Speech Recognition

- **Speech recognition is the task of identifying a sequence of words uttered by a speaker, given** the acoustic signal.
- Speech recognition is difficult because the sounds made by a speaker are ***ambiguous***.
- Example: ”recognize Speech” sounds almost same as “wreck a nice beach”
- Several issues in speech recognition such as ***segmentation, coarticulation, homophones(to, two, too)***

$$\underset{word_{1:t}}{\operatorname{argmax}} P(word_{1:t} \mid sound_{1:t}) = \underset{word_{1:t}}{\operatorname{argmax}} P(sound_{1:t} \mid word_{1:t}) P(word_{1:t}) \, .$$

- P(sound1:t|word1:t) is the ***acoustic model***. P(word 1:t) is known as the ***language model***.
- This approach was named the noise channel model by Claude Shannon(1948).

quickly. Even this short example shows several of the issues that make speech problematic.

2. First, **segmentation**: written words in English have spaces between them, but in fast speech there are no pauses in “wreck a nice” that would distinguish it as a multiword phrase as opposed to the single word “recognize.”
3. Second, **coarticulation**: when speaking quickly the “s” sound at the end of “nice” merges with the “b” sound at the beginning of “beach,” yielding something that is close to a “sp.”Another problem that does not show up in this example is **homophones**—words like “to, ”too,” and “two” that sound the same but differ in meaning.

Acoustic Model

- Sound waves are periodic changes in pressure that propagate through the air.
- Approximates the amplitude of the sound wave—at discrete intervals called the ***sampling rate***.
- The precision of each measurement is determined by the ***quantization factor*** sampling at 8 kHz with 8-bit quantization.
- A ***phone*** is the sound that corresponds to a single vowel or consonant, but there are some complications:
- combinations of letters, such as “th” and “ng” produce single phones, and
- Some letters produce different phones in different contexts (e.g., the “a” in rat and rate)

Acoustic Model

Vowels		Consonants B–N		Consonants P–Z	
Phone	Example	Phone	Example	Phone	Example
[iy]	beat	[b]	bet	[p]	pet
[ih]	bit	[ch]	Chet	[r]	rat
[eh]	bet	[d]	debt	[s]	set
[æ]	bat	[f]	fat	[sh]	shoe
[ah]	but	[g]	get	[t]	ten
[ao]	bought	[hh]	hat	[th]	thick
[ow]	boat	[hv]	high	[dh]	that
[uh]	book	[jh]	jet	[dx]	butter
[ey]	bait	[k]	kick	[v]	vet
[er]	Bert	[l]	let	[w]	wet
[ay]	buy	[el]	bottle	[wh]	which
[oy]	boy	[m]	met	[y]	yet
[axr]	diner	[em]	bottom	[z]	zoo
[aw]	down	[n]	net	[zh]	measure
[ax]	about	[en]	button		
[ix]	roses	[ng]	sing		
[aa]	cot	[eng]	washing	[-]	*silence*

Figure 23.14 The ARPA phonetic alphabet, or **ARPAbet**, listing all the phones used in American English. There are several alternative notations, including an International Phonetic Alphabet (IPA), which contains the phones in all known languages.

Acoustic Model

- A phoneme is the smallest unit of sound that has a distinct meaning to speakers of a particular language. For example, the “t” in “stick” sounds similar enough to the “t” in “tick” that speakers of English consider them the same phoneme.
- Speech systems summarize the properties of the signal over time slices called ***frames***.Each frame is summarized by a vector of ***features.***
- First ***Fourier transform*** is used to determine the amount of acoustic energy at about a dozen frequencies. Then compute a measure called the ***mel frequency ceptral coefficient(MFCC)*** for each frequency.

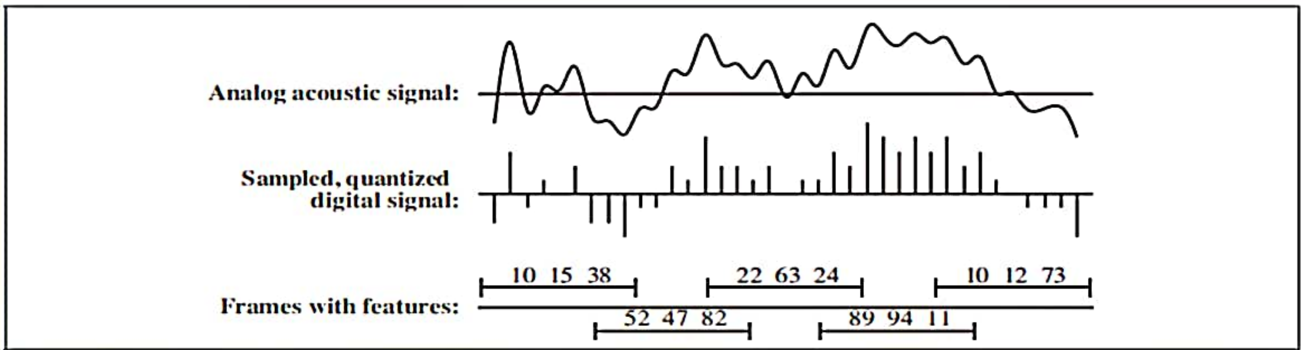

Figure 23.15 Translating the acoustic signal into a sequence of frames. In this diagram each frame is described by the discretized values of three acoustic features; a real system would have dozens of features.

Acoustic Model

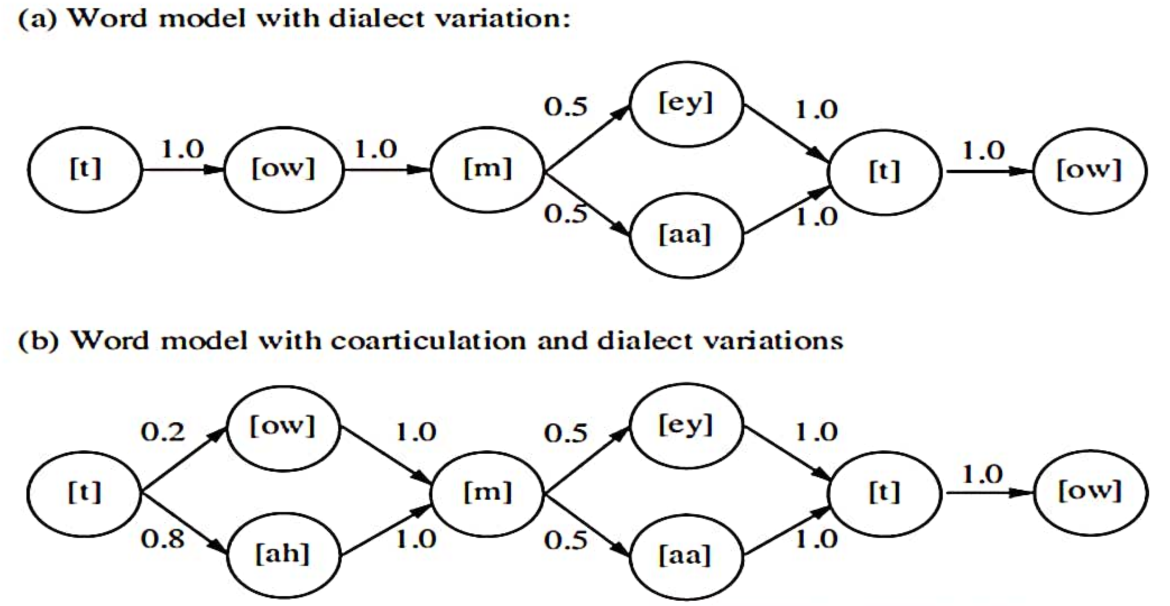

Figure 23.17 Two pronunciation models of the word "tomato." Each model is shown as a transition diagram with states as circles and arrows showing allowed transitions with their associated probabilities. (a) A model allowing for dialect differences. The 0.5 numbers are estimates based on the two authors' preferred pronunciations. (b) A model with a coarticulation effect on the first vowel, allowing either the [ow] or the [ah] phone.

Language Model

- Spoken language has different characteristics than written language, so it is better to get a corpus of transcripts of spoken language.
- For task-specific speech recognition, the corpus should be task-specific.
- To build your airline reservation system, get transcripts of prior calls.
- For example, asking “What city do you want to go to?” elicits a response with a highly constrained language model, while asking “How can I help you?” does not