

## Ch. 5.7 Variation in English pronunciation

### RELATIONSHIP BETWEEN SPELLING CORRECTION & ASR

Motivation: want to use Bayes' rule for spoken language understanding (automated speech recognition).

Spelling: typed sequence of letters --> most probable actual sequence.

ASR: sequence of phones (sounds) --> most probable actual word.

This is a simplification of the actual ASR problem because it assumes we can recognize actual phones.

But there recognizing specific phones is harder than recognizing letters (if we have machine-readable text, recognizing letters is trivial).

a) Ambiguity in defining what a phone is.

Different languages have different pronunciations for what one might consider the same sound.

E.g.: 'R' in American English does not sound like 'r' in British English or French or Arabic. American 'r' is pronounced in the center of the mouth; British 'r' is dental (involves a tongue-teeth combination); French 'r' is uvular (involves the uvula, the hanging part you see in the back of your mouth when you look in the mirror); Arabic 'r' is guttural (pronounced in the throat).

But if you hear a British 'r' in an American context, you still want to code it as 'r'.

And suppose you hear something halfway in between?

Another example: Put your finger on your Adam's apple. When you say 'bin', you will feel it vibrate = 'b' is voiced. For 'pin' you won't = 'p' is unvoiced.

In English, being voiced is a relevant feature of a phone: 'bin' and 'pin' are different words.

The same thing is true in French.

But the French ear (i.e., the ear of someone who has grown up in France--there is no genetic difference!) requires more voicing to hear 'b' than the American ear. So if an American says 'bin' rapidly or casually, a French speaker might hear 'pin'. But not necessarily--there is a lot of variation among speakers and among hearers.

So where should we draw the cutoff between 'p' and 'b'?

b) Different languages require different distinctions.

E.g.: Compare English 'pin' and 'spin'. Put your hand in front of your mouth and say each word out loud. For 'pin', you can feel a puff of air = this 'p' is aspirated. For 'spin', you can't = this 'p' is inaspirated.

All native speakers of American English alternate between these two forms of 'p' in this way, but approximately none of them are aware of it.

Yet this difference is not crucial to speaking English--you can use a different rule, as many speakers from other parts of the world do--and be understood just fine. And no one except a linguist would really notice.

So it's reasonable to say that both aspirated 'p' and inaspirated 'p' should be considered one phone, namely /p/. In fact, we'd prefer to consider all

p's as one phone, i.e., not to make unnecessary differences, because why make the program's life harder trying to differentiate two different kinds of p, when we don't even care about the difference?

But in Thai, for example, aspirate 'p' and inaspirate 'p' are different sounds. If you switch between them, you will change the meaning of the word, just as switching between 'p' and 'b' (e.g. 'pin' vs. 'bin') would in English.

So Thai requires separate phones for aspirate 'p' and inaspirate 'p'.

In French all p's are inaspirate. But if you hear an aspirate 'p', you still want to code it as 'p'.

Summarizing: in reality, we have a probability distribution for each phone.

Another issue: Pronunciation is different in running (continuous) text.

Pronunciation of a sound is different depending on context--beginning vs. middle vs. end of word, surrounding sounds, etc.

Fig. 5.8 = examples of palatalization

Fig. 5.9 = examples of t and d deletion at end of word

This one is especially common in the words *and* and *just*.

Factors increasing probability of deletion:

- faster or more casual speech
- younger speakers, male speakers
- words that form a component
- words that often occur together (even if they're not a component)
  - = high *mutual information*
  - = high *trigram predictability*

Factors decreasing P(deletion):

- -ed past tense suffix

Factors making it harder to see what is happening:

- Rules are different in different varieties of English.
- Different phenomena occur when speakers are aware of their pronunciation (e.g., hyperarticulation).

In connected speech, there are many more pronunciations of common words than you'd expect.

Fig. 5.7 = examples of this.

IPA = International Phonetic Alphabet

ARBAbet = a variant that doesn't require special symbols

We can demonstrate this: how many pronunciations of 'because' do you hear in this sentence:

"We have to change the date because Sara can't make it. It can't be Friday because I won't be here that day. Because Sara won't be here next Monday, we can't meet then. We do have to have a meeting because the dean said so."

## Ch. 5.8 Using Bayes' rule for pronunciation

Consider the phone sequence [ni].

First step: collect the possible words.

Switchboard is a large spoken language corpus collected at CMU. People were given free international phone calls in exchange for giving up their privacy (i.e., having their calls recorded).

There are more words that can be pronounced [ni] than you'd guess:

*neat, need, the, new, knee, to, you*

How can *neat* or *need* be pronounced [ni]?

*t* and *d* are often deleted at the end of a word, especially with certain letters following.

E.g. *neat little* --> [niɫə]

How can *the* be pronounced [ni]?

It occurred in phrases like *in the* [ɪnni], *on the* and *been the*.

Regressive assimilation = a sound moves closer in the mouth to the previous sound spoken.

When is *new* be pronounced [ni]?

In the phrase *New York*.

Vowel is influenced by following *y*.

When is *to* pronounced [ni]?

In a phrase like *talking to you* [tɔkɪniyu].

Here the [u] in *to* is influenced by the following *y*, the final *g* is dropped, and the *n* does double duty. Since the [n] being used in two words, we'll simplify things by ignoring this case.

We'll also ignore *you* as [ni]. (I don't have an example.)

For spelling, we generated candidate words.

For pronunciation, we trade off space for time (and knowledge of rules). We store each pronunciation with all possible variants and their frequencies.

Now we want to calculate:

$$\hat{w} = \operatorname{argmax}_{w \in W} P(y|w)P(w)$$

$P(w)$  = prior probability

$$P(y|w) = \text{likelihood}$$

where  $y$  = sequence of phones

There is no good equivalent of confusion matrices for pronunciation for two reasons:

- Confusion matrices only work for single changes.
- Confusion matrices assume a small set of changes are likely to occur.

Instead, we attach probabilities to rules by counting examples in a corpus.

For example, how often is /th/ (as in *the*) pronounced [n] because the previous word ended in [m] or [n]? How often does /th/ occur?

$$91/617 = .15$$

Fig. 5.10 = results for all the possibilities:

word	p(word is pronounced 'ni')
the	.15 (as above)
neat	.52
need	.11
new	.36

For word frequencies, they use a combination of a written and a spoken corpus. (Is this a good idea?)

Brown corpus = 1,000,000 words of written text.

Switchboard = 1,400,000 words of spoken text.

Denominator = no. of words in the corpus = actual no. of words in corpus + .5 \* no. of distinct words (for smoothing)  
 = 2,486,075 + 30,836 = 2,516,911

So here are the frequencies of our words. We divide by the total word count to get probabilities.

w	freq(w)	p(w)
knee	61	.000024
the	114,834	.046
neat	338	.00013
need	1417	.00056
new	2625	.001

Putting it all together:

w	p(w)	p(y w)	p(y w)p(w)
new	.001	.36	.00036
neat	.00013	.52	.000068
need	.00056	.11	.000062
knee	.000024	1.00	.000024
the	.046	0	0

(Note:  $P(\textit{the})$  is 0 because *the* is only pronounced [ni] after a nasal (e.g. [m], [n]). We are assuming the context (*I* \_\_\_\_). You can see this in Fig. 5.10. "V \_\_\_\_" means after a vowel, which is true in this context.)

To get better results, we need to take context into account.

Looking at one previous word = *bigram* estimate =  $P(\textit{need} \mid \textit{previous word is } T)$  instead of just  $P(\textit{need})$ .

---

Machine learning of decision trees:

In the previous section, we counted examples in a corpus to get the probability that a given rule (e.g., t-deletion when final) was operative.

We'd really like to write a program to get those probabilities.

Fig. 5.11 = decision tree for /t/ induced from Switchboard and pruned by hand.

Machine learning for creating the tree.

Problems w/ machine learning approach.

Improvement (?) from hand-pruning.

Tree didn't include flapping.

But it does show that

- a) /t/ is more likely to be deleted before a consonant than before a vowel.
- b) if /t/ is not deleted before a consonant, it is likely to be unreleased.
- c) /t/ is unlikely to be deleted at the beginning of a syllable.

---

Section 5.9: weighted automata

= efficiently representing multiple pronunciations & their probabilities

Weighted automata = weighted FSA

Sum of probabilities of all arcs leaving a node = 1.

Fig. 5.12 top shows British/Boston pron. vs. rest of country

Bottom shows context-dependent variation too.

Note that the 2 issues are related: the British vowel is less likely to occur with the flap /t/.

Fig 5.13 = 'about' from the Switchboard corpus = data shown previously in table form.