Automatic Detection of Turn-taking Cues in Spontaneous Speech

Kyle Gorman
Department of Linguistics
University of Pennsylvania

Jennifer Cole
Department of Linguistics
University of Illinois at Urbana-Champaign

Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Margaret Fleck
Department of Computer Science
University of Illinois at Urbana-Champaign

- If you remember, I performed a small pilot using a pseudo-gating paradigm for the D+P+C tree, and found that the duration cues provide more information (I'm using that in the technical sense) when less of the pause duration is available. This appears in slide 16 / p. 10 of the thesis (pretty version at http://www.ling.upenn.edu/~kgorman/kgorman_thesis.pdf). I must confess
though, I'm not sure how to evaluate the small decrease observed in delta-accuracy (between P and D+P+C). Is there some way to determine
the confidence with which I can assert that this is not caused by chance?

Introduction

- Dialogue systems must detect the ends of turns and initiate new turns
- We wish to understand the phonetics of spont. speech and discourse structure
- How is turn-taking controlled in speech?
- How does it fit into the overall prosodic structure of dialogue?

Introduction

- Problem: turn-taking is *tightly* coordinated in dialogue (order of ms)

a: I got this cough, I've got a cold because it was eighty degrees up here and I went outside with no coat on. [silence: 85 ms]

b: Oh boy! [laughs] "cough": 623 ms "no": 225 ms

- Suggests turn-taking cues in speech
- Simple model: vocal cue from yielding speaker (ignore multimodal interaction)

Method of Inquiry

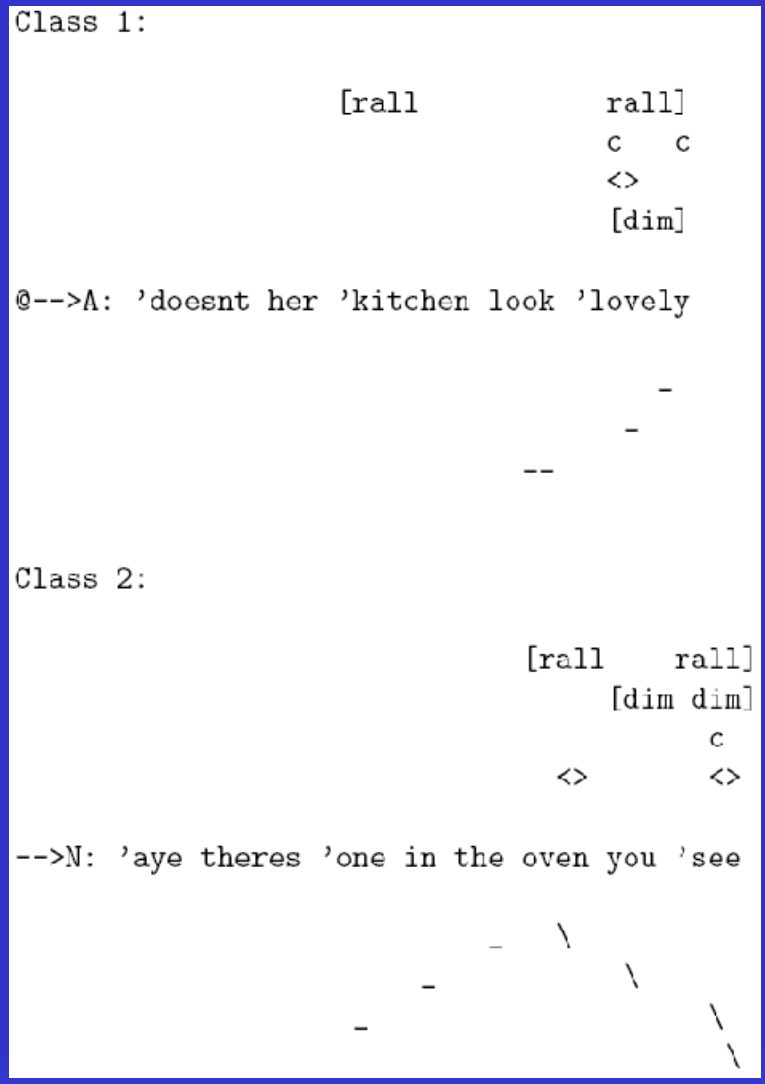
- Phonetic expression of turn-taking (or discourse)
- Investigated using descriptive analysis
- Or by automatic classification methods
- Either way, unscripted speech
- TT an element of prosodic structure

Local et al. (1986)

- Corpus: Tyneside Eng. home recordings
- Descriptive analysis
- Name four features
  - lengthening
  - pitch rise or fall
  - centralization
  - swell

- rall indicates *rallentando* (gradual slow-down)
- dim marks *diminuendo* (gradual decrease in intensity)
- c marks a centralized vowel
- < and > indicate a rapid swell in intensity.
- Lines underneath the transcription mark observed relative pitch
- The primes ' indicate stress.

- Problem
  - What is a cue?
  - What isn't?



Ferrer et al. (2002, 2003)

- Corpus: ATIS (air traffic Wizard of Oz)
- Automatic classification system (CART)
- Autoextracted prosodic features
  - duration feature set
  - filtered (Sönmez 1998) F0 contour features
- Online with multiple decision points
- Baseline: .81 false alarm rate

Ferrer et al. (2002, 2003)

- Final system: .02 false alarm rate with 1.6 s threshold
- With ASR-derived LM and prosody: .049 false alarm with .135 s waiting time
- High information features not reported
- Limited domain, short utterances, not spontaneous, human-computer modality

Experimental design

- Corpus: Switchboard phone convos
- Automatic feature extraction (SRI prosodic database with additions)
  - duration features
  - filtered F0 contour features
  - "context" features
  - pause

CART Classification Method

courtesy A. Moore (http://www.autonlab.org/tutorials/)

- Given some attributes, predict the value of another attribute (output)
  - (in this case, a binary yes/no of whether or not a word is turn-final)
- Decision tree: a plan for attribute testing to predict the output

CART Classification Method

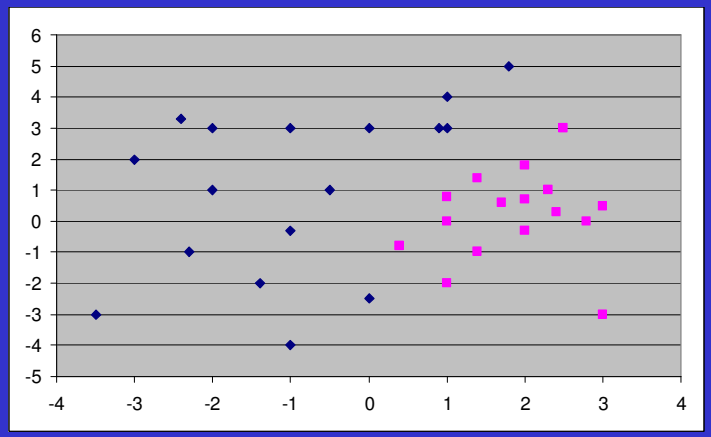
- *Information gain:* a distance measure between observed probabilities and model probabilities

$$IG(Y|X) = H(Y) - H(Y \mid X)$$

- Algorithm:
  - decide the order of attributes to test by choosing the one with the highest IG
  - recurse

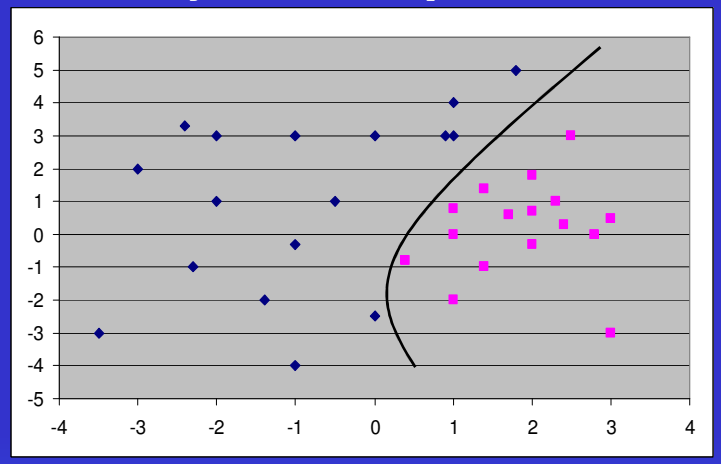
How to classify populations

Two populations

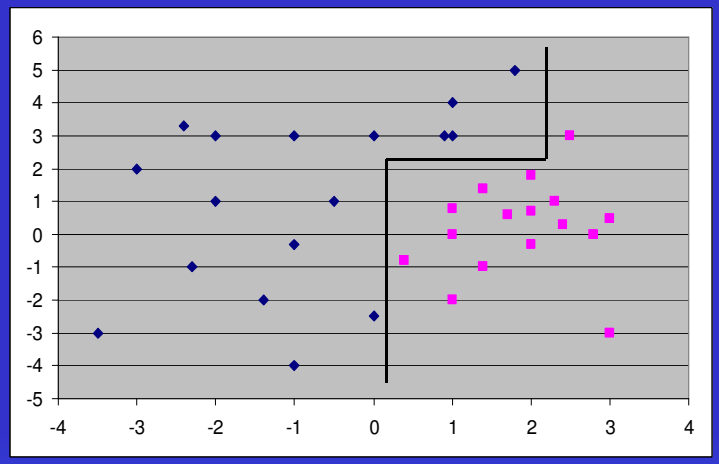


- (two attribute example)
- Higher order equations…
- Or multiple lower order equations (the CART way)

Higher order equations...



Or the CART way: remember to recurse until all data is classified



Experimental Method

- Train on 70% of the corpus
- Prune on 20% (held out)
- Test (evaluate) on 10% (held out) with different combinations of feature sets
- This performed better than using the automatic pruning in CART, despite expectations to the contrary

Results

- Baseline: .5 accuracy (chance)
- Pause only: .898 accuracy
- Duration only: .704 accuracy
- F0: .513 accuracy (Kochanski et al.)
- D+P+C: .938
- F+P+C: .946
- D+F+P+C: .936 (CART limitations)

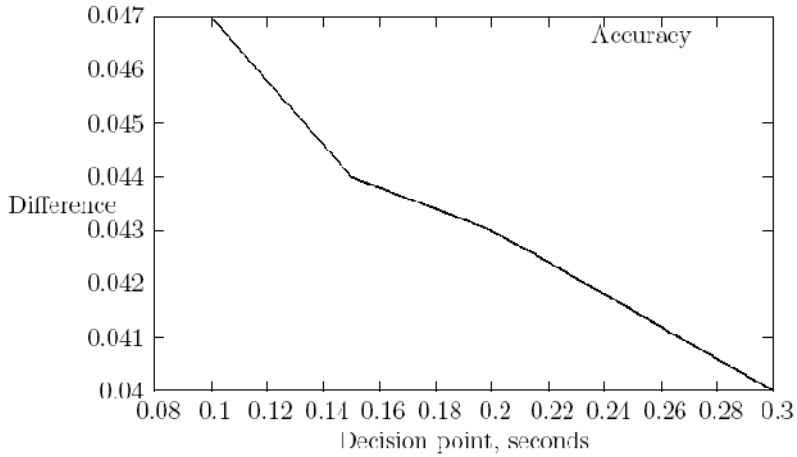
Discussion

- Pause itself a useful cue (but higher decision latency)
- Duration with pause: the stress foot the domain for turn-final lengthening (a useful cue) (cf. Turk & Sawusch 1997)
- F0 useful with other information, but perhaps not a cueing feature

Pilot study - online system

- Simulating an online system
- Similar to gating paradigm
- Reduce latency
- Duration provides more information when pause-length is gated

Figure 3: Difference in accuracy between a P and a D+P+C tree when time-shortened



Future work

- More ASR/classification studies of spontaneous speech
- Particularly disfluency
- You can extract useful prosodic features from a corpus
- Better psycholinguistic studies
- deRuiter et al. 2006

Acknowledgements

- The co-authors
- The Prosody-Disfluency/ASR group at the Beckman Institute @ UIUC
- Mark Liberman and Jiahong Yuan
- The students of the Institute for Research in Cognitive Science @ Penn
- The LSA, fellow students, friends, family