Speaking Rationally

John Pate

Cognitive Modeling

March 12, 2010

1. Background
2. Field Overview
   - Vowel distinctness and H & H Theory: [Lindblom, 1990]
   - Consonant deletion: [Priva, 2008]
   - Placing stress: [Aylett and Turk, 2004]
3. A more in-depth example: [Frank and Jaeger, 2008]
4. Discussion and Open Questions

Big Question

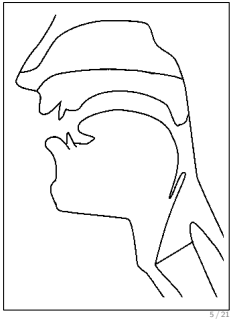
How do talkers decide among different acceptable speech forms?

1. What social class should I sound like I am from?
2. How fast should I talk?
3. How loudly should I talk?
4. Which language should I use?
5. What should I emphasise? How should I emphasise it?
   - I like **chocolate** ice cream.
   - It's chocolate ice cream that I like.
   - **I** like chocolate ice cream.
   - It's me who likes chocolate ice cream.
6. How clearly should I talk?
   - "about" vs. "'bout"
   - "right about oh I guess 'bout four o'clock"
7. How should I put words together?
   - "Is not" or "Isn't"

- Do talkers select different forms in some sort of optimal manner?
  - Longer and more distinct forms ⇒ easier to understand but harder to produce.
  - Shorter and more ambiguous forms ⇒ harder to understand but easier to produce.
- Other explanations
  - Register: Academic speech is different from speech to friends is different from speech to parents is different from speech to your boss. . .
  - Sociolinguistics: Use of certain forms can affirm one's membership in a particular social group.

Vowel Production Basics

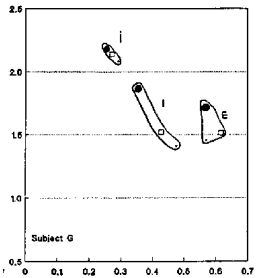
- Voicebox buzzes.
- The size of the cavities in back of the throat and the mouth make certain frequencies from the buzzing noise louder.
- Moving the tongue changes the size of these cavities ⇒ the pattern of loud frequencies
- Large tongue movements result in more acoustically distinct vowels but require more effort.



- [Lindblom, 1990] seeks to explain why talkers sometimes produce very clear vowels and other times produce indistinct vowels.
- H & H theory:
  - talkers **h**ypo-articulate when context provides a lot of disambiguating information.
  - talkers **h**yper-articulate when context provides little disambiguating information.
- [Lindblom, 1990] includes a review of many studies looking at articulatory effort in different contexts.

- Subjects read a word list twice:
  1. First, just given the list and read it.
  2. Second, told to articulate as clearly as possible.
- Subjects also recorded saying "heed," "hid," "head," &c.

- NULL CONTEXT : /h_d/
- CITATION FORM
- CLEAR SPEECH



- Consonant Deletion: "explanation" or "explanatio'"
- When can we delete a consonant?
  - Intuition: Predictable consonants can be deleted.
- Two kinds of predictability (figures from Switchboard):
  1. Overall probability of /t/ and /ŋ/ ("ng"):

$$P(/\text{ŋ}/) = \frac{779}{99{,}280} = 0.78\%$$

$$P(/\text{t}/) = \frac{779}{99{,}280} = 3.61\%$$

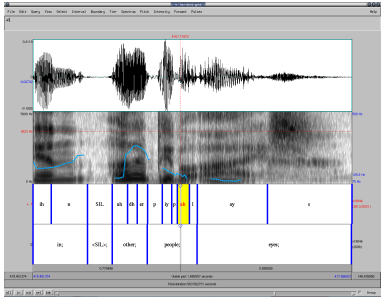
  2. Probability of /t/ and /ŋ/ ("ng") after ɪ:

$$P(/\text{ŋ}/|/\text{ɪ}/) = \frac{542}{11{,}919} = 4.55\%$$

$$P(/\text{t}/|/\text{ɪ}/) = \frac{351}{11{,}919} = 2.94\%$$

- **Informativity** is a measure that combines these two sources of predictability (similar to weighted entropy).

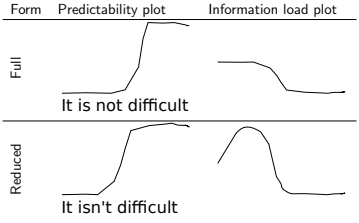
- Corpus study looking at the predictive power of informativity:
  - Buckeye Corpus: speech corpus of a range of talkers from Columbus, Ohio which has been hand-annotated by phoneticians on the phone level.
  - A pronunciation dictionary is used to identify when phones have been deleted.
  - Informative phones are less likely to delete.
  - Phones following highly informative phones are more likely to delete.
  - Phones preceding highly informative phones are more likely to delete.



- Stress:
  - "I want to wear the **blue** shoes."
  - "I want to wear the blue **shoes**."
- Stress is generally realized with longer duration, louder volume, and more distinct vowels, among other things.
- When should a word be stressed?
- [Aylett and Turk, 2004] provide an explanation that should be relatively familiar by now:
  - We should provide lots of cues in the acoustic signal only when necessary.
  - Stress leads to more acoustic cues
  - So we should find stress appears on less predictable words.
- Their paper includes a corpus study backing this up.

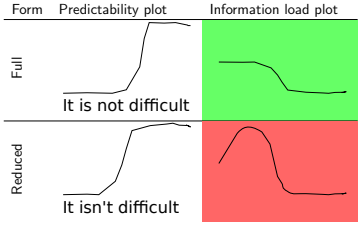
- How do people choose between the "full" and "reduced" forms of **be**, **have**, and **not**?
  - "I am" or "I'm"
  - "We have" or "We've"
  - "is not" or "isn't"
- **Uniform Information Density (UID)**: Talkers prefer to produce speech which does not have sudden spikes in information content.
  - If a region of speech contains more information, take longer to say it (i.e. use a full form instead of a reduced form)
- Explicit Hypothesis: Talkers use full forms in less probable contexts, and reduced forms in more probable contexts.

- Low-probability reducible target, high-probability following word:



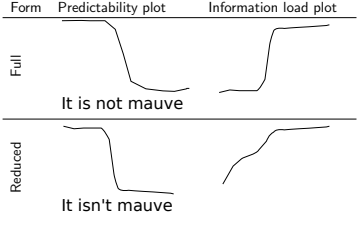
Speaking Rationally
A more in-depth example: [Frank and Jaeger, 2008]

- Low-probability reducible target, high-probability following word:



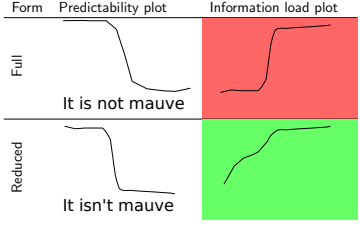
Speaking Rationally
A more in-depth example: [Frank and Jaeger, 2008]

- High-probability reducible target, low-probability following word:



Speaking Rationally
A more in-depth example: [Frank and Jaeger, 2008]

- High-probability reducible target, low-probability following word:



- Method: Look at real speech, see if talkers take longer to say less probable things.
  - The probability of a sequence of words is based on an unsmoothed maximum likelihood estimate from the corpus (i.e. count and divide).

$$P(w_i = \text{not} | w_{i-2,i-1} = \text{it,is}) = \frac{P(w_{i-2..i}=\text{it,is,not})}{P(w_{i-2,i-1}=\text{it,is})} \approx \frac{|w_{i-2..i}=\text{it,is,not}|}{|w_{i-2,i-1}=\text{it,is}|}$$

  - Rare events are thrown out (maximum likelihood is inaccurate with rare events).
  - Duration is not explicitly measured – Full forms are just assumed to take more time than reduced forms.

- Data set
  - Many conversations between two participants by telephone.
  - Relatively naturalistic setting.
  - Spontaneous (not read or practiced) speech.
  - Automatically extract sentences with reducible elements (hand annotation of grammatical structure allows this).

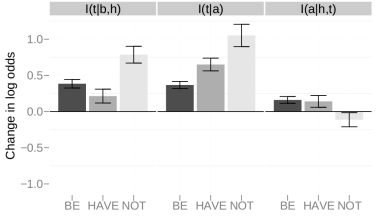
Time (s)	Speaker	Turn
191.9 - 197.2	A	but uh the guy winds up getting hurt every other game and you can't do that and stay
194.7 - 207.1	B	yeah i i tell you it's it's difficult in in that guy's position coming into that because there he was just so highly touted by the press and everybody expected so many big things you know
205.9 - 215.5	A	yeah they did they put a lot of pressure on him from the outside and from the inside uh it's funny watching them them play [vocalized-noise] he's probably like a lot of quarterbacks uh
215.5 - 223.2	A	when the pressure is really on when it's down to the last few minutes of the game for the season is when the guys seem to really do their best
222.4 - 223.7	B	uh-huh
223.2 230.2	A	and i haven't quite figured that out if if they figure they have got it won or if there's no real hurry because the first three quarters
230.2 - 241.8	A	or uh if [vocalized-noise] if something happens that that adrenaline really starts flowing they say hey we got to do something now and and start playing the game the way the game should be played toward the last few [vocalized-noise] few minutes

- Analysis
  - Three conditional probabilities are computed for the reducible elements found in the corpus.

before	host	target	after
it	is	not	difficult
it	is	n't	difficult

- $p(\text{target}|\text{before,host}) = p(\text{not}|\text{it,is})$
  - Is the reducible target likely to follow the first two words?
- $p(\text{target}|\text{after}) = p(\text{not}|\text{difficult})$
  - Is the reducible element likely to be followed by the word we see following it?
- $p(\text{after}|\text{host,target}) = p(\text{difficult}|\text{is,not})$
  - Is the following word likely to follow the preceding two words?

- How well do these predict the coice of full form vs reduced?
- Results: full forms used in strings with higher information load.



- There are many ways to say the same thing.
- We've looked at some studies arguing that people pick what they're going to say to optimize communicative efficiency.
  - Produce more informative vowels more clearly.
  - Delete consonants only when other phones disambiguate.
  - Stress informative words.
  - Use full forms when context does not disambiguate.
- Very impoverished notion of information–trigrams are a small part of the whole story!
  - What have we been talking about so far?
  - What things are being contrasted with eachother?
  - What counts as "common knowledge"?
- Are talkers doing these things for themselves or for the hearer?
  - Do talkers want to make sure the hearer has an easier time?
  - Do talkers themselves have a harder time accessing language items when encoding lots of information?

Aylett, M. and Turk, A. (2004).
The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech.
Language and Speech, 47(1):31 – 56.

Frank, A. and Jaeger, T. F. (2008).
Speaking rationally: Uniform information density as an optimal strategy for language production.
In Proceedings of The 30th Annual Meeting of the Cognitive Science Society (CogSci08), pages 933–938.

Lindblom, B. (1990).
Explaining phonetic variation: a sketch of the h & h theory.

Priva, U. C. (2008).
Using information content to predict phone deletion.
In Proceedings of the 27th west coast conference on formal linguistics, pages 90 – 98.