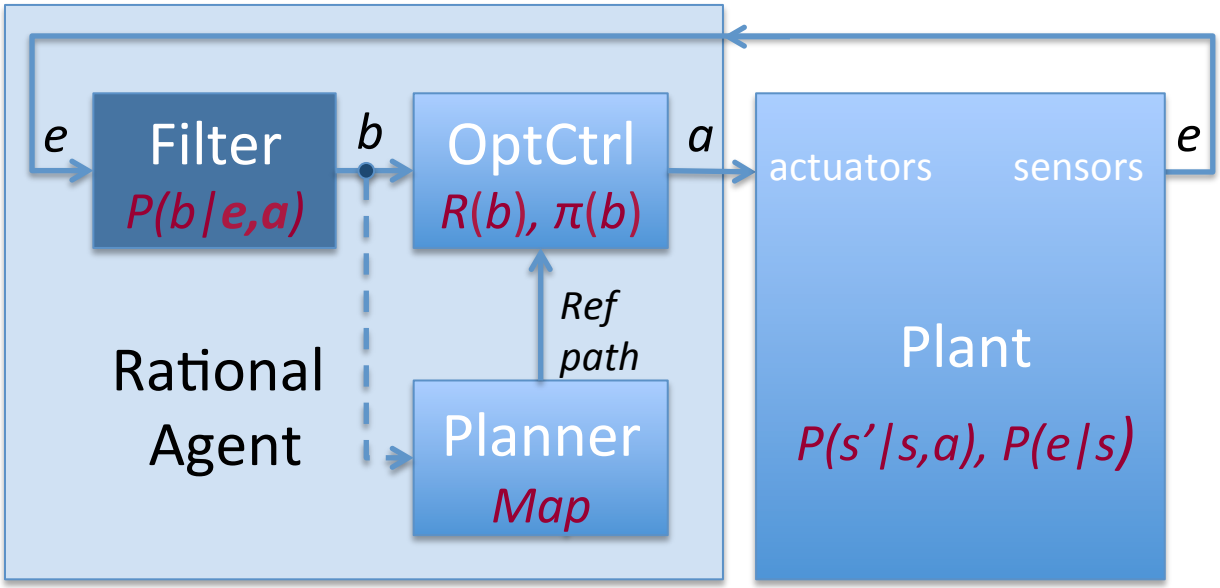
Speech Recognition

Chapter 23.5



Outline

- Speech as probabilistic inference
- Speech sounds
- Word pronunciation
- Word sequences

Challenges in Speech Recognition

Speech signals are noisy, variable, ambiguous

- Example: "recognize speech" and "wreck a nice beach"

Issues making speech problematic

- Segmentation: written words have spaces between them
- Not in fast speach: "wreck a nice" and "recognize"
- Coarticulation: sounds of successive words merge
- Sound "s" from "nice" merges "b" from "beach" $\approx$ "sp"
- Homophones: words with different meaning sound the same
- Words "to", "too", and "two" differ in meaning but sound the same

Speech as Probabilistic Inference

Since mid 1970s

- Formulated as a probabilistic inference problem

Most likely word sequence, given the speech signal?

- Choose Words to maximize P(Words | signal)
- $\underset{word_{1:t}}{\text{argmax}}\, P(word_{1:t} \mid sound_{1:t}) = \underset{word_{1:t}}{\text{argmax}}\, P(sound \mid word_{1:t})\, P(word_{1:t})$

acoustic model

language model

Approach: Use Bayes' rule

- $P(Words \mid signal) = \alpha P(signal \mid Words) P(Words)$
- Decompose into: Acoustic model + Language model
- Words: Are the hidden state sequence
- Signal: Is the observation sequence

Noisy channel model: By Claude Shannon (1948)

- Original msg: words, corrupted msg: sounds, noisy cnl: telephone line
- Applied to: Speech recogn, machine transl, spelling correction, etc

Speech Sounds

Raw signal: Microphone displacement over time

- Processed into overlapping 30ms frames
- Each described by features

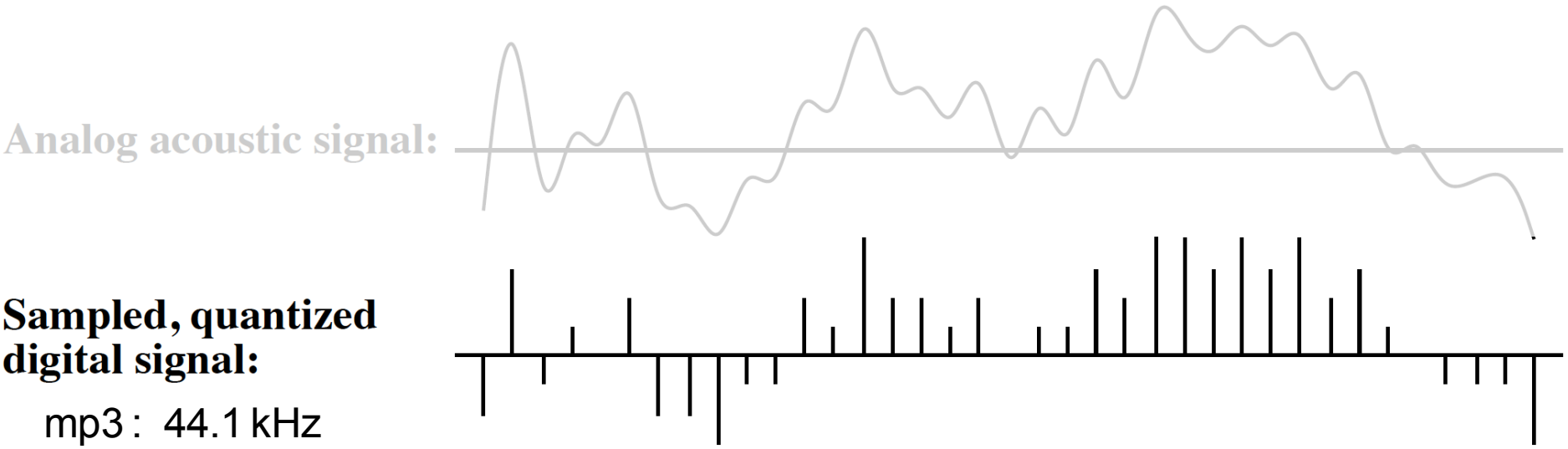
mp3 : 44.1 kHz

MFCC for each ν

Frame features: here three acoustic features per frame

Speech Sounds

Raw signal: Microphone displacement over time



mp3 : 44.1 kHz

MFCC for each ν

Frame features formants: peaks in the power spectrum

- MFCC: mel frequency cepstral coefficient for each υ + total energy (13)
- Diff between: this and previous frame, diff between diff (39 features)

Phones

All human speech is composed from 40-50 phones

- Determined by conguration of articulators
- Articulators: Lips, teeth, tongue, vocal cords, air flow

Form intermediate level of HS between words and signal

- Acoustic model = pronunciation model + phone model

ARPAbet designed for American English

[iy]	beat	[b]	bet	[p]	pet
[ih]	bit	[ch]	Chet	[r]	rat
[ey]	bet	[d]	debt	[s]	set
[ao]	bought	[hh]	hat	[th]	thick
[ow]	boat	[hv]	high	[dh]	that
[er]	Bert	[l]	let	[w]	wet
[ix]	roses	[ng]	sing	[en]	button
⋮	⋮	⋮	⋮	⋮	⋮

- Example: "ceiling" is [s iy l ih ng] / [s iy l ix ng] / [s iy l en]

Phones Models

Frame features in P(features | phone) summarized by:

- Vector quantization: an integer in [0 : 255]
- Mixture of Gaussians: associated parameters

3-state phones having three phases (Onset, Mid, End)

- Example: [t] has silent Onset, explosive Mid, hissing End
- P(features | phone, phase)

Triphone context: each phone becomes n^2 distinct phones

- Depending on the phones to its left and right
- Example: [t] in "star" is written [t(s,aa)] (different from "tar"!)

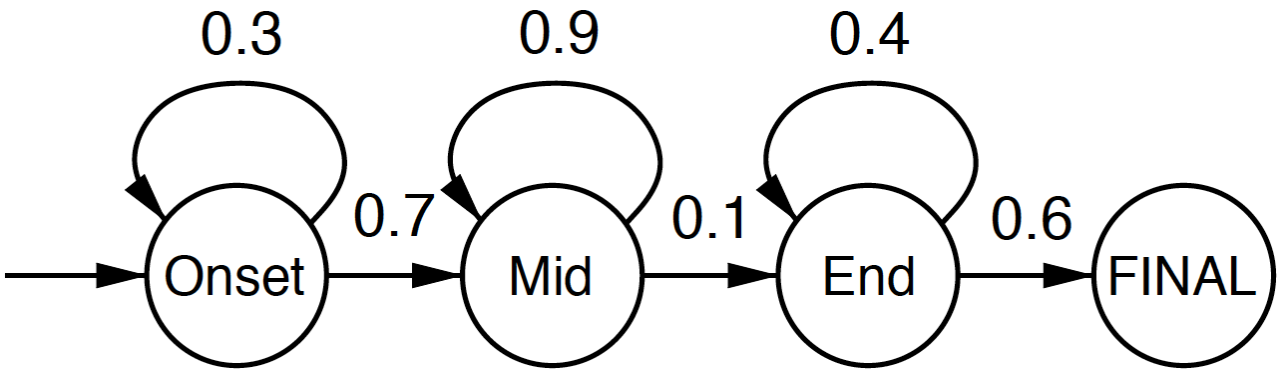
Triphones useful for handling coarticulation effects

- Inertia: articulators cannot switch instantaneously
- Example: [t] in "eighth" has tongue against front teeth

Phone Model Example

Phone HMM for [m]:

- Duration: normal speech 5-100 ms, 5-10 frames (self loops)



Output probabilities for the phone HMM (MFCC features):

- $C_1 - C_7$: Some arbitrary combination of feature values

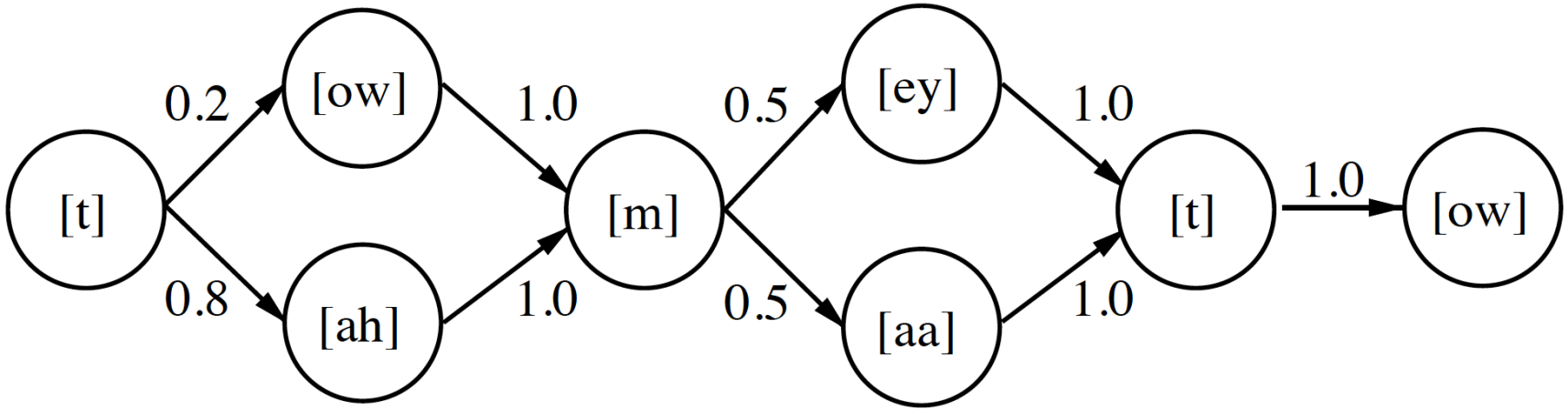
Onset:	Mid:	End:
C1: 0.5	C3: 0.2	C4: 0.1
C2: 0.2	C4: 0.7	C6: 0.5
C3: 0.3	C5: 0.1	C7: 0.4

Word-Pronunciation Models

Words described as distributions over phone sequences

Distribution represented as an HMM transition model

- With coarticulation: [t] tong at the top of the mouth and [ow] at bottom
- With dialect variation: [t] followed by [ow] or [ah], [m] by [ey] or [aa]



P([towmeytow] | "tomato") = P([towmaatow] | "tomato") = 0.1

P([tahmeytow] | "tomato") = P([tahmaatow] | "tomato") = 0.4

Structure created manually, transition probabilities learned from data

Isolated Words

Phone models + word models

- Fix likelihood $P(e_{1:t} \mid \text{word})$ for isolated word
- $P(\text{word} \mid e_{1:t}) = \alpha P(e_{1:t} \mid \text{word}) P(\text{word})$

Prior probability P(word) obtained by counting word frequencies

$P(e_{1:t} \mid \text{word})$ can be computed recursively

- Define: $l_{1:t} = P(X_t, e_{1:t})$
- Use the recursive update: $l_{1:t+1} = \text{Forward}(l_{1:t}, e_{t+1})$
- Finally: $P(e_{1:t} \mid \text{word}) = \sum_{x_t} l_{1:t}(x_t)$

Trained isolated-word dictation systems 95-99% accuracy

Continuous Speech Systems (CSS)

Not just a sequence of isolated-word recognitions!

- Adjacent words highly correlated
- Sequence of most likely words $\neq$ most likely sequence of words
- Segmentation: there are few gaps in speech
- Cross-word coarticulation: for example "next thing"

CSS manage 60-80% accuracy on a good day

Language Model

Prior probability of a word sequence is given by chain rule:

- $P(w_1,...,w_n) = \prod_{i=1}^{n} P(w_i \mid w_1,...,w_{i-1})$

Bigram model:

- $P(w_i \mid w_1,...,w_{i-1}) \approx P(w_i \mid w_{i-1})$

Train by counting all word pairs in a large text corpus

More sophisticated models (3grams, grammars,...) help a bit

Combined HMM

Combined language+word+phone model

- State labels: the word we're in +
  the phone in that word +
  the phone state in that phone

The most likely phone-state sequence

- Found by using the Viterbi algorithm

Does segmentation by

- Considering all possible word sequences and boundaries

Doesn't always give the most likely word sequence because

- Each word sequence is the sum over many state sequences

Jelinek invented A* in '69 to find most likely word sequence

- Here "step cost" is $\log P(w_i \mid w_{i-1})$

DBNs and Speech Recognition

Speech model = acoustic model + pronunciation model, too

- Acoustic model includes articulatory-context variables
- Capture the sate of the articulatory apparatus of the speaker
- Depend on current phonetic state and previous articulatory context

$$\text{Hidden} = \text{Phones} \cup \text{Articulation}$$

$$P(\text{Sound}, \text{Hidden} \mid \text{Words}) = P(\text{Sound}, \text{Phones}, \text{Articulation} \mid \text{Words})$$

$$= P(\text{Phones} \mid \text{Words})\, P(\text{Sound}, \text{Articulation} \mid \text{Phones})$$

Bayesian structure consists of two layers

- One that models $P(\text{Phones} \mid \text{Words})$
- One that models $P(\text{Sound}, \text{Articulation} \mid \text{Phones})$

DBN: Pronunciation Model

Assumptions

- Each word liear sequence of phonetic units "cat" [k ae t]
- Average duration of phones given by $t_{q_q} : q_1 \rightarrow q_2$
- Depend on current phonetic state and previous articulatory context

Index nodes

- Keeps track of the position in the phonetic transcription
- All words go through same sequence 1,2,...,k
- Assignment of values specifies a time alignment
- Deterministic map from index to actual phonetic value
- Value 1 index value is increased by one

$$P(\text{EOW} = 1 \mid \text{index} = \text{last}, \text{transition} = 1) = 1$$

DBN: Acoustic Model

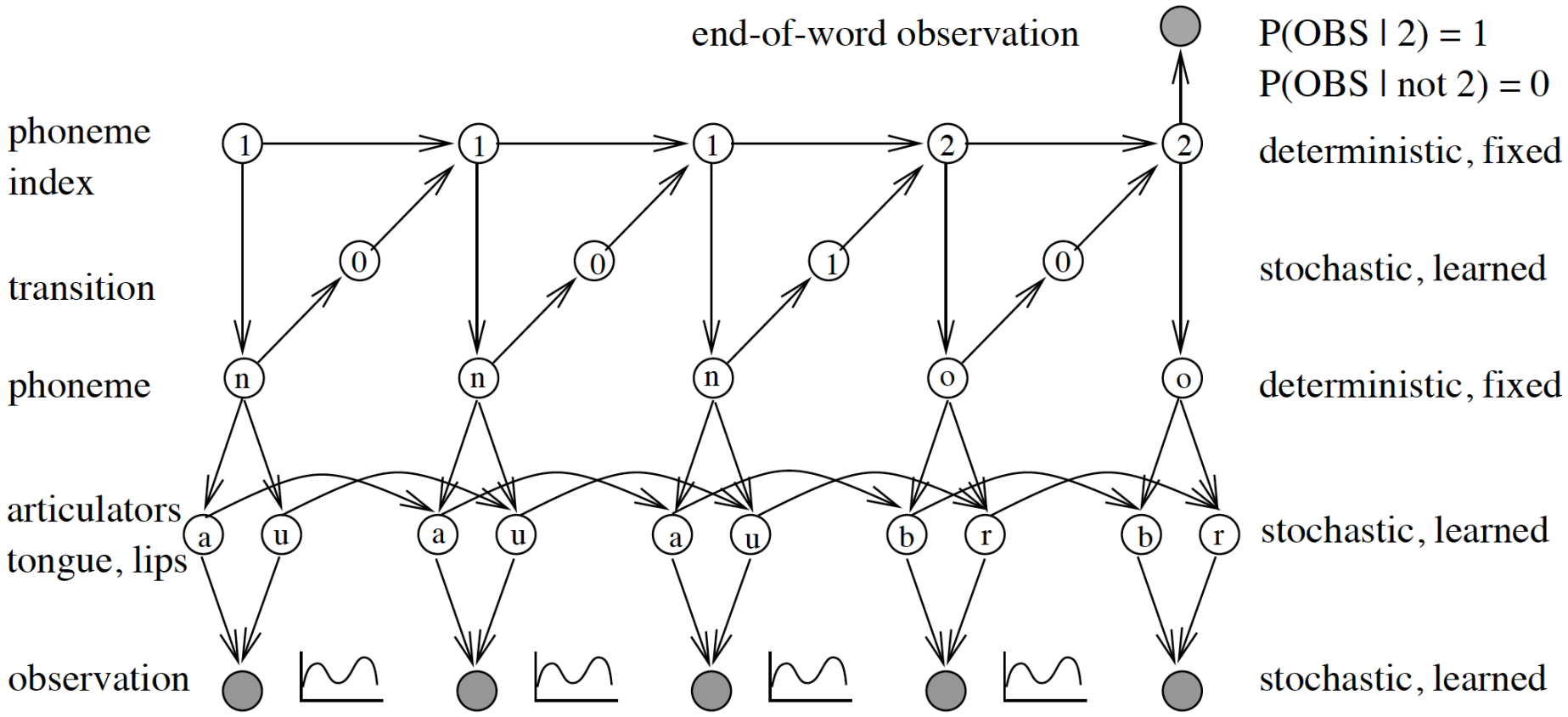
A DBN allows the hidden state to be factored

- Augment phonetic state vars with articulatory context vars

Context variable serves two purposes

- Long-term correlations among observations accross time frames
- Short-term correlations among observations within one time frame

DBNs for Speech Recognition



Also easy to add variables for, e.g., gender, accent, speed

Zweig,Russell (1998) show up to 40% error reduction over HMMs

Summary

Speech recognition (since mid 1970)

- Has been formulated as probabilistic inference

Evidence:

- Speech signal, hidden variables = word and phone sequences

Context effects (coarticulation etc.)

- Handled by augmenting state

Variability in human speech

- Speed, timbre, etc., and background noise
- Make continuous speech recognition in real settings an open problem

Want to know more?

- L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition