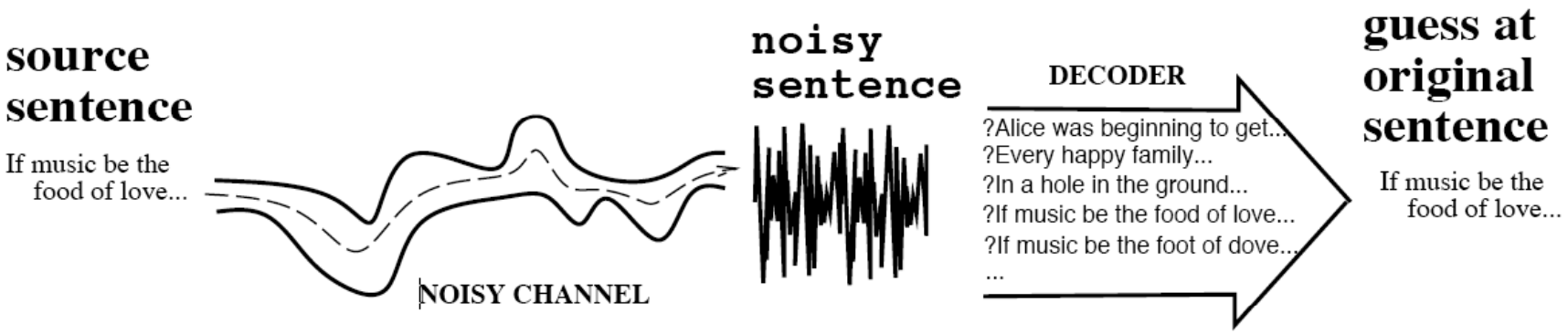
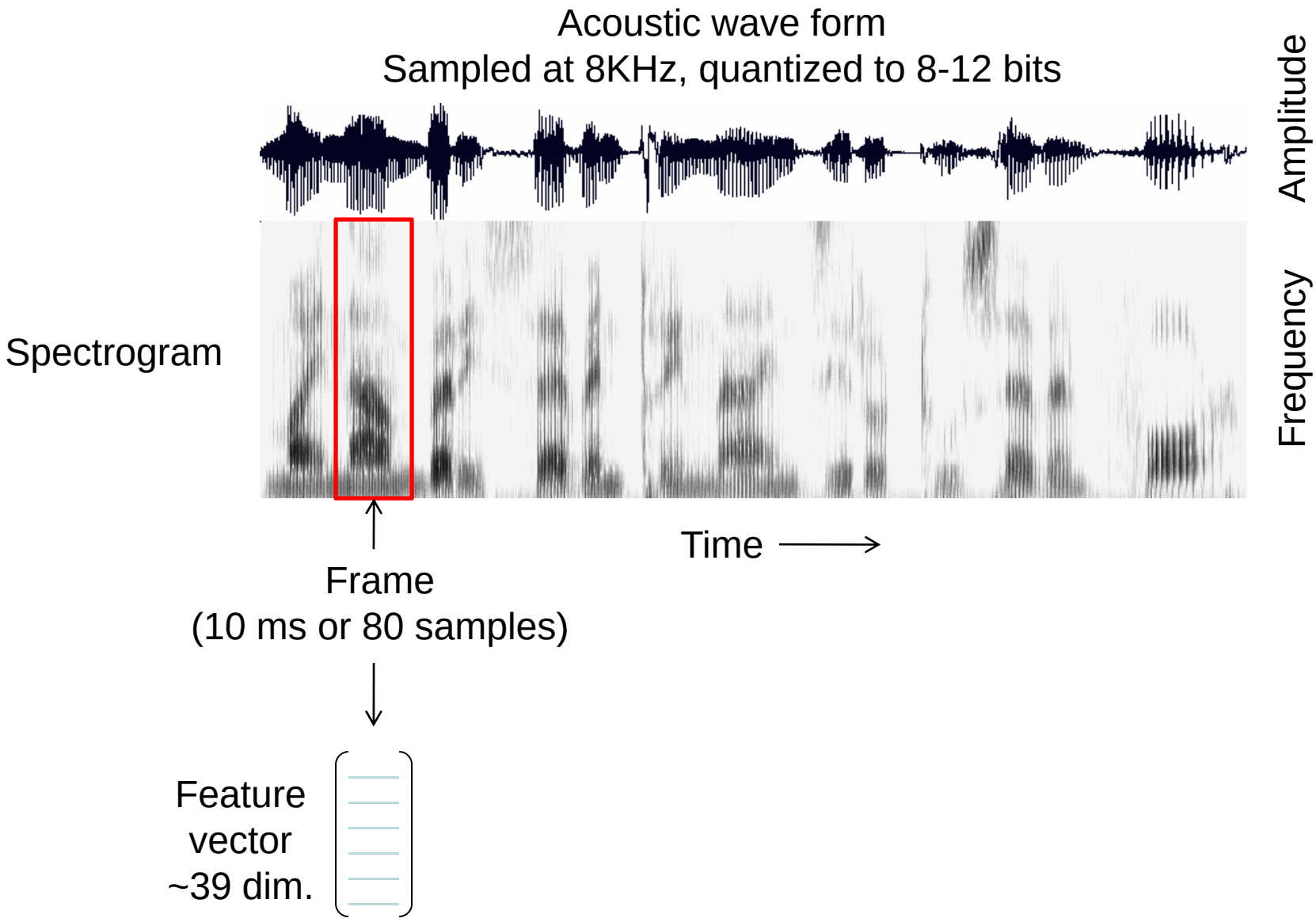
Application of HMMs: Speech recognition

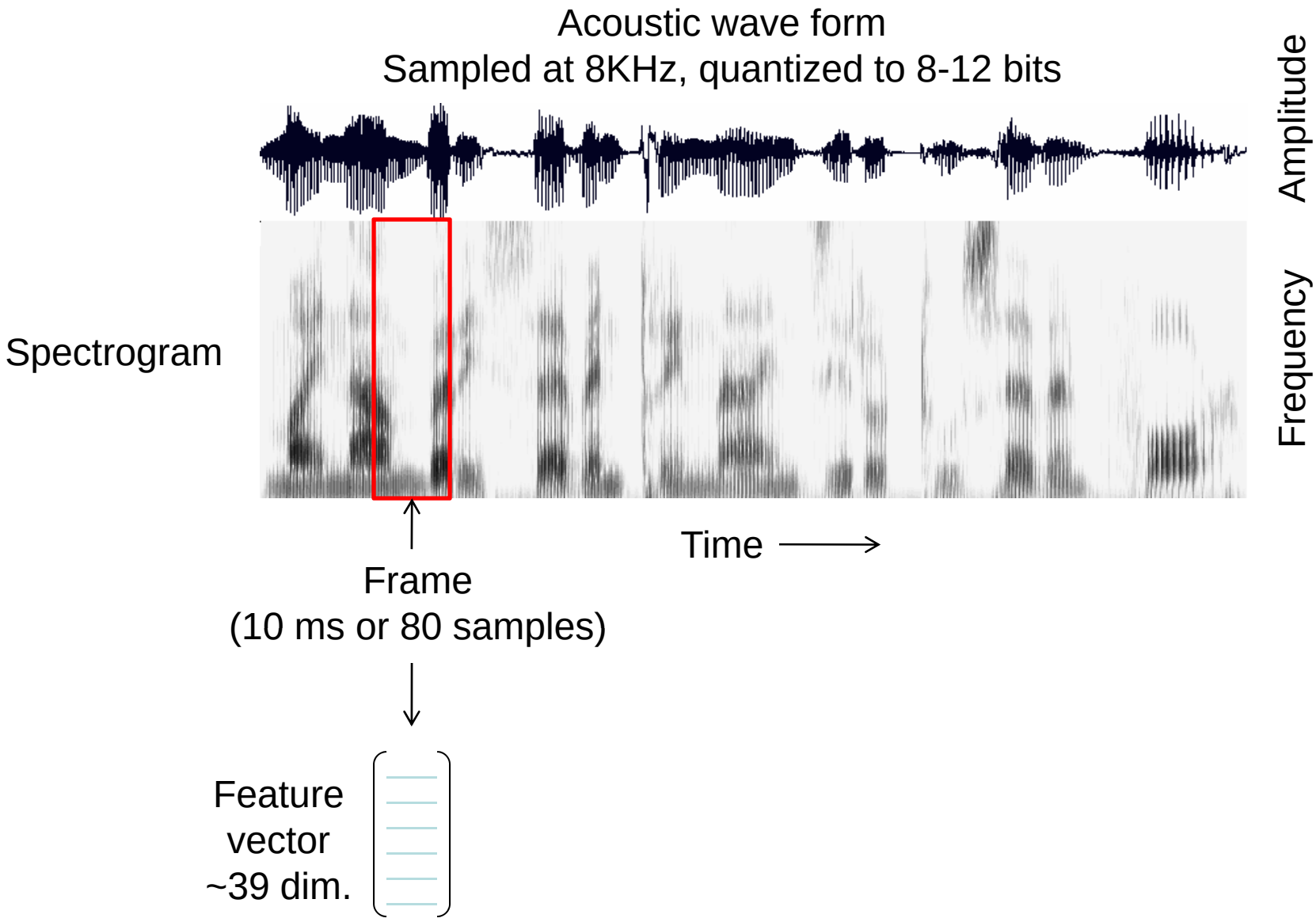
- "Noisy channel" model of speech



Speech feature extraction



Speech feature extraction



Phonetic model

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[p]	[p]	parsley	['parsli]	[p aa r s l iy]
[t]	[t]	tarragon	['tærəgɑn]	[t ae r ax g aa n]
[k]	[k]	catnip	['kætnɪp]	[k ae t n ix p]
[b]	[b]	bay	[beɪ]	[b ey]
[d]	[d]	dill	[dɪl]	[d ih l]
[g]	[g]	garlic	['gɑrlɨk]	[g aa r l ix k]
[m]	[m]	mint	[mɪnt]	[m ih n t]
[n]	[n]	nutmeg	['nʌtmɛg]	[n ah t m eh g
[ŋ]	[ng]	ginseng	['dʒɪnsɨŋ]	[jh ih n s ix ng]
[f]	[f]	fennel	['fɛnl]	[f eh n el]
[v]	[v]	clove	[kloʊv]	[k l ow v]
[θ]	[th]	thistle	['θɪsl]	[th ih s el]
[ð]	[dh]	heather	['hɛðɚ]	[h eh dh axr]
[s]	[s]	sage	[seɪdʒ]	[s ey jh]
[z]	[z]	hazelnut	['hɛɪzl̩nʌt]	[h ey z el n ah t]
[ʃ]	[sh]	squash	[skwɑʃ]	[s k w a sh]
[ʒ]	[zh]	ambrosia	[æm'broʊʒə]	[ae m b r ow zh ax]
[tʃ]	[ch]	chicory	['tʃɪkɚi]	[ch ih k axr iy]
[dʒ]	[jh]	sage	[seɪdʒ]	[s ey jh]
[l]	[l]	licorice	['lɪkɚɨʃ]	[l ih k axr ix sh]
[w]	[w]	kiwi	['kiwi]	[k iy w iy]
[r]	[r]	parsley	['pɑrsli]	[p aa r s l iy]
[j]	[y]	yew	[yu]	[y uw]
[h]	[h]	horseradish	['hɔrsrædɪʃ]	[h ao r s r ae d ih sh]
[ʔ]	[q]	uh-oh	[ʔʌʔoʊ]	[q ah q ow]
[ɾ]	[dx]	butter	['bʌɾɚ]	[b ah dx axr]
[ɾ̃]	[nx]	wintergreen	[wɪɾ̃ɚgrin]	[w ih nx axr g r i n]
[l̩]	[el]	thistle	['θɪsl̩]	[th ih s el]

Figure 4.1 IPA and ARPAbet symbols for transcription of English consonants.

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[i]	[iy]	lily	['lɪli]	[l ih l iy]
[ɪ]	[ih]	lily	['lɪli]	[l ih l iy]
[eɪ]	[ey]	daisy	['deɪzi]	[d ey z i]
[ɛ]	[eh]	poinsettia	[pɔɪn'sɛɾiə]	[p oy n s eh dx iy ax]
[æ]	[ae]	aster	['æstɚ]	[ae s t axr]
[ɑ]	[aa]	poppy	['papi]	[p aa p i]
[ɔ]	[ao]	orchid	['ɔrkɨd]	[ao r k ix d]
[ʊ]	[uh]	woodruff	['wʊdrʌf]	[w uh d r ah f]
[oʊ]	[ow]	lotus	['loʊɾəs]	[l ow dx ax s]
[u]	[uw]	tulip	['tuɫɨp]	[t uw l ix p]
[ʌ]	[uh]	buttercup	['bʌɾɚ,kʌp]	[b uh dx axr k uh p]
[ɝ]	[er]	bird	[bɝd]	[b er d]
[aɪ]	[ay]	iris	['aɪɾɨs]	[ay r ix s]
[aʊ]	[aw]	sunflower	['sʌnflaʊɚ]	[s ah n f l aw axr]
[oɪ]	[oy]	poinsettia	[pɔɪn'sɛɾiə]	[p oy n s eh dx iy ax]
[ju]	[y uw]	feverfew	[fivɚfju]	[f iy v axr f y u]
[ə]	[ax]	woodruff	['wʊdrəf]	[w uh d r ax f]
[ɚ]	[axr]	heather	['hɛðɚ]	[h eh dh axr]
[ɨ]	[ix]	tulip	['tuɫɨp]	[t uw l ix p]
[ʉ]	[ux]		[]	[]

Figure 4.2 IPA and ARPAbet symbols for transcription of English vowels

Phonetic model

- **Phones:** speech sounds
- **Phonemes:** groups of speech sounds that have a unique meaning/function in a language (e.g., there are several different ways to pronounce “t”)

HMM models for phones

- HMM states in most speech recognition systems correspond to *subphones*
  - There are around 60 phones and as many as 60^3 context-dependent *triphones*

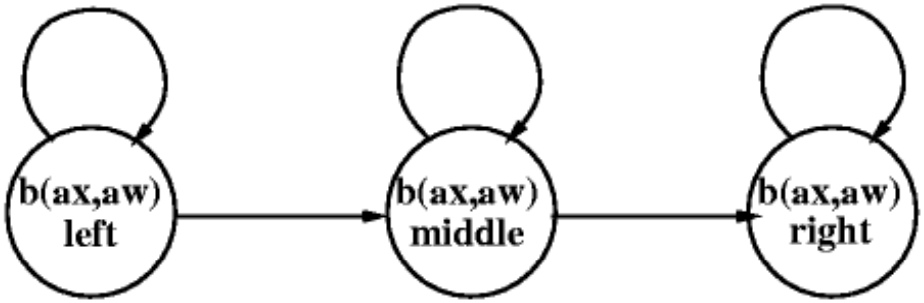


Figure 7.11 An example of the context-dependent triphone b(ax,aw) (the phone [b] preceded by a [ax] and followed by a [aw], as in the beginning of *about*, showing its left, middle, and right subphones.

HMM models for words

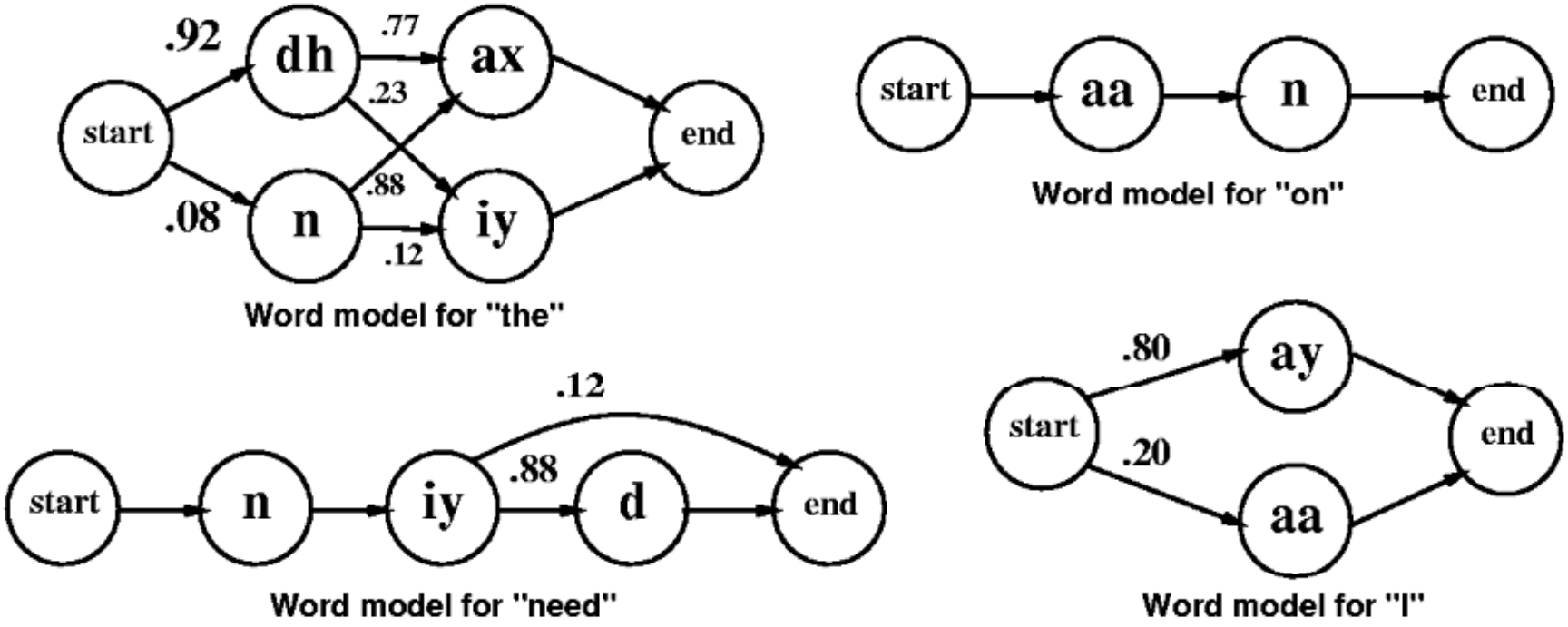
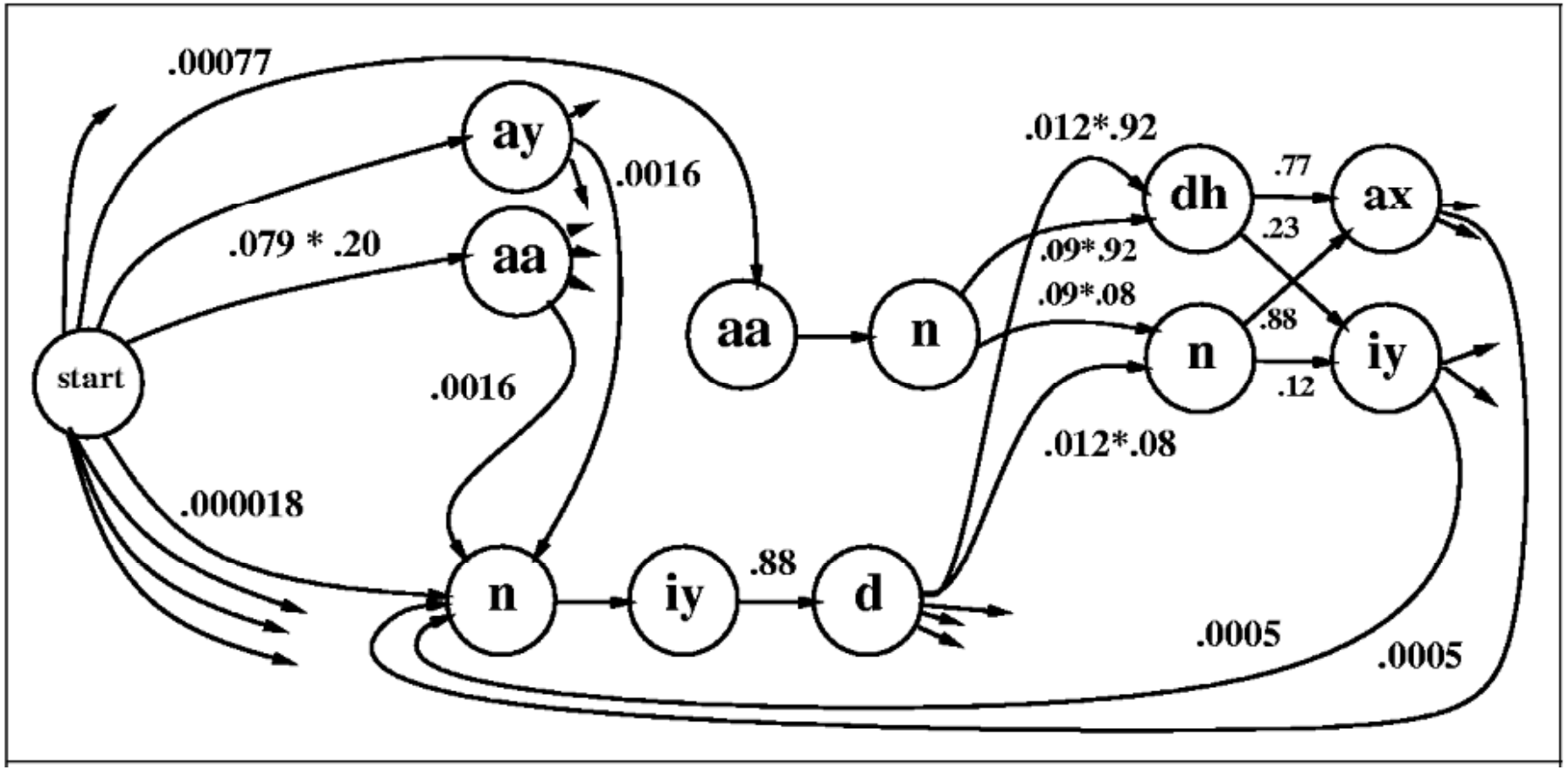


Figure 7.5 Pronunciation networks for the words *I*, *on*, *need*, and *the*. All networks (especially *the*) are significantly simplified.

Putting words together



- Given a sequence of acoustic features, how do we find the corresponding word sequence?

Decoding with the Viterbi algorithm

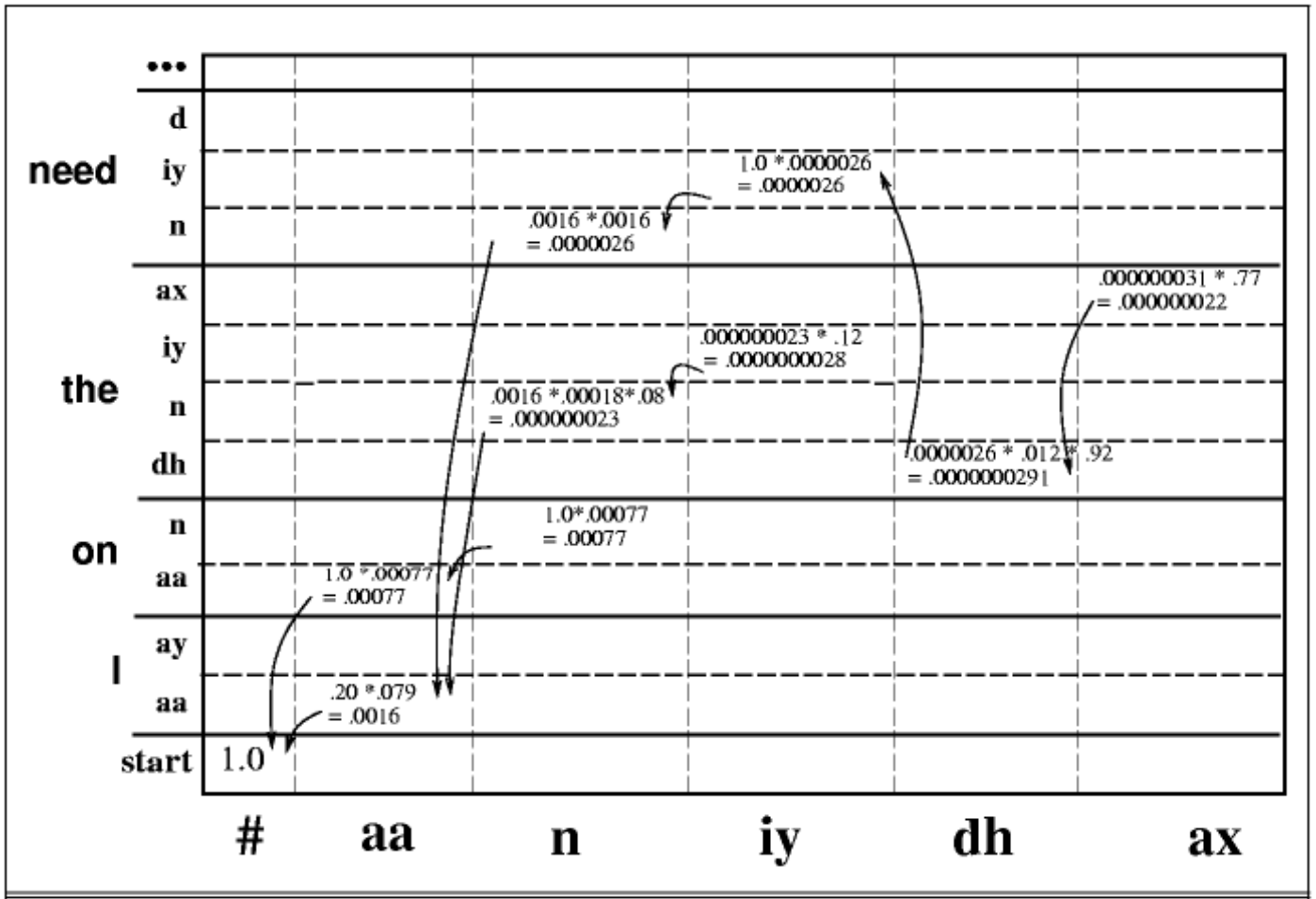
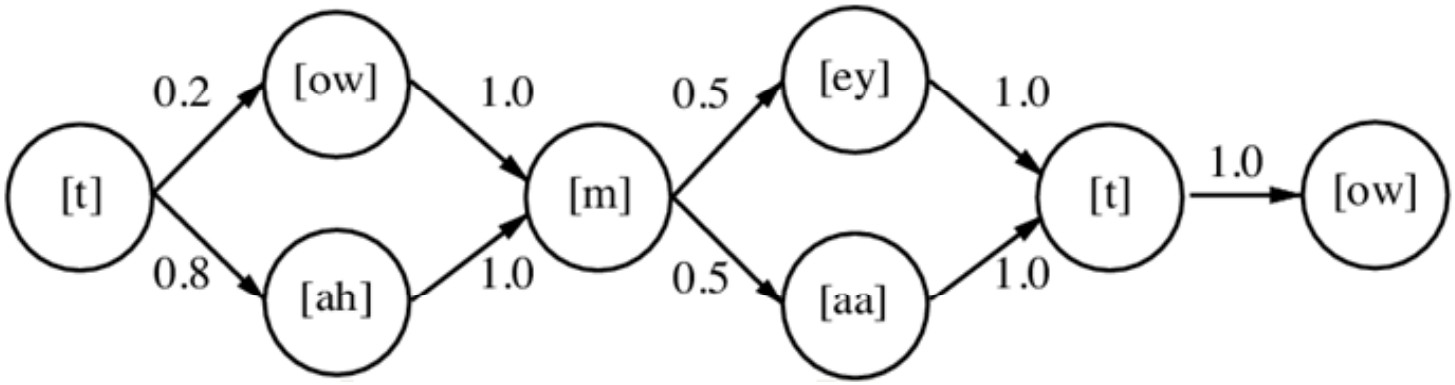


Figure 7.10 The entries in the individual state columns for the Viterbi algorithm. Each cell keeps the probability of the best path so far and a pointer to the previous cell along that path. Backtracing from the successful last word (*the*), we can reconstruct the word sequence *I need the*.

Limitations of Viterbi decoding

- Number of states may be too large
  - **Beam search:** at each time step, maintain a short list of the most probable words and only extend transitions from those words into the next time step
- Words with multiple pronunciation variants may get a smaller probability than incorrect words with fewer pronunciation paths



Word model for “tomato”

Limitations of Viterbi decoding

- Number of states may be too large
  - **Beam search:** at each time step, maintain a short list of the most probable words and only extend transitions from those words into the next time step
- Words with multiple pronunciation variants may get a smaller probability than incorrect words with fewer pronunciation paths
  - Use the forward algorithm instead of Viterbi algorithm
- The Markov assumption is too weak to capture the constraints of real language

Advanced techniques

- Multiple pass decoding
  - Let the Viterbi decoder return multiple candidate utterances and then re-rank them using a more sophisticated language model, e.g., *n-gram model*

Figure 7.12 The use of *N*-best decoding as part of a two-stage decoding model. Efficient but unsophisticated knowledge sources are used to return the *N*-best utterances. This significantly reduces the search space for the second pass models, which are thus free to be very sophisticated but slow.

Advanced techniques

- Multiple pass decoding
  - Let the Viterbi decoder return multiple candidate utterances and then re-rank them using a more sophisticated language model, e.g., *n-gram model*
- A* decoding
  - Build a search tree whose nodes are words and whose paths are possible utterances
  - Path cost is given by the likelihood of the acoustic features given the words inferred so far
  - Heuristic function estimates the best-scoring extension until the end of the utterance

Reference

- D. Jurafsky and J. Martin, “Speech and Language Processing,” 2nd ed., Prentice Hall, 2008