Hidden Markov Models: lecture 8

Application to speech recognition

Xavier Didelot

Introduction to speech recognition

- Understanding spoken language understanding is a difficult task, and it is remarkable that humans do as well at it as we do
- The goal of automatic speech recognition research is to address this problem
- Hidden Markov Models have a long successful history of being used in this field computationally by building systems that map from an acoustic signal to a string of words
- Comparison of error rates:

Task	Vocabulary	Machine	Human
Digits	11	0.5	0.009
Wall Street Journal clean	5k	3	0.9
Wall Street Journal noisy	5k	9	1.1
Broadcast new	65k	20	4

Prior probability of a word

- Even before we hear the beginning of a word, not all words are equally likely...
- An N-gram probability is the conditional probability of a word given the previous N-1 words
- N-gram probabilities can be computed by simply counting in a corpus and normalizing (the Maximum Likelihood Estimate) or they can be computed by more sophisticated algorithms
- The advantage of N-grams is that they take advantage of lots of rich lexical knowledge
- A disadvantage for some purposes is that they are very dependent on the corpus they were trained on

Bigram

Bigram for eight of the words (out of $V = 1446$) in a corpus of 9332 sentences.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Random sentences from Shakespeare's work

Unigram

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have

Every enter now severally so, let

Hill he late speaks; or! a more to leg less first you enter

Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.

Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.

What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.

This shall forbid it should be branded, if renown made it empty.

Indeed the duke; and had a very good friend.

Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;

Will you not tell me who I am?

It cannot be but so.

Indeed the short and the long. Marry, 'tis a noble Lepidus.

Random sentences from Wall Street Journal

Unigram
Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
Bigram
Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
Trigram
They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Phones

- We can represent the pronunciation of words in terms of units called phones
- The standard system for representing phones is the International Phonetic Alphabet or IPA

International Phonetic Alphabet (IPA) ˌɪntəˈnæʃn̩l fəˈnɛtɪk ˈælfəˌbɛt

Consonants (pulmonic)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

- The most common computational system for transcription of English is the ARPAbet, which conveniently uses ASCII symbols

Phones for consonants

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]
[t]	[t]	tea	[t iy]
[k]	[k]	cook	[k uh k]
[b]	[b]	bay	[b ey]
[d]	[d]	dill	[d ih l]
[g]	[g]	garlic	[g aa r l ix k]
[m]	[m]	mint	[m ih n t]
[n]	[n]	nutmeg	[n ah t m eh g]
[ng]	[ŋ]	baking	[b ey k ix ng]
[f]	[f]	flour	[f l aw axr]
[v]	[v]	clove	[k l ow v]
[th]	[θ]	thick	[th ih k]
[dh]	[ð]	those	[dh ow z]
[s]	[s]	soup	[s uw p]
[z]	[z]	eggs	[eh g z]
[sh]	[ʃ]	squash	[s k w aa sh]
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]
[ch]	[tʃ]	cherry	[ch eh r iy]
[jh]	[dʒ]	jar	[jh aa r]
[l]	[l]	licorice	[l ih k axr ix sh]
[w]	[w]	kiwi	[k iy w iy]
[r]	[r]	rice	[r ay s]
[y]	[j]	yellow	[y eh l ow]
[h]	[h]	honey	[h ah n iy]
Less commonly used phones and allophones			
[q]	[ʔ]	uh-oh	[q ah q ow]
[dx]	[ɾ]	butter	[b ah dx axr]
[nx]	[ɾ̃]	winner	[w ih nx axr]
[el]	[l̩]	table	[t ey b el]

Phones for vowels

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɝ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]
Reduced and uncommon phones			
[ax]	[ə]	lotus	[l ow dx ax s]
[axr]	[ɚ]	heather	[h eh dh axr]
[ix]	[ɨ]	tulip	[t uw l ix p]
[ux]	[ʉ]	dude[1]	[d ux d]

Words are not deterministic sequences of phones

- For example there are several pronunciations possible for "because" and "about":

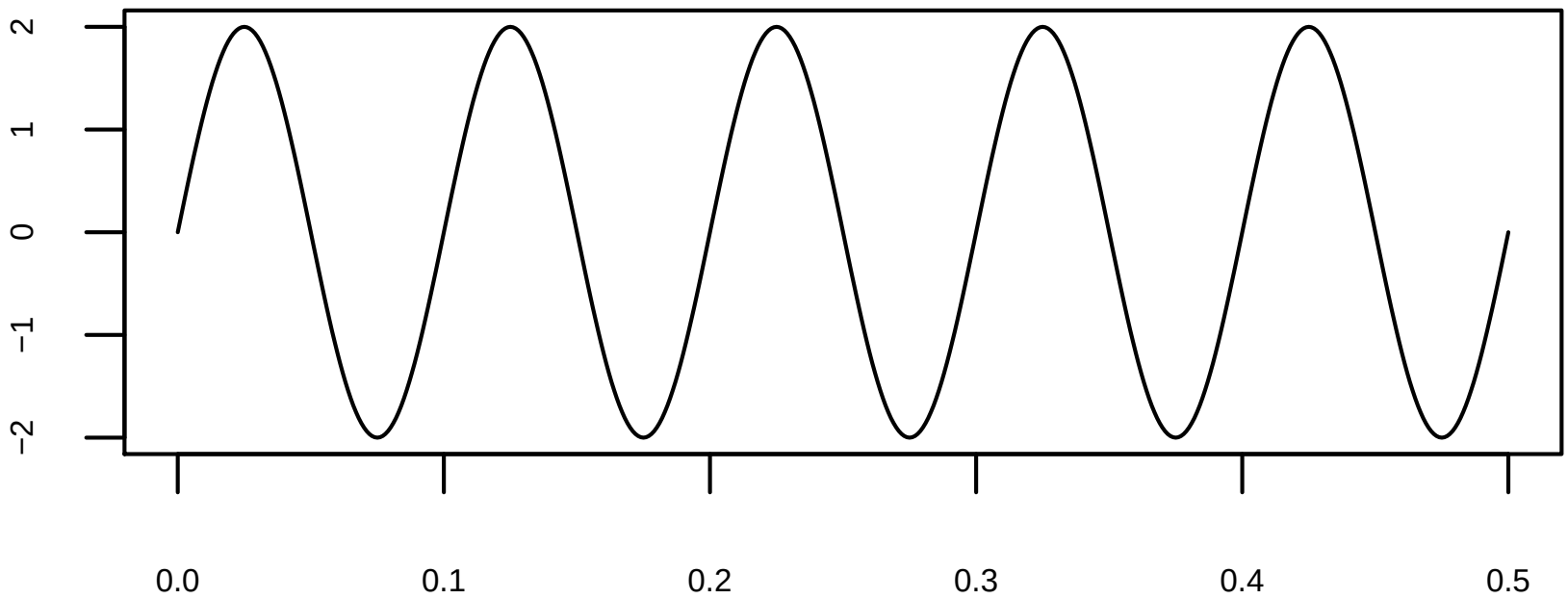
because				about			
ARPAbet	**%**	**ARPAbet**	**%**	**ARPAbet**	**%**	**ARPAbet**	**%**
b iy k ah z	27%	k s	2%	ax b aw	32%	b ae	3%
b ix k ah z	14%	k ix z	2%	ax b aw t	16%	b aw t	3%
k ah z	7%	k ih z	2%	b aw	9%	ax b aw dx	3%
k ax z	5%	b iy k ah zh	2%	ix b aw	8%	ax b ae	3%
b ix k ax z	4%	b iy k ah s	2%	ix b aw t	5%	b aa	3%
b ih k ah z	3%	b iy k ah	2%	ix b ae	4%	b ae dx	3%
b ax k ah z	3%	b iy k aa z	2%	ax b ae dx	3%	ix b aw dx	2%
k uh z	2%	ax z	2%	b aw dx	3%	ix b aa t	2%

Waves

- Acoustic analysis is based on trigonometric functions.

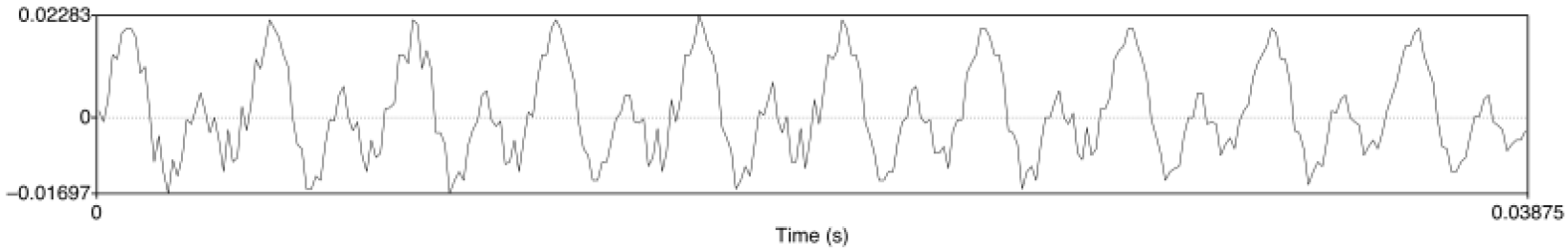
$$y = A\sin(2\pi ft)$$

- x is time measured in seconds
- y represents air pressure above and below the normal atmospheric pressure
- A is the signal amplitude
- f is the signal frequency
- For example if $A = 2$ and $f = 10$ Hertz:



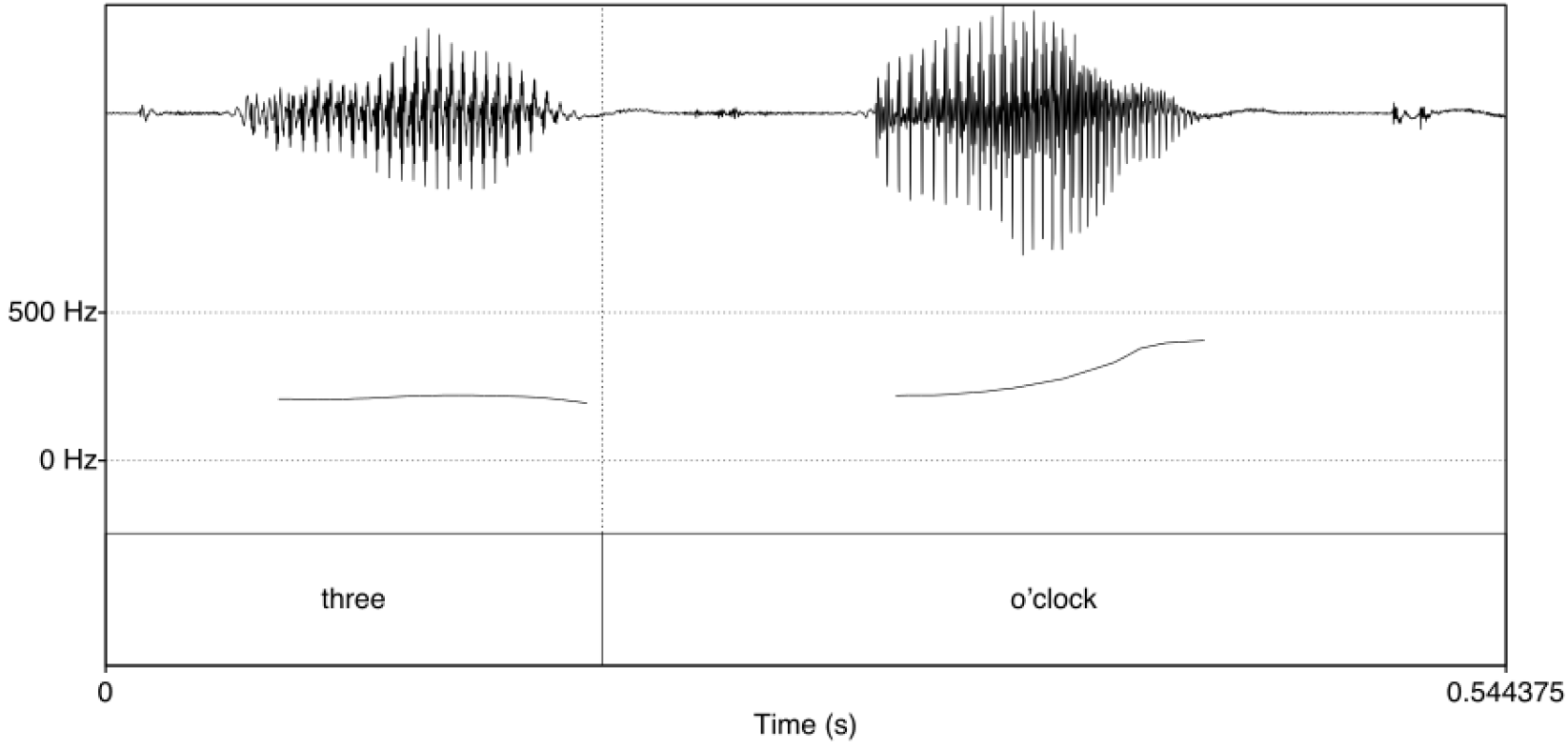
Real sounds

- Real spoken sounds are complex series of change in air pressure
- These changes in air pressure are caused by the specific way that air passes through the glottis and out the oral or nasal cavities
- For example the vowel [iy] from someone saying “She just had a baby”



Fundamental frequency

- Real sounds are more complex than the simple sine, but the signal still looks periodic with a frequency which is called the fundamental frequency
- Perceived pitch is (roughly) proportional to the fundamental frequency
- The fundamental frequency varies over time:

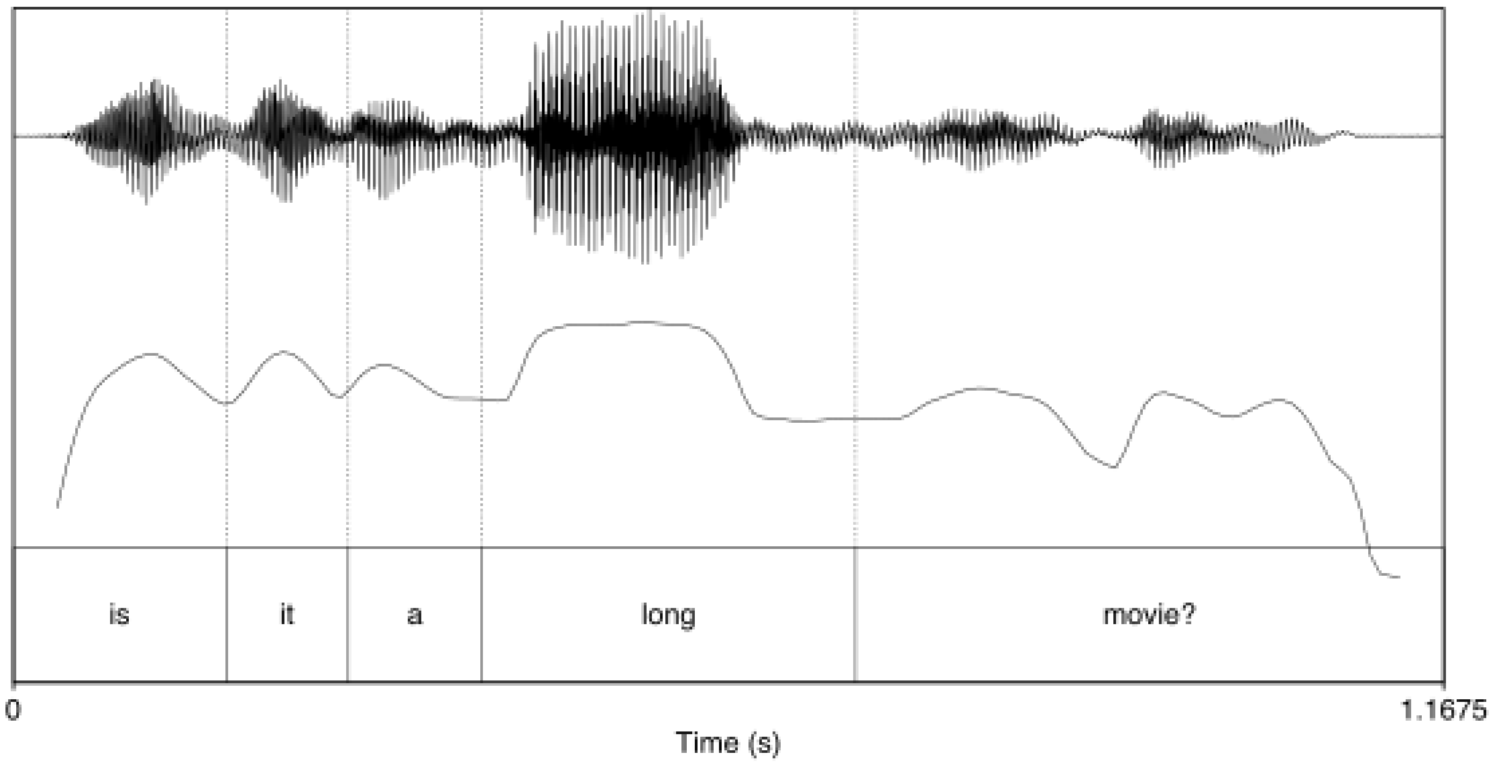


Power and intensity over time

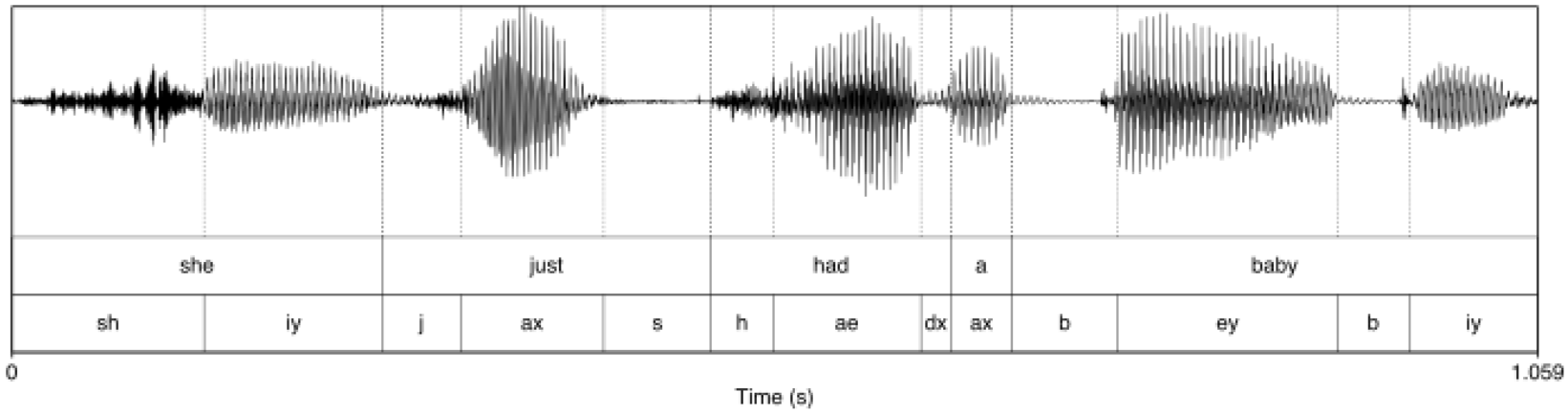
- The power and intensity of the signal are defined as:

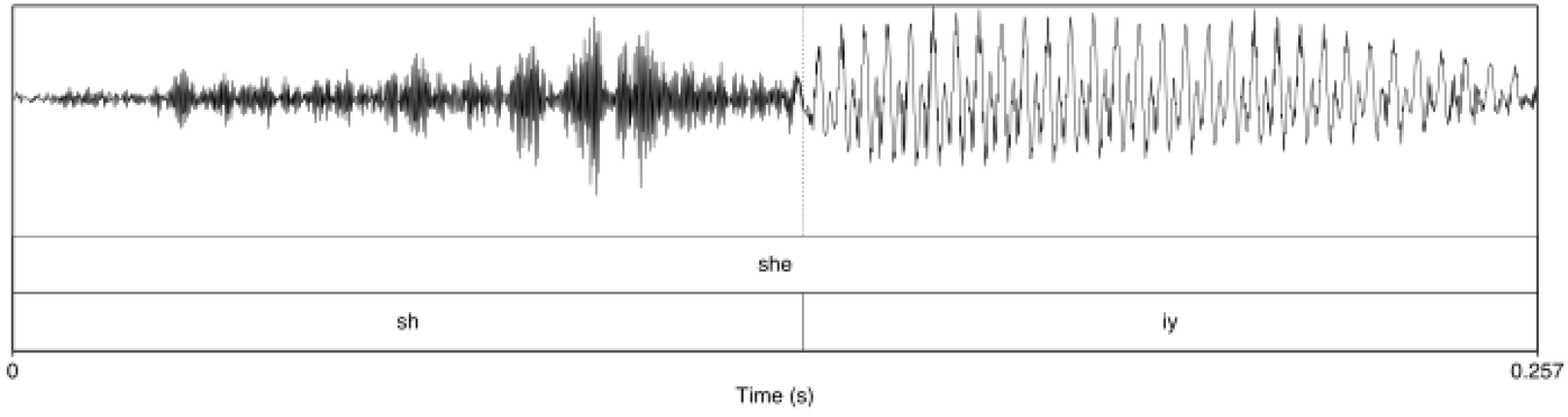
$$P = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \text{ and } I = 10\log_{10}(P/P_0)$$

- Perceived loudness is (roughly) proportional to power
- Variation of intensity over time:



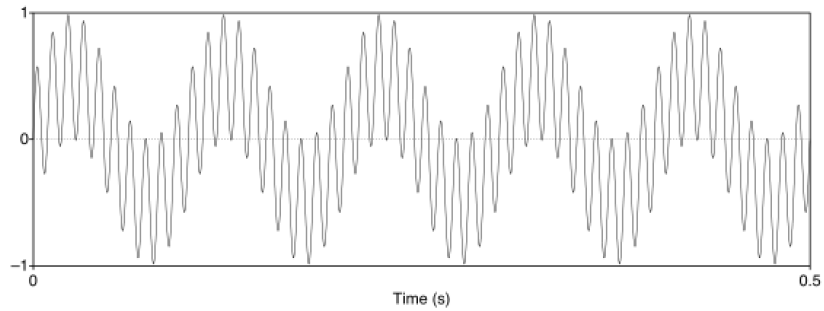
She just had a baby



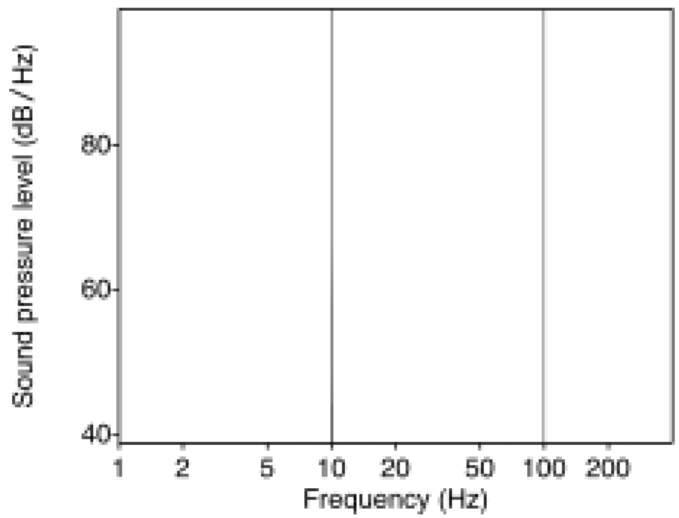


Fourier analysis and spectrum

- Idea: decompose complex waves into a sum of many sine waves of different frequencies
- For example the sum of two sine waveforms, one of frequency 10 Hz and one of frequency 100 Hz, both with amplitude 1:

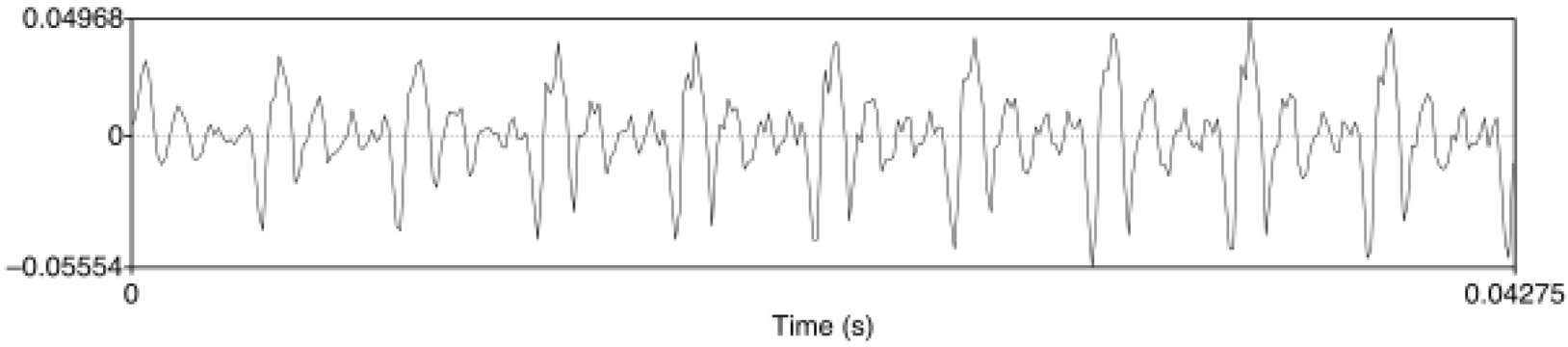


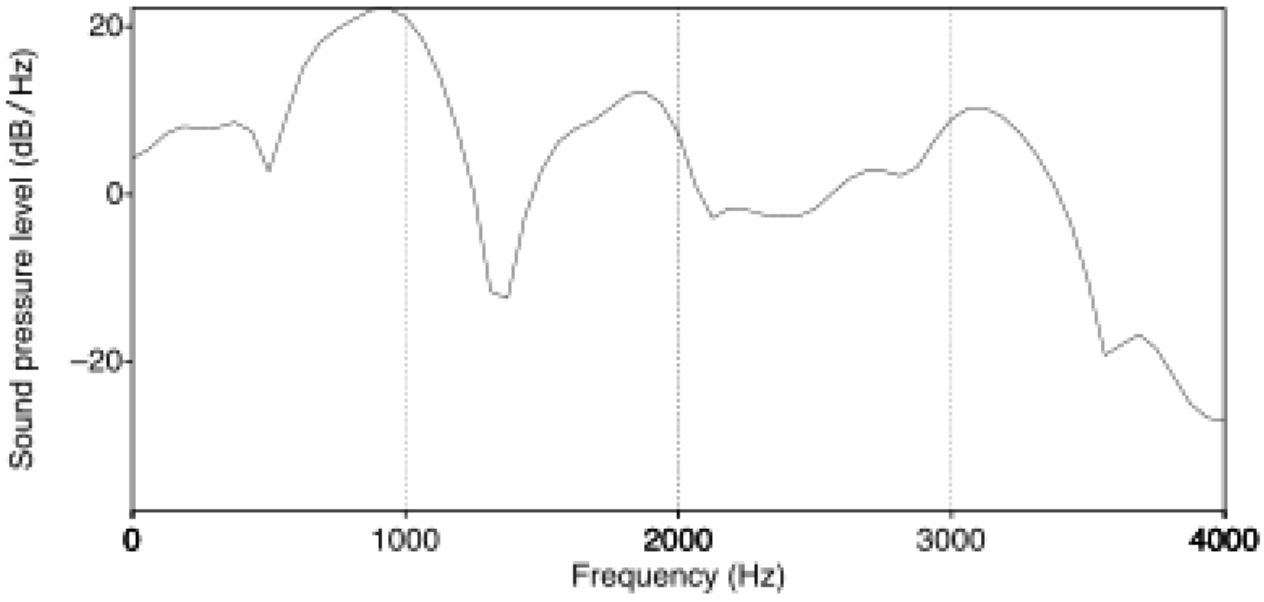
- We can represent the two component frequencies with a spectrum.



Discrete Fourier Transform

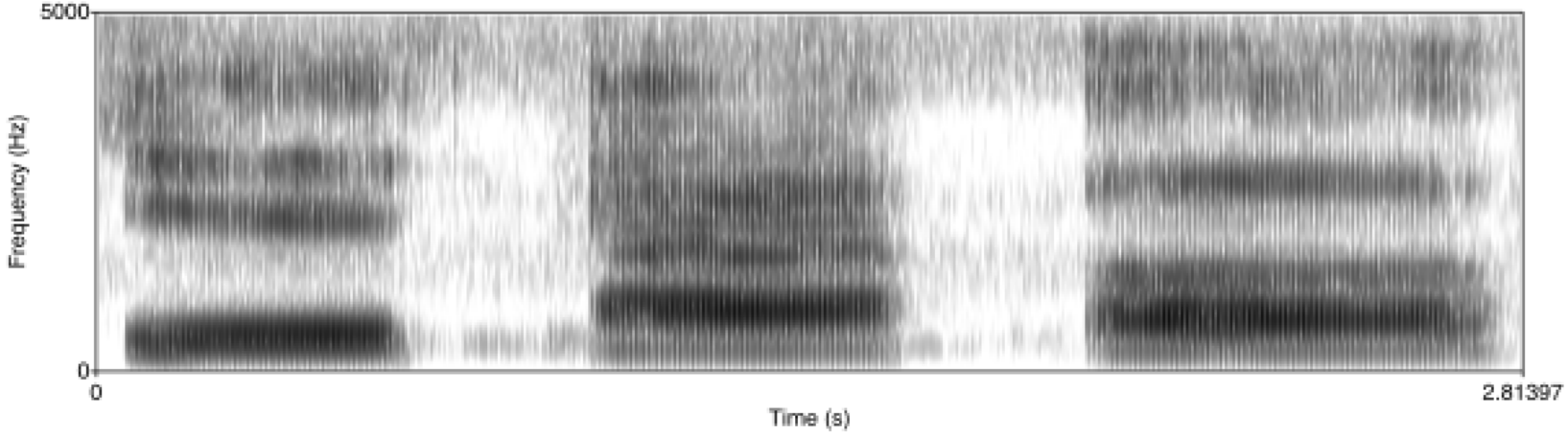
- Discrete Fourier Transform (DFT) computes a spectrum
- For example the vowel [ae] from the word "had"





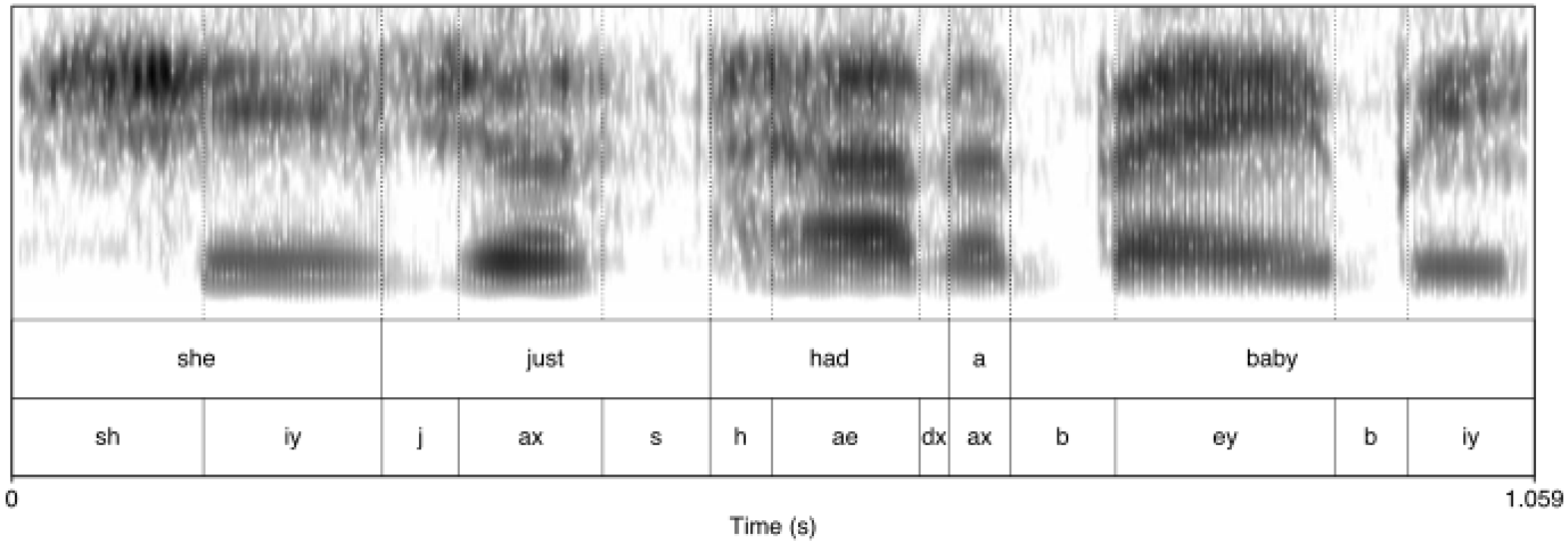
Spectrogram

- A spectrogram shows how the spectrum changes over time
- The x-axis is time
- The y-axis is the frequency shown on the x-axis of the spectrum
- The darkness represents the amplitude shown on the y-axis of the spectrum
- For example [ih], [ae], and [uh]

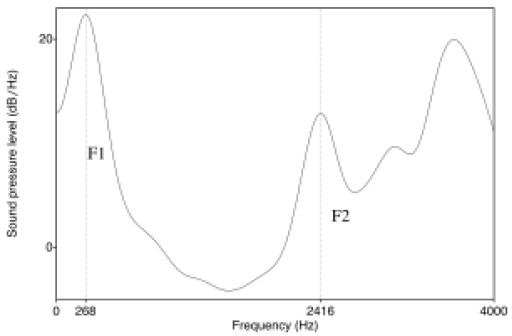


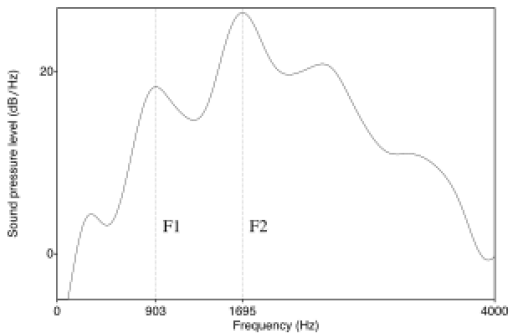
Formants

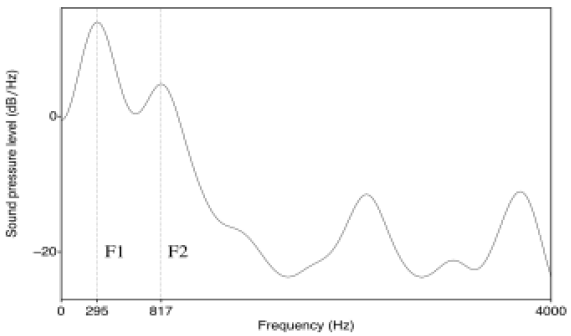
- Black bars on spectrograms correspond to peaks in spectrums and are called formants
- “She just had a baby”

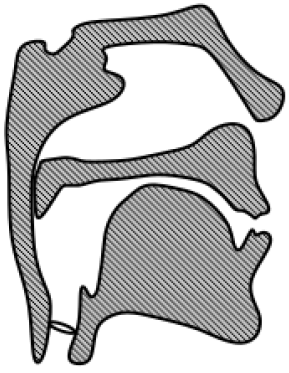


Comparing three phones

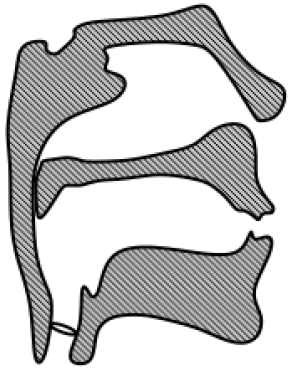




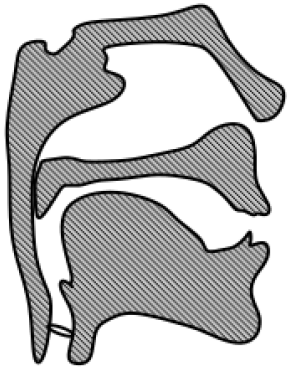




[iy] (tea)

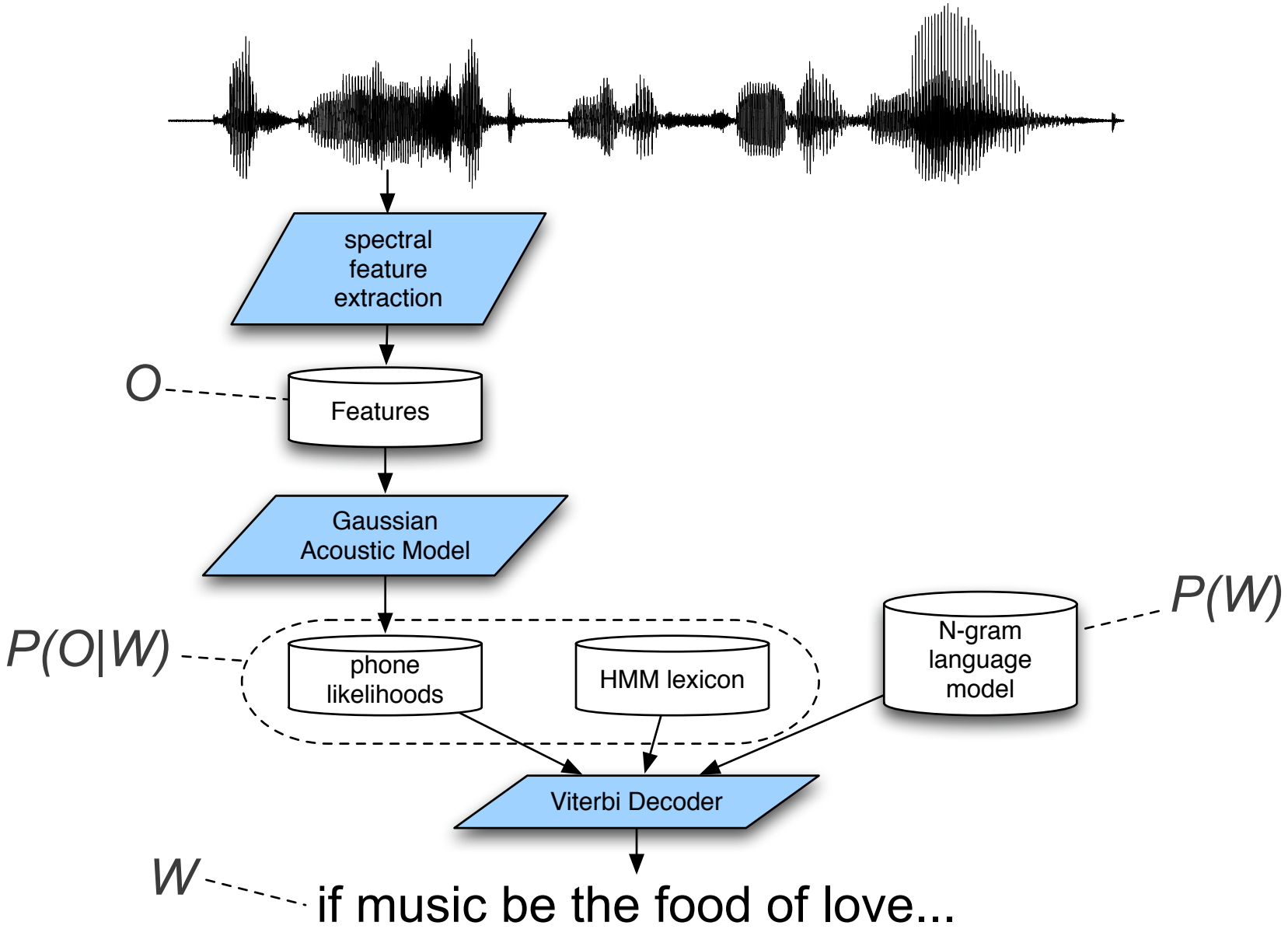


[ae] (cat)



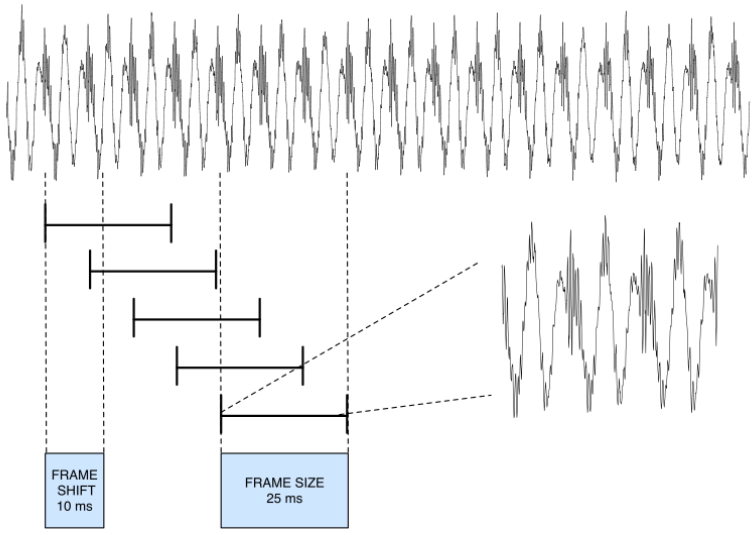
[uw] (moo)

Schematic for speech recognition of a sentence

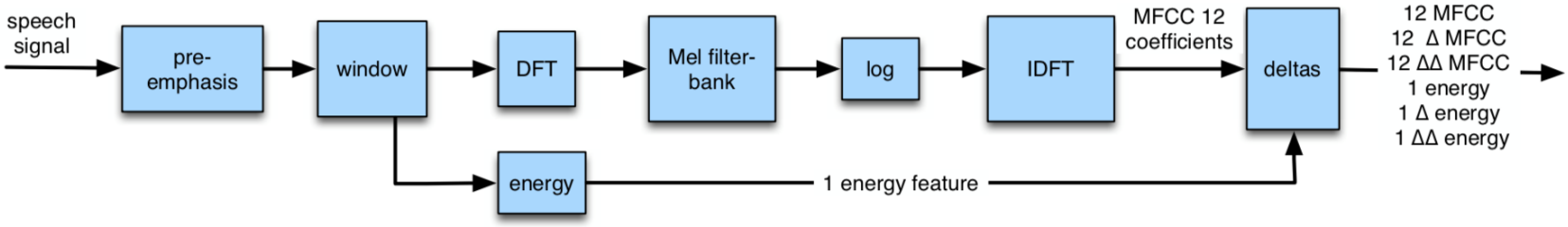


Features extraction

- Every ~10ms a spectrum is computed based on the next ~25ms

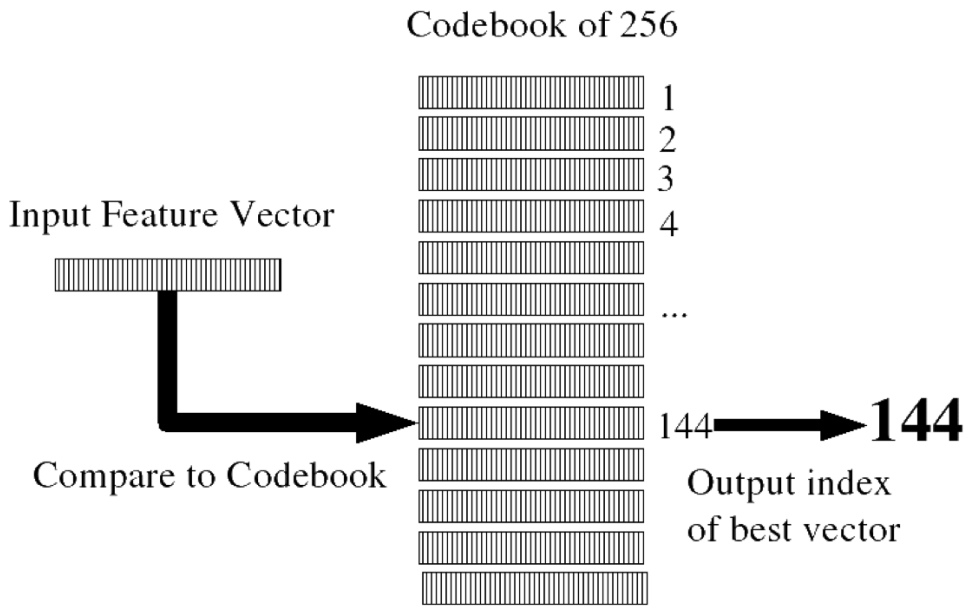


- From each spectrum, 39 features are computed (MFCC)



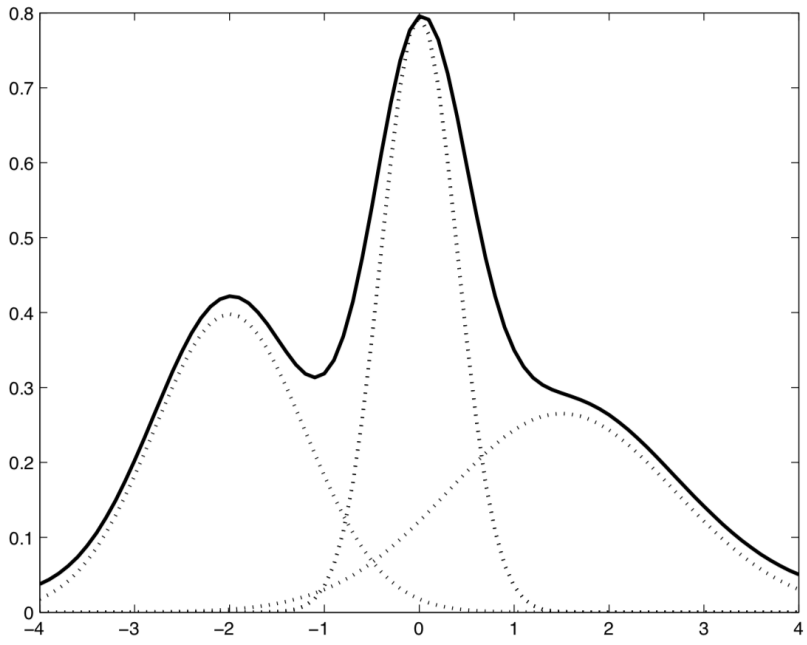
Vector quantization

- Old method, no longer used
- Compare feature vector to codebook
- Given an observation, select nearest vector from codebook



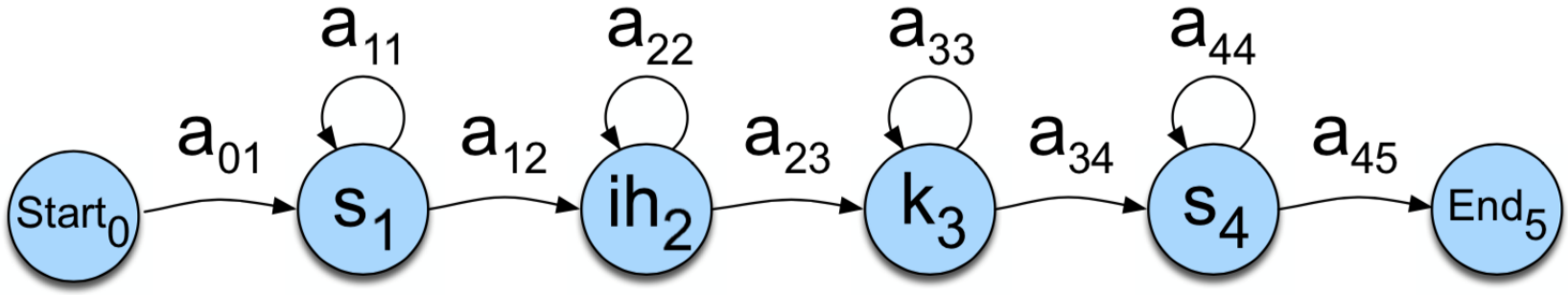
Gaussian acoustic model

- Consider that each feature is Normally distributed
- Learn parameters from training set
- Multivariate Gaussian: consider covariance between features
- Even better: mixture of Gaussians
- GMM: Gaussian Mixture Model
- Training similar to Baum-Welch algorithm



Building HMMs for speech recognition

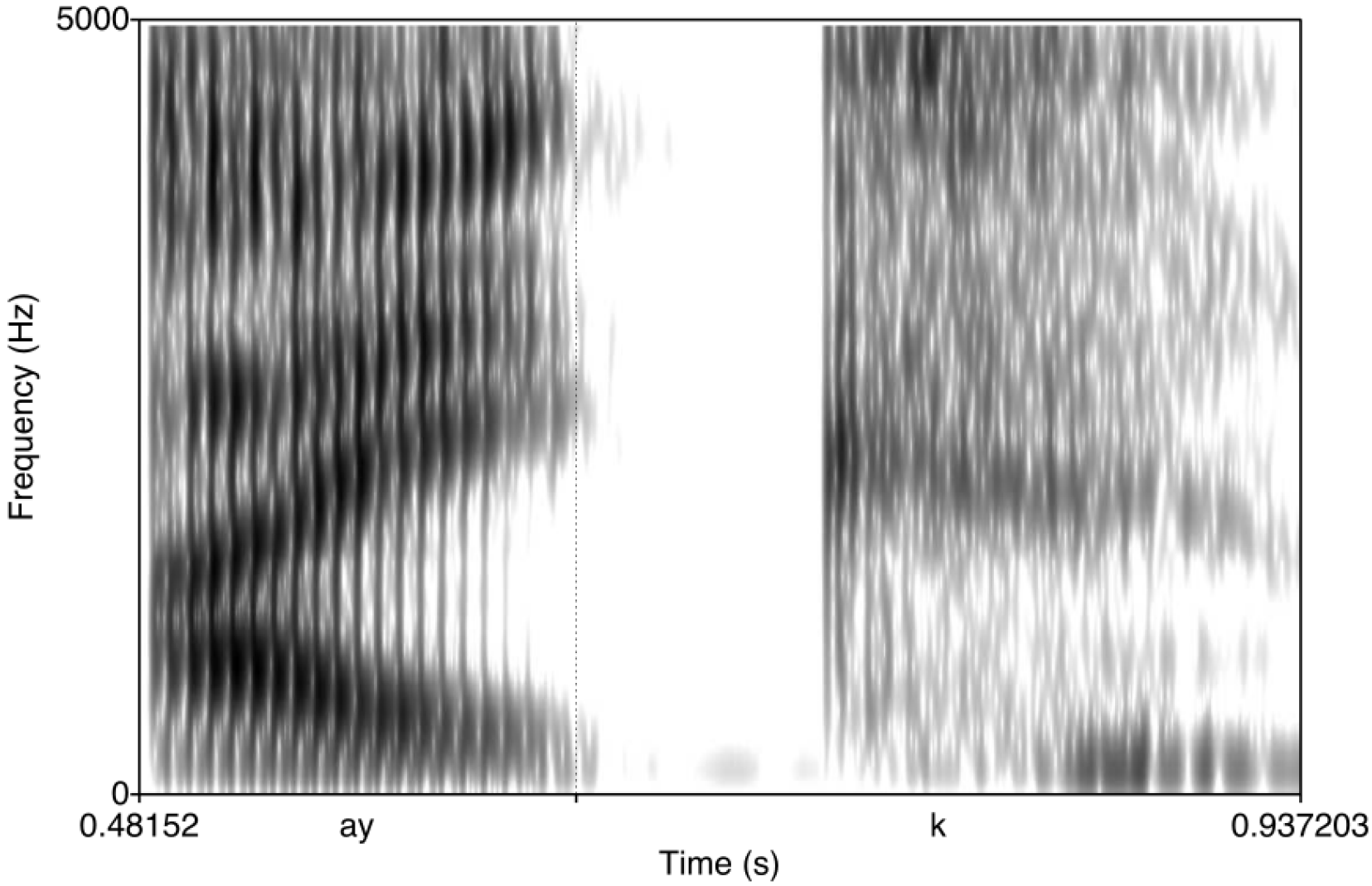
- Since words are made of phones, we consider a HMM in which each state is a phone
- For example for the word "six":



- Note the self-loops, which allow a single phone to repeat so as to cover a variable amount of the acoustic input
- Phone durations vary hugely, dependent on the phone identify, the speaker's rate of speech, the phonetic context, and the level of prosodic prominence of the word

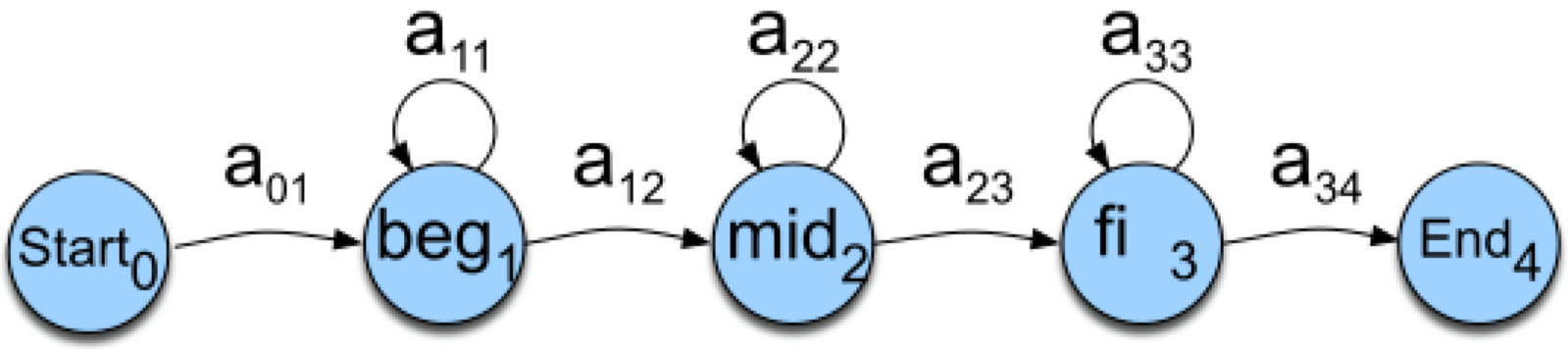
Spectrum variation within a phone

- The spectrum of a phone is not always constant, and the variation is sometimes really important
- For example the word “Ike” pronounced [ay k]:

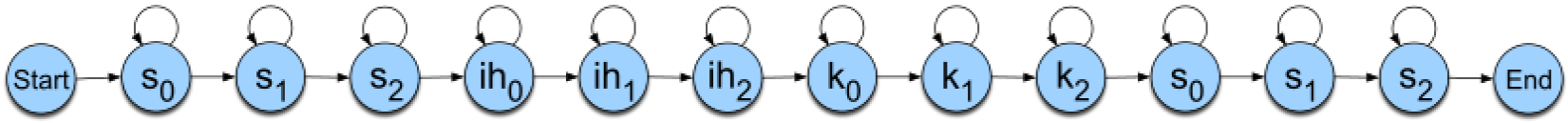


Spectrum variation within a phone

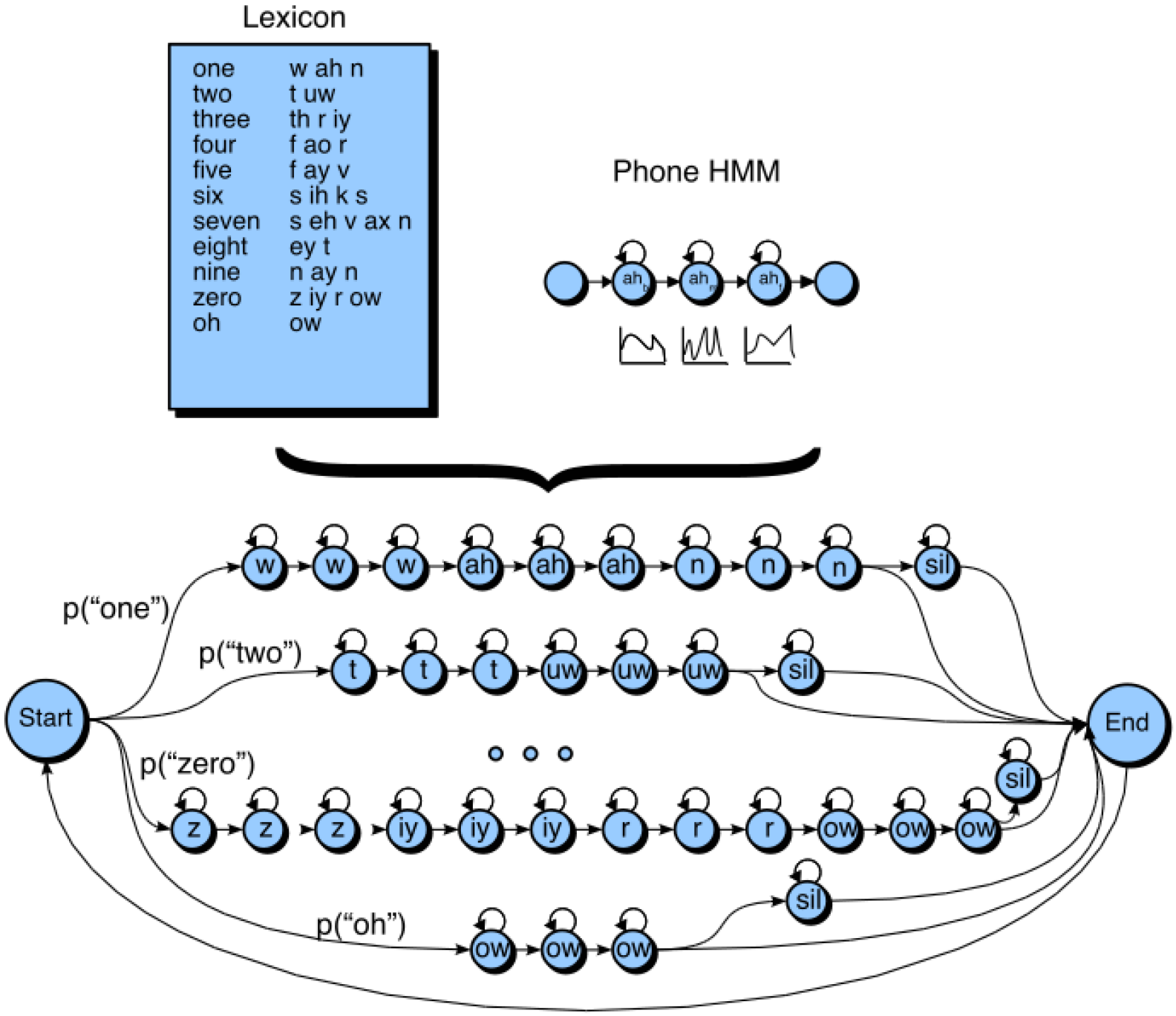
- To account for this, we consider that a phone is made of three states:



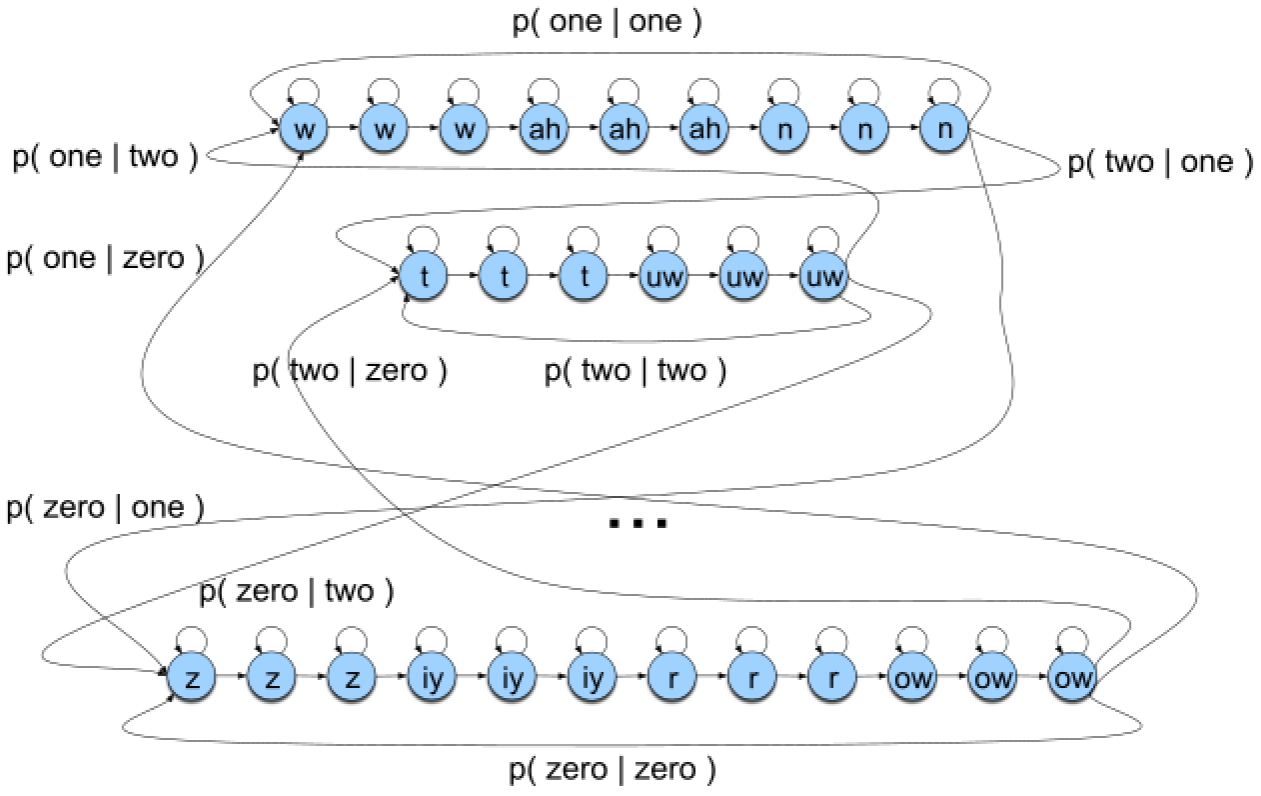
- For the word “six”, this gives:



HMM for the digit recognition task



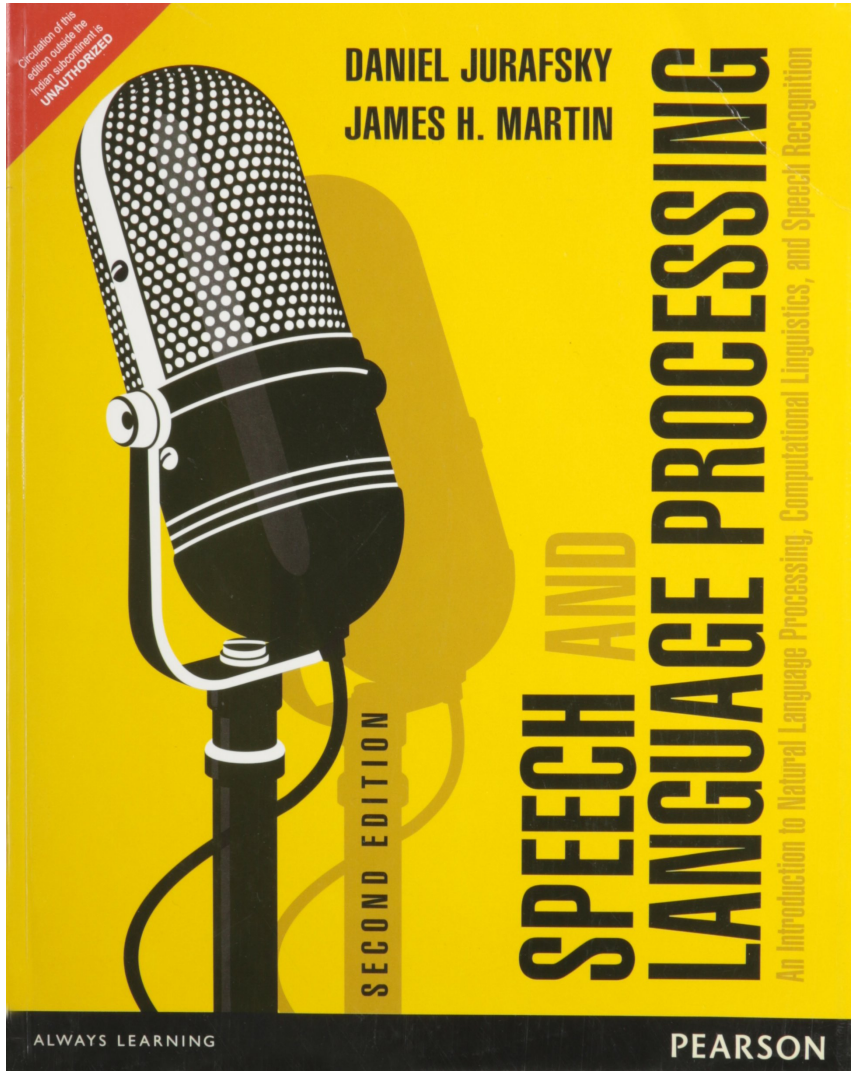
Bigram grammar network for the digit recognition task



HMM algorithms

- Now that we have a well defined HMM, we can use the algorithms we have learnt in previous lectures
- Training using the Baum-Welch algorithm on a large number of known words
- Forward algorithm to calculate $p(O|W)$ for several words
- Viterbi algorithm for global decoding

To find out more. . .



Conclusions

- The input to a speech recognizer is a series of acoustic waves
- Waveform, spectrum and spectrogram are useful visualization tools
- In the first step in speech recognition, sound waves are sampled, quantized, and converted into spectral feature vectors (often MFCC)
- GMM acoustic models are used to estimate the phonetic likelihoods of these feature vectors for each frame
- HMM models combine phone (or parts of phones) into words
- Can also account for independence of successive words via N-gram
- Efficiency of HMM algorithms is key to application in automatic speech recognition