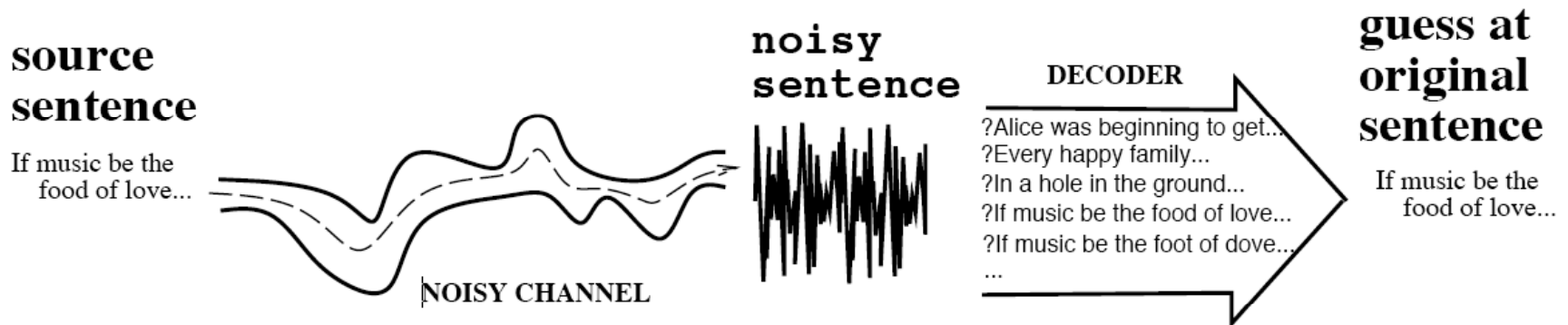
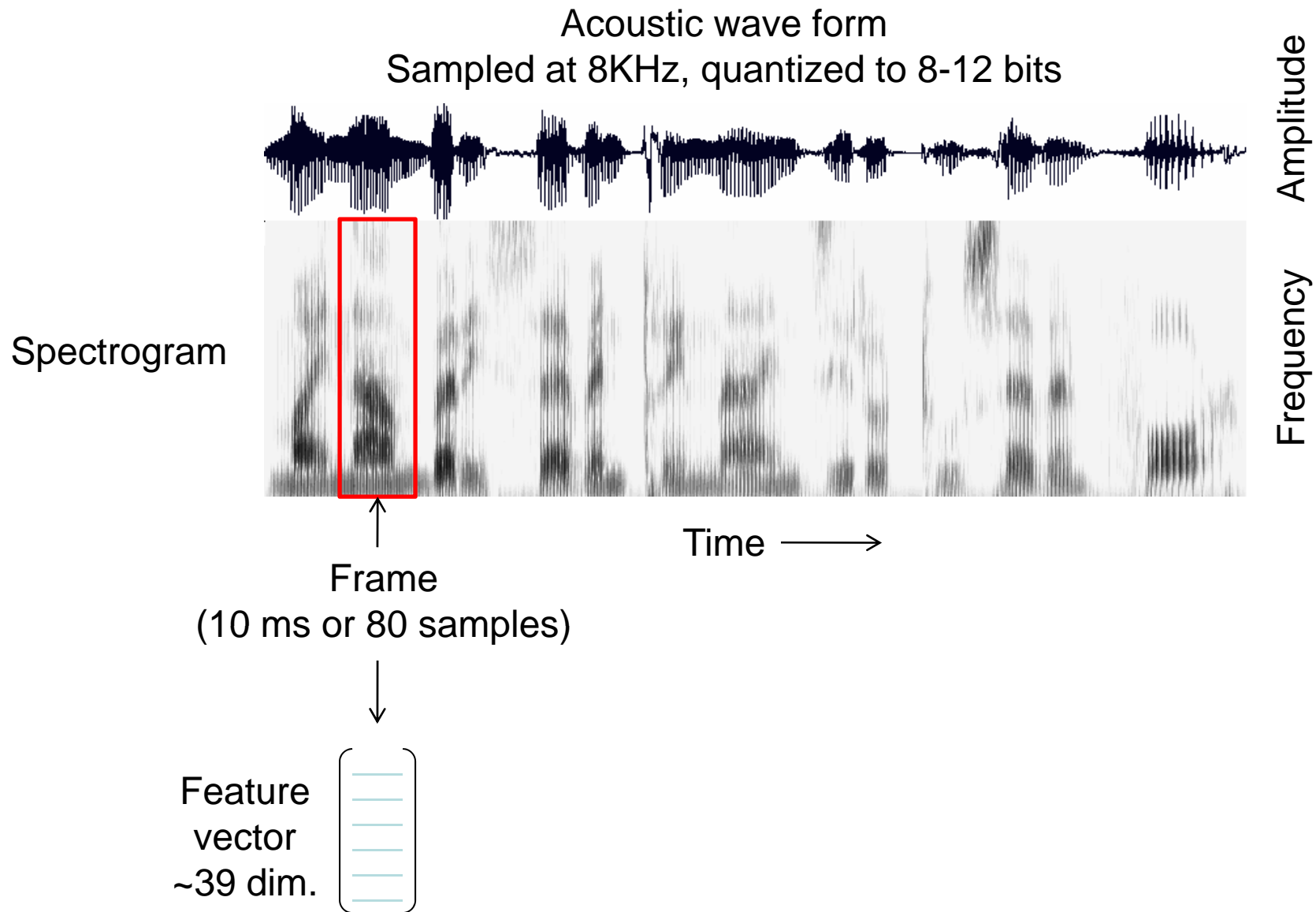


Application of HMMs: Speech recognition

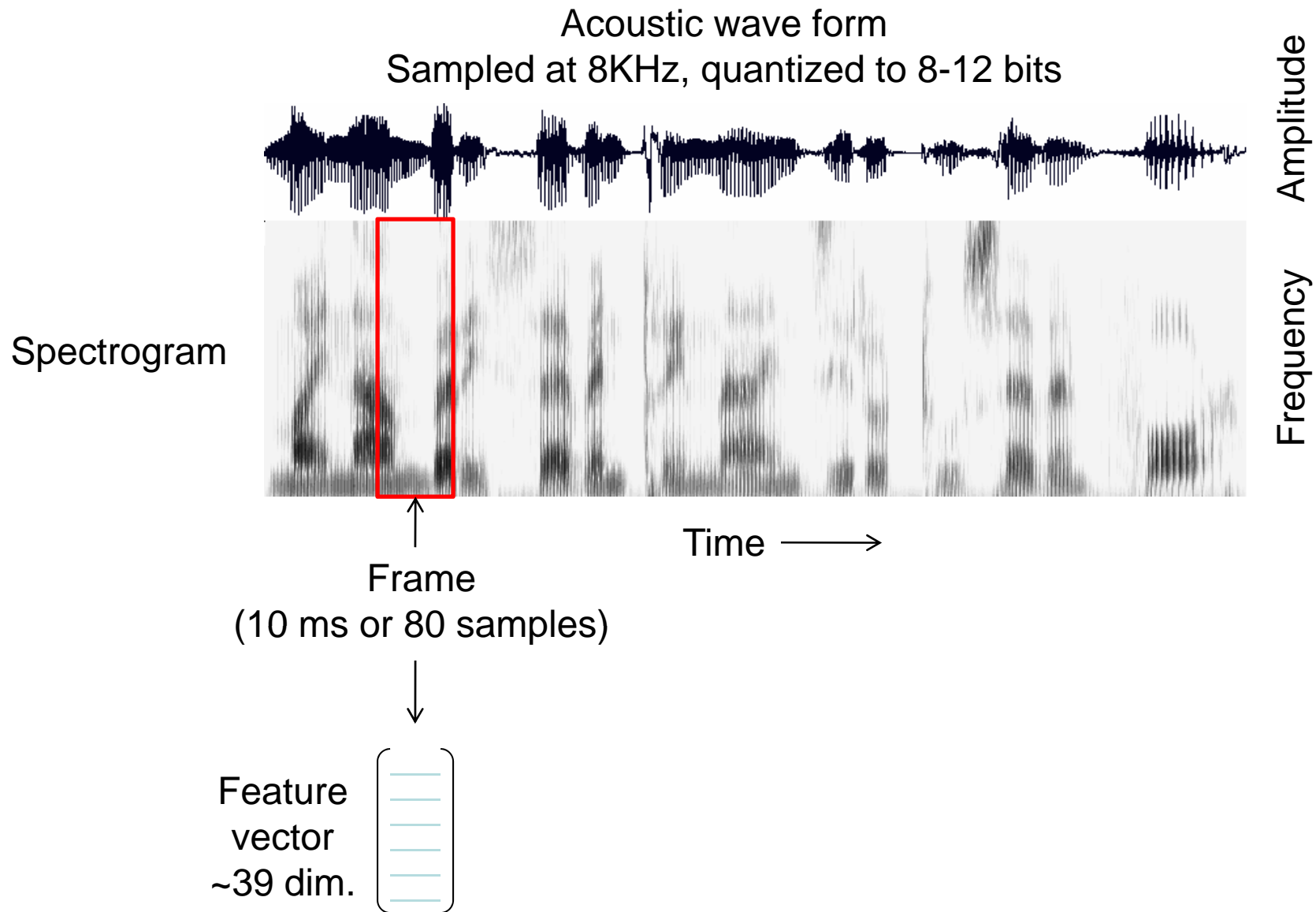
- “Noisy channel” model of speech



Speech feature extraction



Speech feature extraction



Phonetic model

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[p]	[p]	<u>p</u> arsley	[ˈpɑrsli]	[p aa r s l iy]
[t]	[t]	<u>t</u> arragon	[ˈtærəɡən]	[t ae r ax g aa n]
[k]	[k]	<u>c</u> atnip	[ˈkætnip]	[k ae t n ix p]
[b]	[b]	<u>b</u> ay	[ber]	[b ey]
[d]	[d]	<u>d</u> ill	[dɪl]	[d ih l]
[g]	[g]	<u>g</u> arlic	[ˈɡɑrlɪk]	[g aa r l ix k]
[m]	[m]	<u>m</u> int	[mɪnt]	[m ih n t]
[n]	[n]	<u>n</u> utmeg	[ˈnʌtmæg]	[n ah t m eh g]
[ŋ]	[ng]	<u>g</u> inseng	[ˈdʒɪnsɪŋ]	[jh ih n s ix ng]
[f]	[f]	<u>f</u> ennel	[ˈfenl]	[f eh n el]
[v]	[v]	<u>c</u> love	[kloʊv]	[k l ow v]
[θ]	[th]	<u>t</u> histle	[ˈθɪsl]	[th ih s el]
[ð]	[dh]	<u>h</u> eather	[ˈhedðə]	[h eh dh axr]
[s]	[s]	<u>s</u> age	[seɪdʒ]	[s ey jh]
[z]	[z]	<u>h</u> azelnut	[ˈheɪzlnʌt]	[h ey z el n ah t]
[ʃ]	[sh]	<u>s</u> quash	[skwɔʃ]	[s k w a sh]
[ʒ]	[zh]	<u>a</u> mbrosia	[æmˈbrʊʒə]	[ae m b r ow zh ax]
[tʃ]	[ch]	<u>c</u> hicory	[ˈtʃɪkəri]	[ch ih k axr iy]
[dʒ]	[jh]	<u>s</u> age	[seɪdʒ]	[s ey jh]
[l]	[l]	<u>l</u> icorice	[ˈlɪkəriʃ]	[l ih k axr ix sh]
[w]	[w]	<u>k</u> iwi	[ˈkiwi]	[k iy w iy]
[r]	[r]	<u>p</u> arsley	[ˈpɑrsli]	[p aa r s l iy]
[j]	[y]	<u>y</u> ew	[ju]	[y uw]
[h]	[h]	<u>h</u> orseradish	[ˈhɔrsrædɪʃ]	[h ao r s r ae d ih sh]
[ʔ]	[q]	uh-oh	[ʔʌʔou]	[q ah q ow]
[ɹ]	[dx]	<u>b</u> utter	[ˈbʌtə]	[b ah dx axr]
[ɹ̩]	[nx]	<u>w</u> intergreen	[wɪntəˈɡriːn]	[w ih nx axr g r i n]
[l]	[el]	<u>t</u> histle	[ˈθɪsl]	[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]	<u>l</u> ily	[ˈlɪli]	[l ih l iy]
[ɪ]	[ih]	<u>l</u> ily	[ˈlɪli]	[l ih l iy]
[er]	[ey]	<u>d</u> aisy	[ˈdeɪzi]	[d ey z i]
[ɛ]	[eh]	<u>p</u> oinsettia	[pɔɪnˈsetiə]	[p oy n s eh dx iy ax]
[æ]	[ae]	<u>a</u> ster	[ˈæstə]	[ae s t axr]
[ɑ]	[aa]	<u>p</u> oppy	[ˈpɑpi]	[p aa p i]
[ɔ]	[ao]	<u>o</u> rchid	[ˈɔrkɪd]	[ao r k ix d]
[u]	[uh]	<u>w</u> oodruff	[ˈwʊdrʌf]	[w uh d r ah f]
[ou]	[ow]	<u>l</u> otus	[ˈloʊəs]	[l ow dx ax s]
[u]	[uw]	<u>t</u> ulip	[ˈtulɪp]	[t uw l ix p]
[ʌ]	[uh]	<u>b</u> uttercup	[ˈbʌtəˌkʌp]	[b uh dx axr k uh p]
[ɜ]	[er]	<u>b</u> ird	[ˈbɜrd]	[b er d]
[aɪ]	[ay]	<u>i</u> ris	[ˈaɪrɪs]	[ay r ix s]
[aʊ]	[aw]	<u>s</u> unflower	[ˈsʌnflaʊə]	[s ah n f l aw axr]
[oɪ]	[oy]	<u>p</u> oinsettia	[pɔɪnˈsetiə]	[p oy n s eh dx iy ax]
[ju]	[y uw]	<u>f</u> everfew	[ˈfɪvəfju]	[f iy v axr f y u]
[ə]	[ax]	<u>w</u> oodruff	[ˈwʊdrʌf]	[w uh d r ax f]
[ð]	[axr]	<u>h</u> eather	[ˈhedðə]	[h eh dh axr]
[i]	[ix]	<u>t</u> ulip	[ˈtulɪp]	[t uw l ix p]
[ɪ]	[ux]		[]	[]

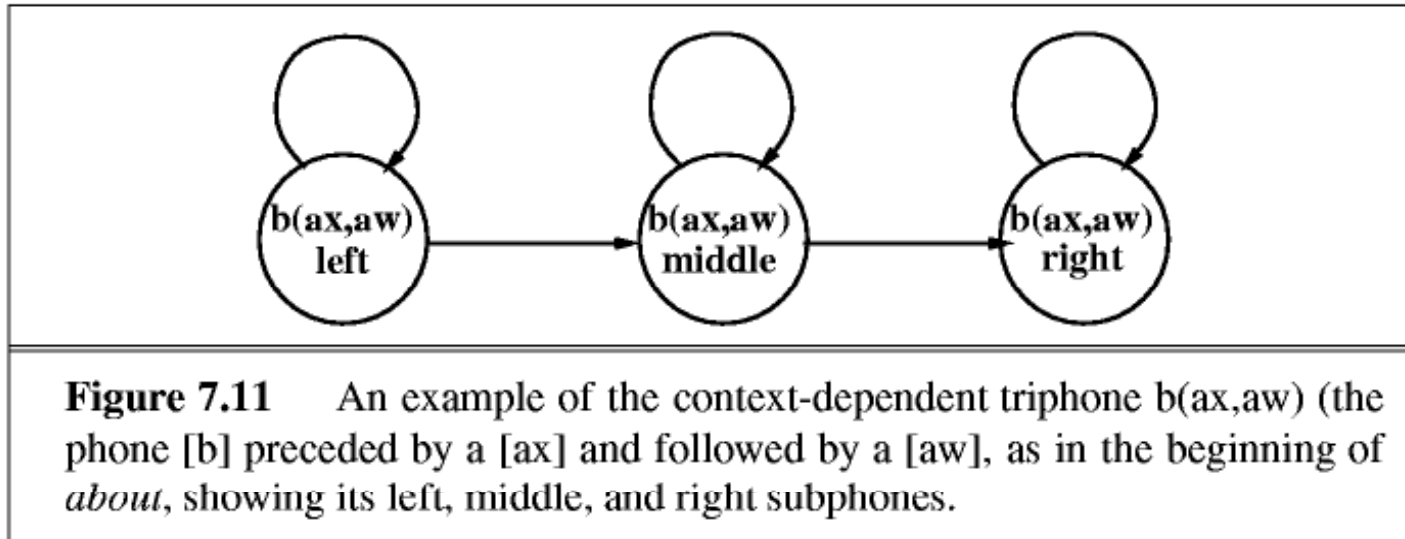
Figure 4.2 IPA and ARPAbet symbols for transcription of English vowels

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*



HMM models for words

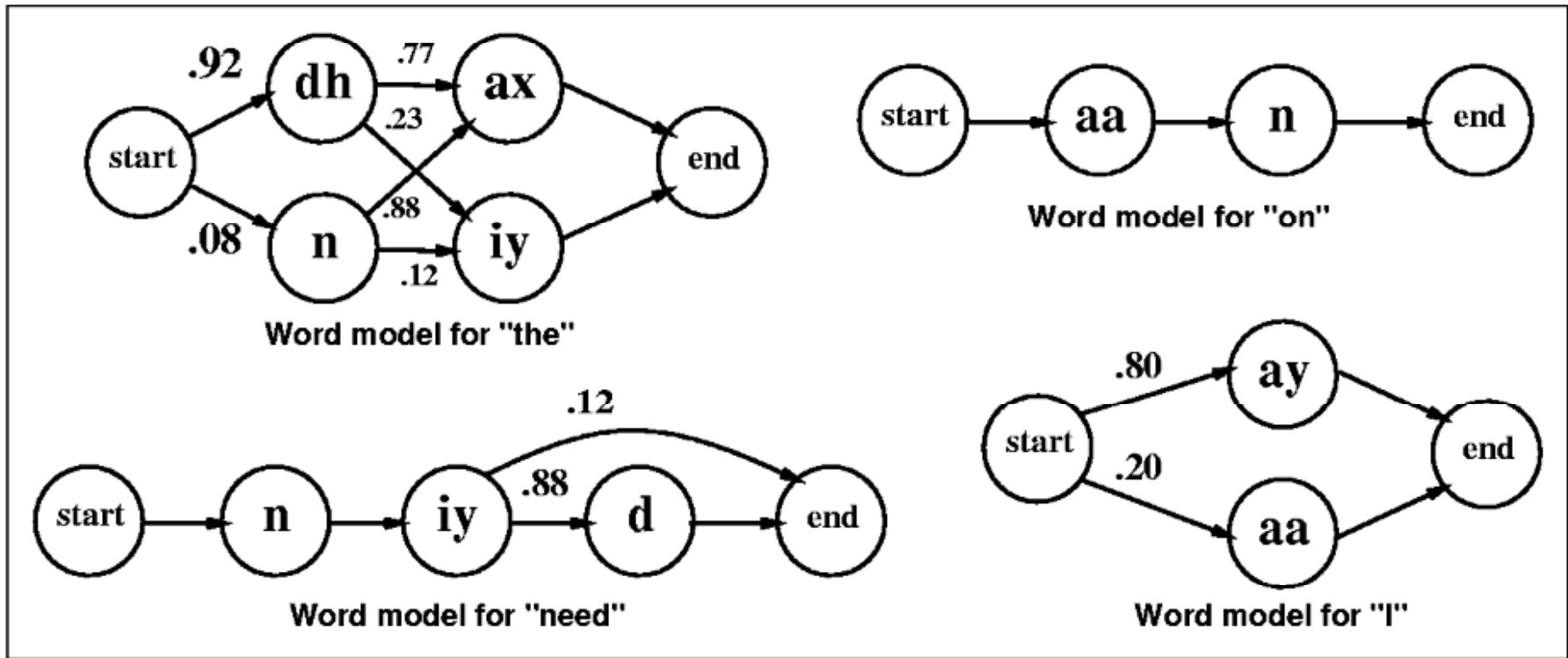
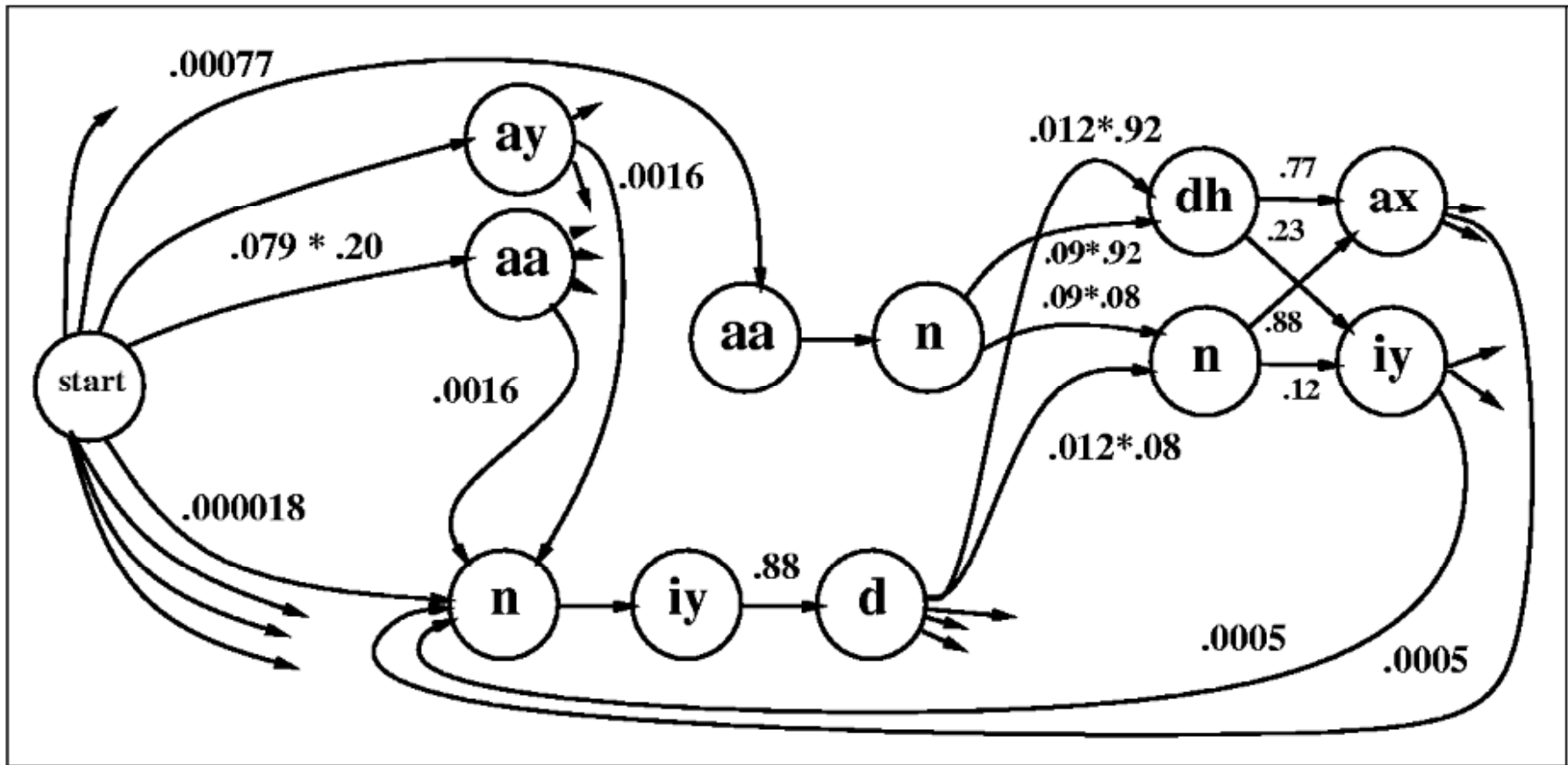


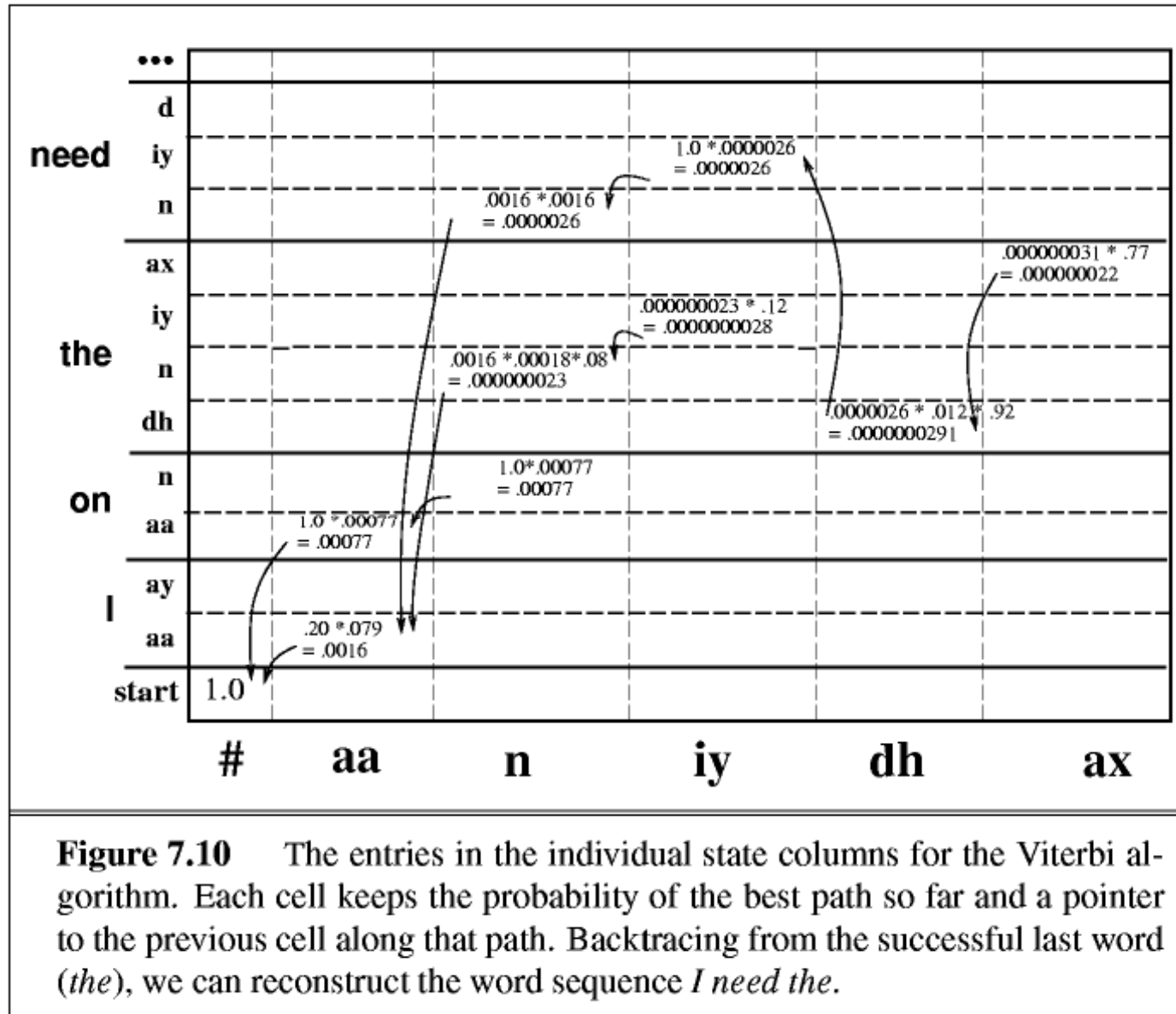
Figure 7.5 Pronunciation networks for the words *I*, *on*, *need*, and *the*. All networks (especially *the*) are significantly simplified.

Putting words together



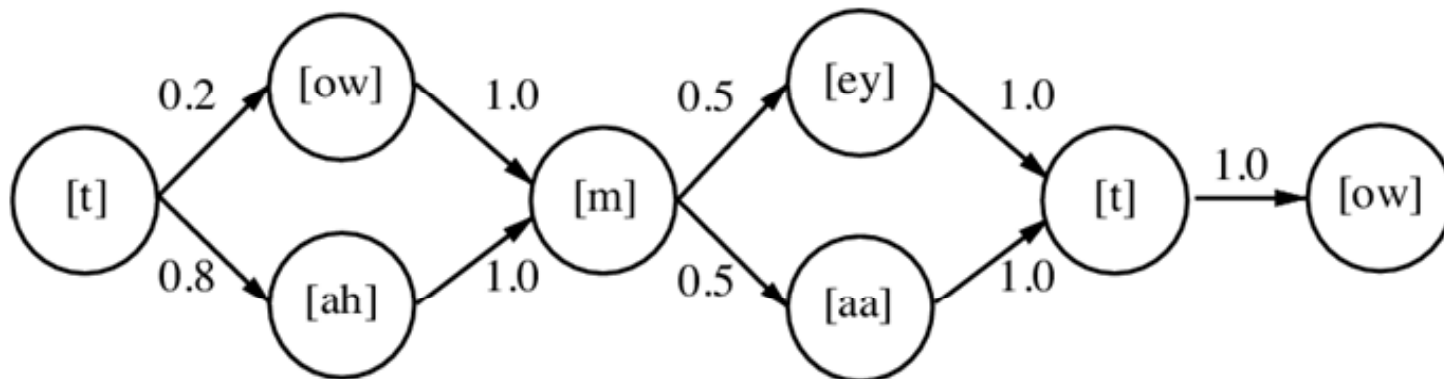
- Given a sequence of acoustic features, how do we find the corresponding word sequence?

Decoding with the Viterbi algorithm



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



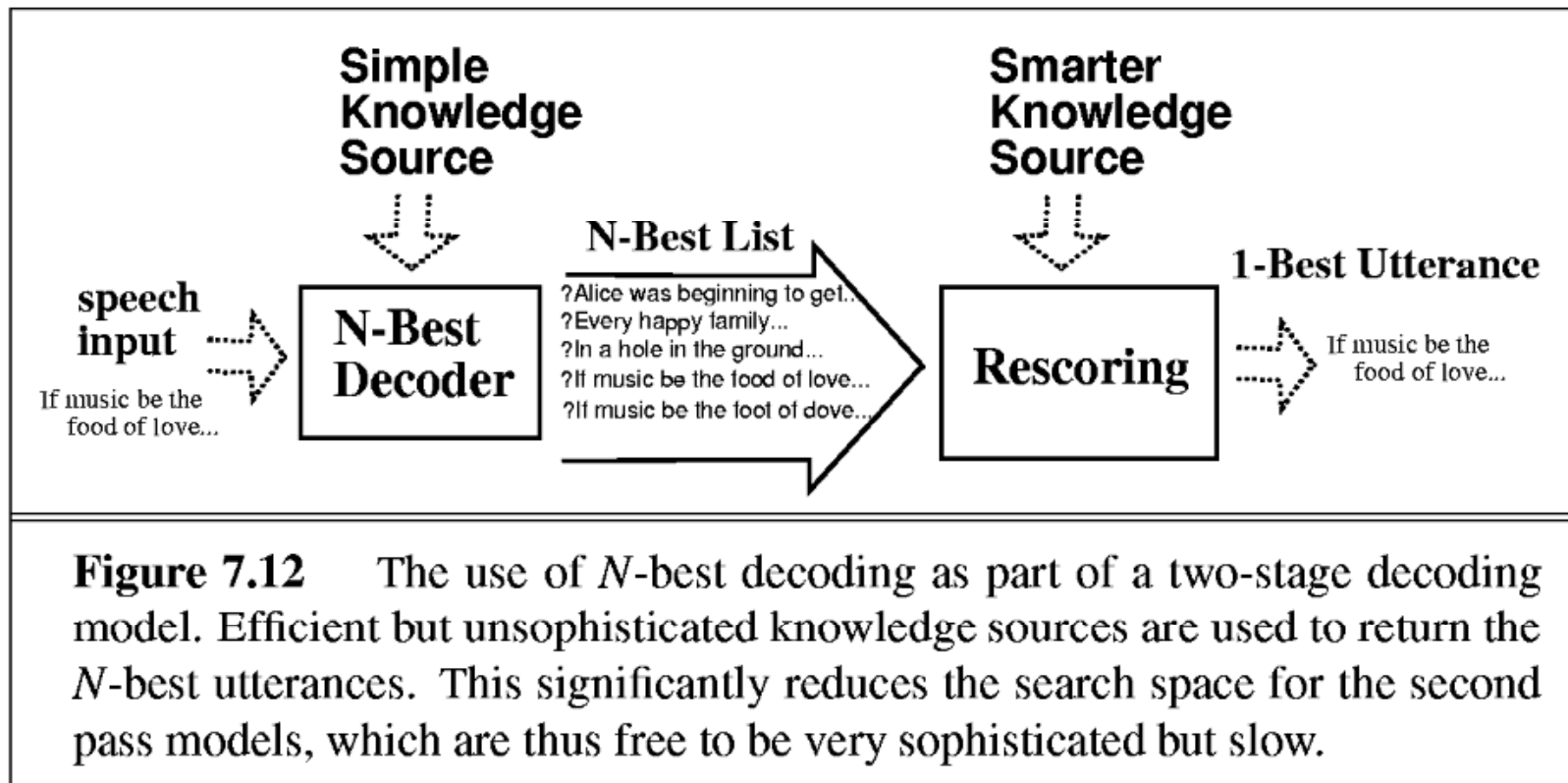
Word model for "tomato"

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*



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

