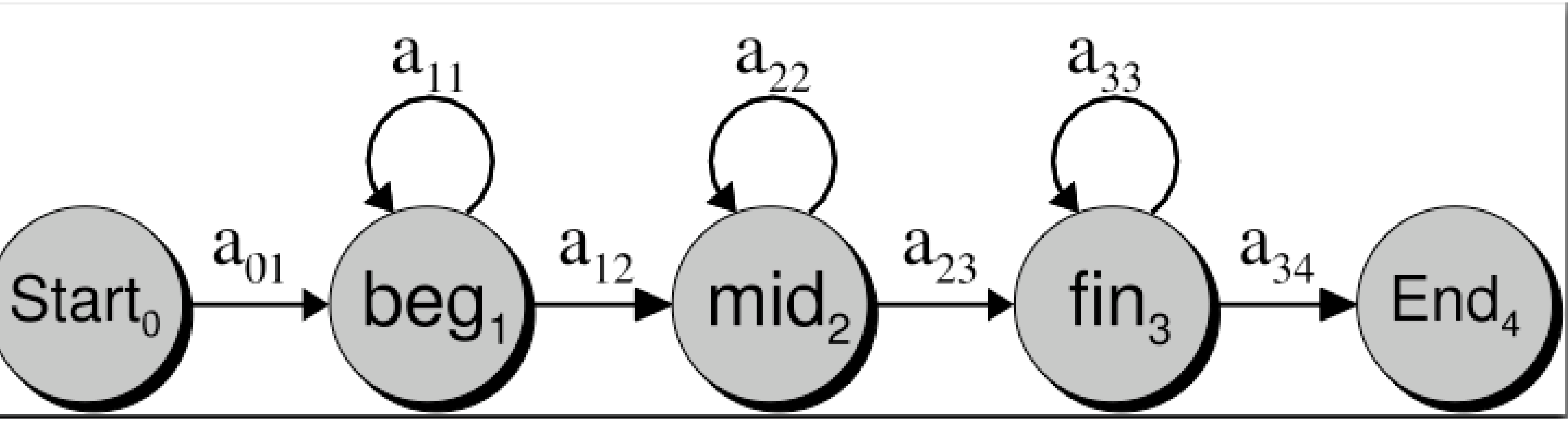
HMMs for speech



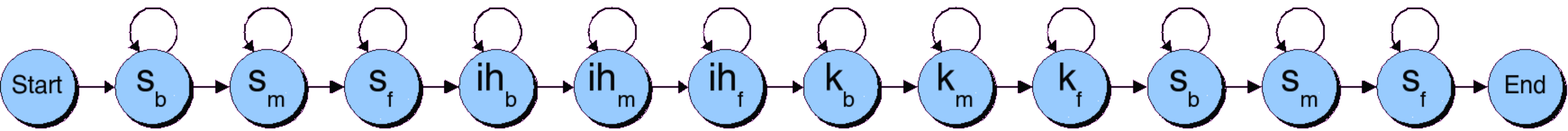
Phones are not homogeneous!



Each phone has 3 subphones



Resulting HMM word model for "six"



HMMs more formally

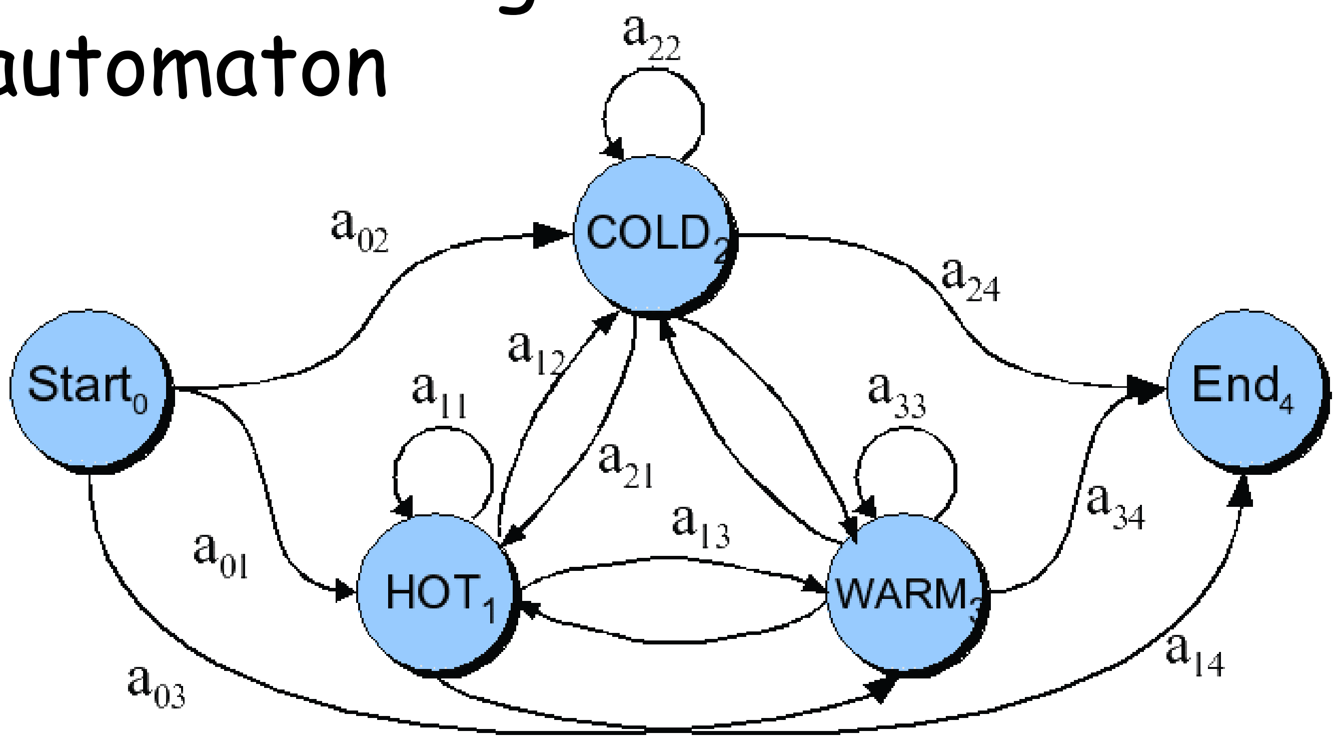
- Markov chains
- A kind of weighted finite-state

$Q = q_1 q_2 \ldots q_N$	a set of **states**
$A = a_{01} a_{02} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
q_0, q_{end}	a special **start and end state** which are not associated with observations.

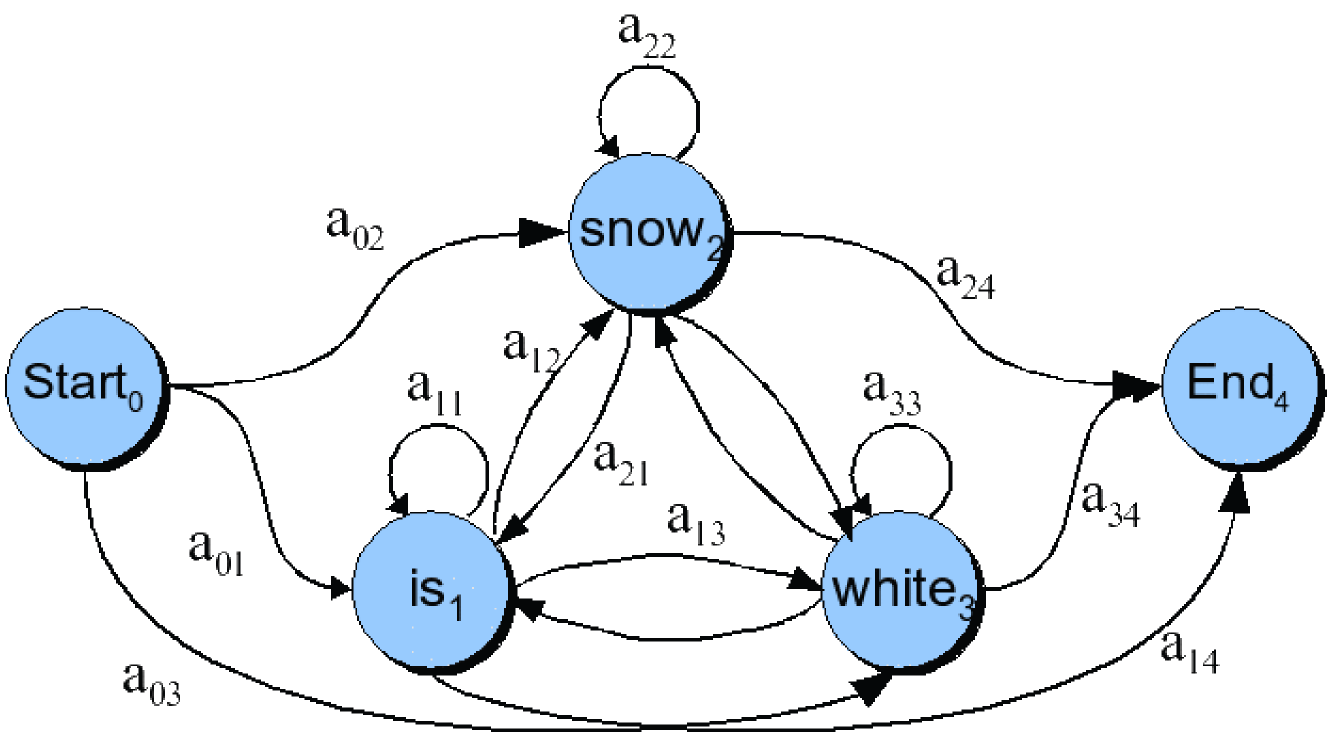
Markov Assumption: $P(q_i | q_1 \ldots q_{i-1}) = P(q_i | q_{i-1})$

HMMs more formally

- Markov chains
- A kind of weighted finite-state automaton



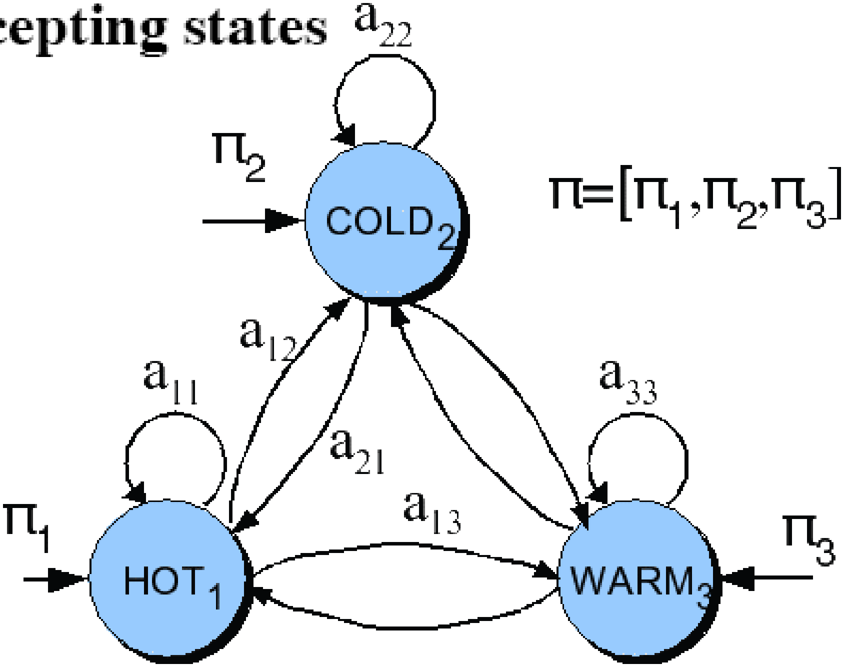
Another Markov chain



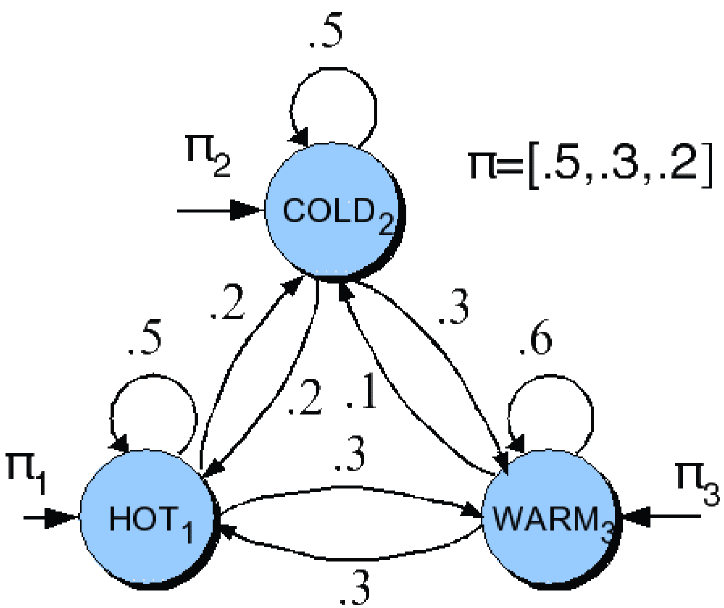
Another view of Markov chains

$\pi = \pi_1, \pi_2, ..., \pi_N$ an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

$QA = \{q_x, q_y...\}$ a set $QA \subset Q$ of legal **accepting states**



An example with numbers:



- What is probability of:
  - Hot hot hot hot
  - Cold hot cold hot

Hidden Markov Models

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$.
$B = b_i(o_t)$	A sequence of **observation likelihoods:**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i.
q_0, q_F	a special **start state** and **end (final) state** which are not associated with observations, together with transition probabilities $a_{01} a_{02} .. a_{0n}$ out of the start state and $a_{1F} a_{2F} ... a_{nF}$ into the end state.

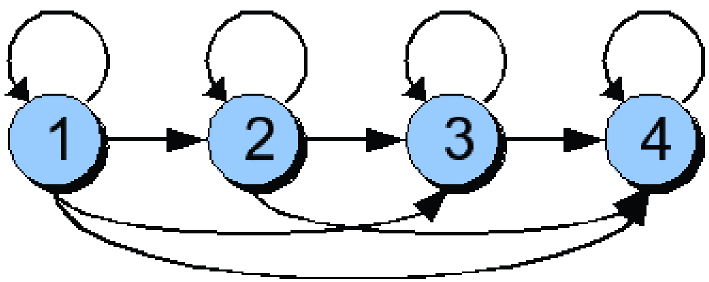
Hidden Markov Models

Markov Assumption: $P(q_i|q_1...q_{i-1}) = P(q_i|q_{i-1})$

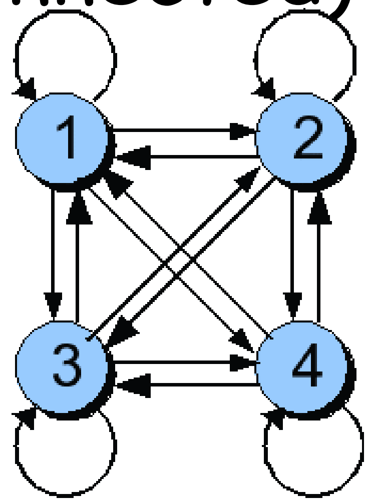
Output Independence Assumption: $P(o_i|q_1 \ldots q_i, \ldots, q_n, o_1, \ldots, o_i, \ldots, o_n) = P(o_i|q_i)$

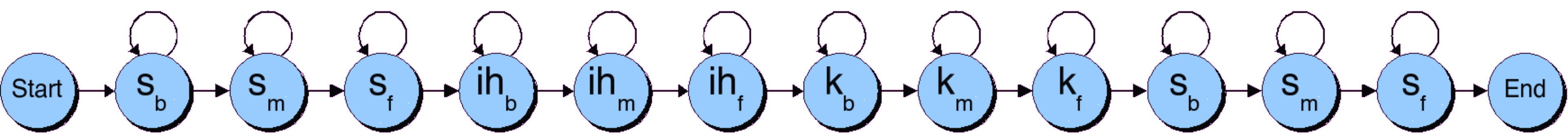
Hidden Markov Models

- Bakis network



Ergodic (fully-connected) network





- Left-to-right network

HMMs more formally

- Three fundamental problems
  - Jack Ferguson at IDA in the 1960s

1) Given a specific HMM, determine likelihood of observation sequence.
2) Given an observation sequence and an HMM, discover the best (most probable) hidden state sequence
3) Given only an observation sequence, learn the HMM parameters (A, B matrix)

The Three Basic Problems for HMMs

- Problem 1 (**Evaluation**): Given the observation sequence $O=(o_1 o_2 \ldots o_T)$, and an HMM model $\Phi = (A,B)$, how do we efficiently compute $P(O|\Phi)$, the probability of the observation sequence, given the model
- Problem 2 (**Decoding**): Given the observation sequence $O=(o_1 o_2 \ldots o_T)$, and an HMM model $\Phi = (A,B)$, how do we choose a corresponding state sequence $Q=(q_1 q_2 \ldots q_T)$ that is optimal in some sense (i.e., best explains the observations)
- Problem 3 (**Learning**): How do we adjust the model parameters $\Phi = (A,B)$ to maximize $P(O|\Phi)$?

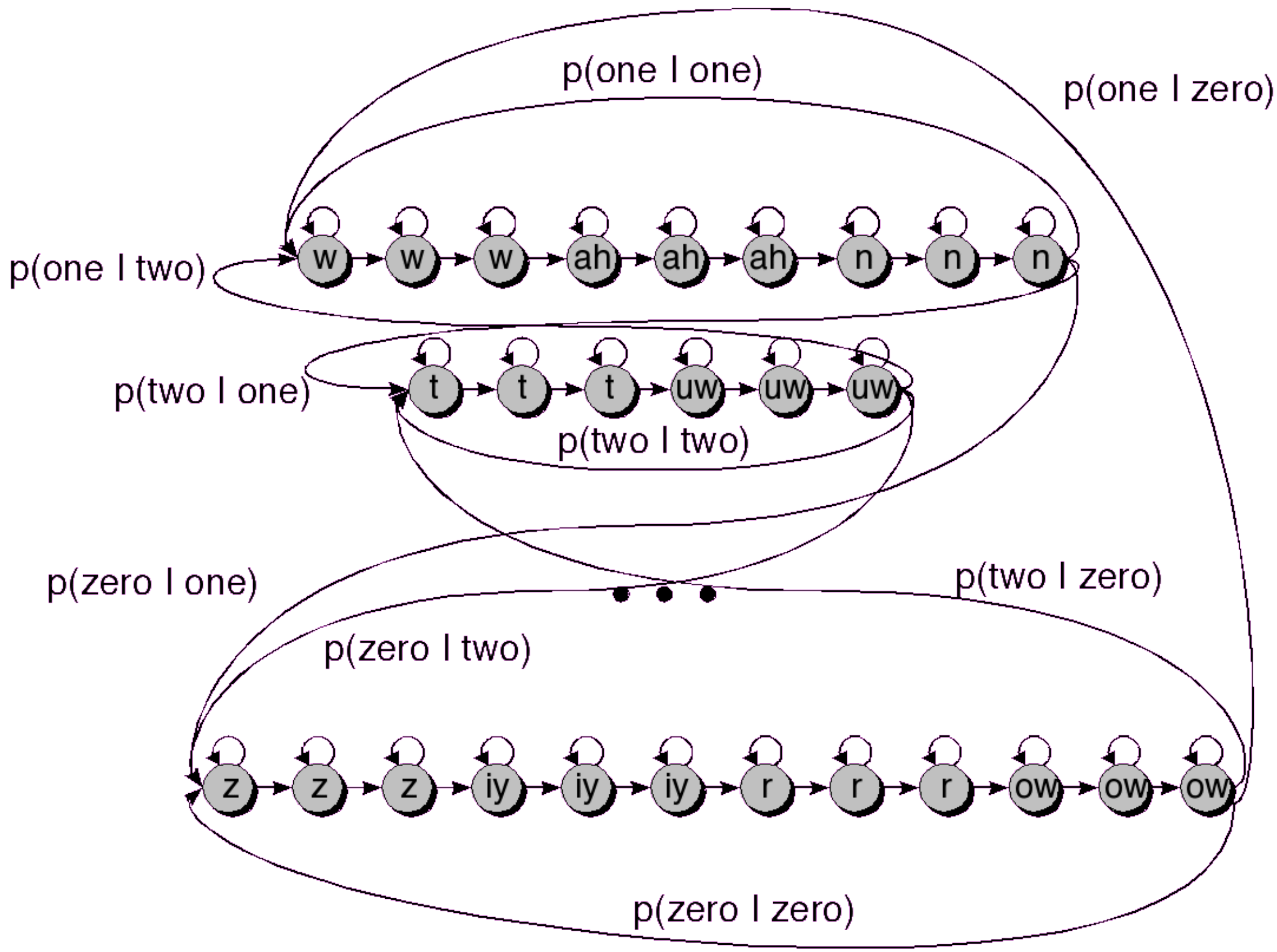
The Forward problem for speech

- The observation sequence O is a series of feature vectors
- The hidden states W are the phones and words
- For a given phone/word string W, our job is to evaluate $P(O|W)$
- Intuition: how likely is the input to have been generated by just that word string W

Evaluation for speech: Summing over all different paths!

- f ay ay ay ay v v v v
- f f ay ay ay ay v v v
- f f f f ay ay ay ay v
- f f ay ay ay ay ay ay v
- f f ay ay ay ay ay ay ay ay v
- f f ay v v v v v v v

Search space with bigrams



Summary: ASR Architecture

Five easy pieces: ASR Noisy Channel architecture

1) Feature Extraction:
   39 "MFCC" features

2) Acoustic Model:
   Gaussians for computing $p(o|q)$

3) Lexicon/Pronunciation Model
   - HMM: what phones can follow each other

4) Language Model
   - N-grams for computing $p(w_i|w_{i-1})$

5) Decoder
   - Viterbi algorithm: dynamic programming for combining all these to get word sequence from speech!

Evaluation of ASR Quality

- 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

Word Error Rate

Word Error Rate =

100 (Insertions + Substitutions + Deletions)

Total Words in Correct Transcript

Aligment example:

REFERENCE:	portable		PHONE	UPSTAIRS	last	night	so
HYPOTHESIS:	portable	FORM	OF	STORES	last	night	so
Evaluation:		I	S	D			

WER = 100 (1+2+0)/6 = 50%

NIST sctk-1.3 scoring software: Computing WER with sclite

- http://www.nist.gov/speech/tools/
- Sclite aligns a hypothesized text (HYP) (from the recognizer) with a correct or reference text (REF) (human transcribed)

```
id: (2347-b-013)
Scores: (#C #S #D #I) 9 3 1 2
REF:  was an engineer SO   I    i was always with ****  ****   MEN UM   and they
HYP:  was an engineer **  AND i was always with THEM THEY ALL THAT and they
Eval:                     D  S                      I    I     S   S
```

Better metrics than WER?

- WER has been useful
- But should we be more concerned with meaning ("semantic error rate")?
  - Good idea, but hard to agree on
  - Has been applied in dialogue systems, where desired semantic output is more clear

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