

VSLP-2011

Jan 29, UPenn

# **Acoustic Classification of Focus in a Web Corpus of Comparatives**

Jonathan Howell, Mats Rooth, Michael Wagner



**Cornell University**  
Linguistics Department



**Cornell University**  
Computing and Information Science



**McGill**

prosody.lab



# Support



National Science Foundation



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada

# Domain of Inquiry

- Observed correlation between
  - i. what has been uttered or is salient in a discourse
  - ii. which phonetic material in an utterance is realized with prosodic prominence

## Challenges in studying this correlation

- Empirical challenges
- Methodological challenges

## **Empirical challenge:**

What are the objects of study?

How do we identify and/or measure them?

- Pragmatics/semantics

## Assume: Focus anaphoricity (Rooth 2008)

- focus involves a relation to context which is a kind of anaphora

### Licensing condition

The antecedent entails the union of the alternative set (focus existential closure)

- (1) A: I heard **[the quarterback took steroids]<sub>2</sub>**  
B: Yes. In fact, **[the WHOLE TEAM<sub>F</sub> took steroids]** ~2

Focus existential closure: 'some person x took steroids'

Antecedent: 'the quarterback took steroids'

- Phonetics / Phonology
  - abstract, categorical objects  
nuclear pitch accent?  
stress?
  - directly observable, gradient measures  
fundamental frequency?  
formant extrema?  
duration?

all (or some combination) of the above?

“a complex of properties that can be related to greater force of articulation, including increased intensity and duration, and shallower spectral tilt” (Ladd 1996:58)



## **Methodological Challenge:**

- We want:
  - to study speech from natural discourse
  - to control for grammatical/pragmatic conditioning
- difficult to recreate real-world context in the lab environment
- traditional speech corpora lack specific focus-sensitive constructions in sufficient numbers to allow a quantitative analysis

# Our goals

- Assemble large, focused datasets of examples where prominence varies in a way that correlates with syntax, semantics, or pragmatics
- Study correlation between lexical/grammatical/pragmatic context and acoustic realization of prominence

# Outline for rest of presentation

- the first dataset
- our datasource
- web harvest methodology (Howell & Rooth 2010)
- classification experiments

# First dataset

Search query: ... *than I did* ...

- Theory makes predictions for the location of prominence
- Antecedent for comparative clause is syntactically obligatory

*he stayed longer than I did*

-er [[ he stayed d long]<sub>2</sub>

than [ I<sub>F</sub> stayed d long ] ~2]

[ y stayed d-long ]

antecedent clause

[ speaker stayed d-long ]

scope of focus

Licensing condition for focus

The antecedent entails the union of the alternative set (focus existential closure).

‘He stayed  $d$  long’ entails ‘someone stayed  $d$  long’

## Our Datasource

|                                       |             |                 |
|---------------------------------------|-------------|-----------------|
| search.everyzing.com<br>(now defunct) | <i>web1</i> | 91 true tokens  |
| multimedia.play.it<br>(now defunct)   | <i>web2</i> | 127 true tokens |
| mediasearch.wnyc.org                  |             |                 |

Ramp (formerly Everyzing) is a commercial audio indexing service using ASR


Content includes radio and tv programs, podcasts

EveryZing - WAMU: The Diane Rehm Show Podcast - Hans Blix: "Why Nuclear Disarmament Matters" (Boston Rev... ⏏ ⏏ ⏏

File Edit View History Bookmarks Tools Help

⏮ ⏪ ⏩ ⏭ 🏠 🔍 <http://search.everyzing.com/viewMedia.jsp?res=214863535&dedu> 📡 ▶ 🔍 graph powerpoint

EveryZing - WAMU: The Diane Re... ⏏ 📄 Graphs in PowerPoint XP.pdf (applicati... ⏏



TO FIND THE MEDIA YOU WANT, DON'T JUST SEARCH FOR IT...ZING IT!

SEARCH:


☐ AUDIO ☐ VIDEO ☒ BOTH

[Sports](#) [News & Politics](#) [Lifestyle & Health](#) [2008 Presidential Election](#) [Gaming](#) [Technology](#) [Business](#) [Food & Videos](#) [Entertainment](#) [Finance](#) [Home Improvement](#) [Comedy Shows](#)

---

**WAMU: The Diane Rehm Show Podcast - Hans Blix: "Why Nuclear Disarmament Matters" (Boston Review)** Apr 08, 2008

The former U.N. chief weapons inspector says the world is sleepwalking towards a new arms race. He joins us to discuss his proposals for how the nuclear powers can [show more](#)



Your Search Term: "than I did"

Key Terms: ["US senate"](#) ["Julie Andrews"](#) ["Iran Iraq"](#) ["Iraq war"](#)

[Full Text](#)

▶ [Play Here 0:45:57](#) ... While you did better **than I did** I didn't really know how long the war would be but which you failed at that time. That they were eyes -- ...

Done



WNYP Audio Search for "than I did" - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://mediasearch.wnyc.org/search?q="than I did" wnyc

Most Visited Customize Links RefGrab-It Site Title Mp3Realm Google

WNYP - New York Public Radio WNYP Audio Search for "t...

Showing 1 - 10 of 23 results for ""than I did""

Sort By: Date | Relevance

**The Takeaway for Friday, August 14, 2009 (hour 4)**  
 The Takeaway for August 14, 2009 [View original source »](#)  
 Audio | Fri, 14 Aug 2009 | **than I did** found at 6:52

"...I have to tell yet that I heard more in your laugh **than I did** -- what you just said so if we can help me understand. ..."

**June 19, 2009 (On The Media: Friday, 19 June 2009)**  
[View original source »](#)  
 Audio | Fri, 19 Jun 2009 | **than I did** found at 47:53

"...I spent reading little door I bet I spend more hours. Listening **than I did** to any of the other formats and that I think is because of the flexibility. Of an audio book. You know you're ..."

**The Takeaway for Monday, April 27, 2009 (hour 2)**  
 The Takeaway for April 27, 2009 [View original source »](#)  
 Audio | Mon, 27 Apr 2009 | **than I did** found at 36:27

"... You're absolutely right you know I definitely feel differently **than I did** you know the last time now we spoke and I feel great if we are. Has -- sales increases now with what's ..."

**The Takeaway for Tuesday, April 07, 2009 (hour 1)**  
 The Takeaway for April 07, 2009 [View original source »](#)  
 Audio | Tue, 7 Apr 2009 | **than I did** found at 47:52

"...Are you okay. Okay honey okay where did better **than I did** okay. ..."

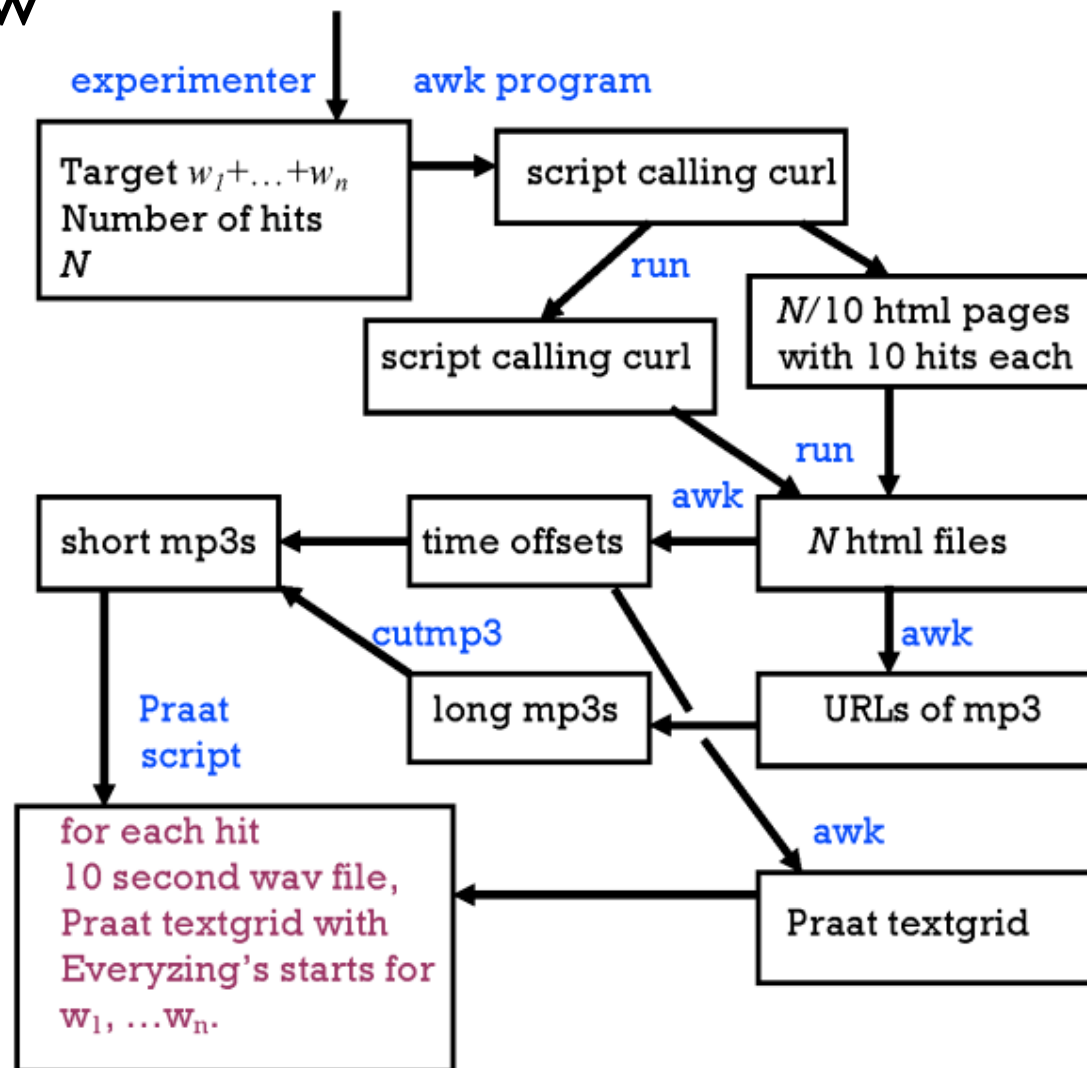
**Follow Up Friday: 100 Days (The Brian Lehrer Show: Friday, 01 May 2009)**  
 On Wednesday Barack Obama explained what he's done in the last 100 days. Now it's

# Web Harvest Methodology

Howell & Rooth 2010

- Turn-key procedure using Unix tools (e.g. curl, cutmp3, awk, bash, make)
  - replicates user interaction with website

# Workflow

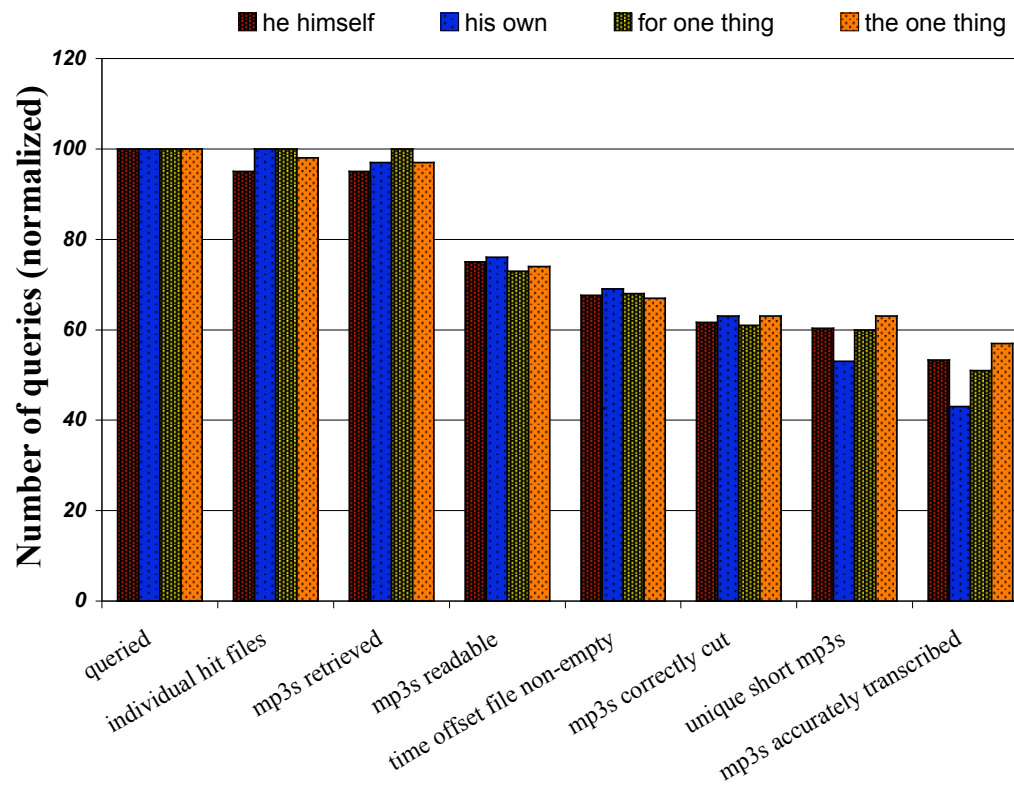


## Some steps yet to be automated

- elimination of duplicates common in radio programming
  - > measure similarity between ASR transcripts and acoustic signals
- phonetic-level annotation
  - > segmentation by forced alignment using HTK hidden Markov model toolkit
- sentence-level annotation
  - > sentence-level annotation with commercial transcription service or Amazon Turk

# Retrieval efficacy

- roughly half or more of hits yield correct, unique speech tokens



# Binary Classification Experiment

- Task: Predict focus class from acoustic parameters alone

|   |          |
|---|----------|
| (1) He stayed longer than I <sub>F</sub> did<br>antecedent: he stayed x long  | s class  |
| (2) I should have liked that song more than I did <sub>F</sub><br>antecedent: I should have liked that song x much. | ns class |
| (3) I understand even less than I did before <sub>F</sub><br>antecedent: I understand even x little                 | ns class |

# Input for statistical classifiers

- Semantic class:
  - Determined from the TEXT ALONE by co-reference of subjects

If subjects co-refer, ns class

Else, s class

# Input for statistical classifiers

- Acoustic features:
  - 308 parameters extracted from string *than I did*
    - incl. duration,  $f_0$ ,  $f_1$ ,  $f_2$ , intensity, amplitude, voice quality, spectral tilt
    - means, extrema, range
    - over a vowel, at regular intervals, at times of other events



# Questions

- Is the theoretical prediction for location of prominence in comparative clauses confirmed?
- Will the variability in speakers, recording conditions, audio compression, levels of formality, discourse conditions, etc. help or hinder classification compared to laboratory

# Questions

- Which set of acoustic features is most predictive?
  - Pitch-first theories privilege  $f_0$ -related features
  - Stress-first theories privilege stress-related features (duration, intensity, formant extrema)
    - if pitch accents align with stress, then F0 correlates expected
    - however, pitch cues might not be necessary or may be secondary to other stress cues

# Machine Learning

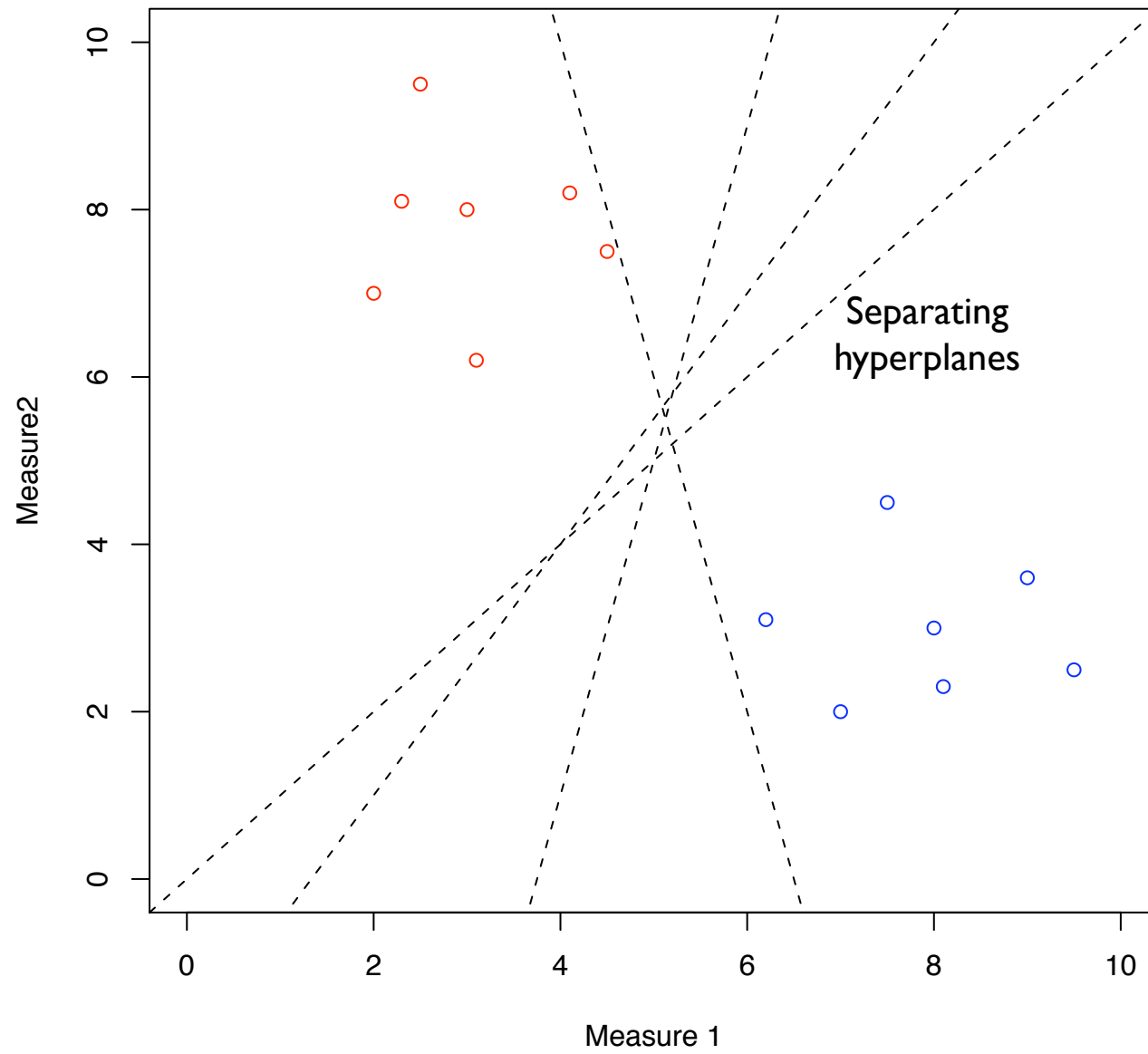
- Linear Discriminant Analysis (LDA)
  - minimize within-class distance and maximize between-class distance
  - assumes normal distribution of classes, homogeneity of classes
  - performs best with small number of attributes

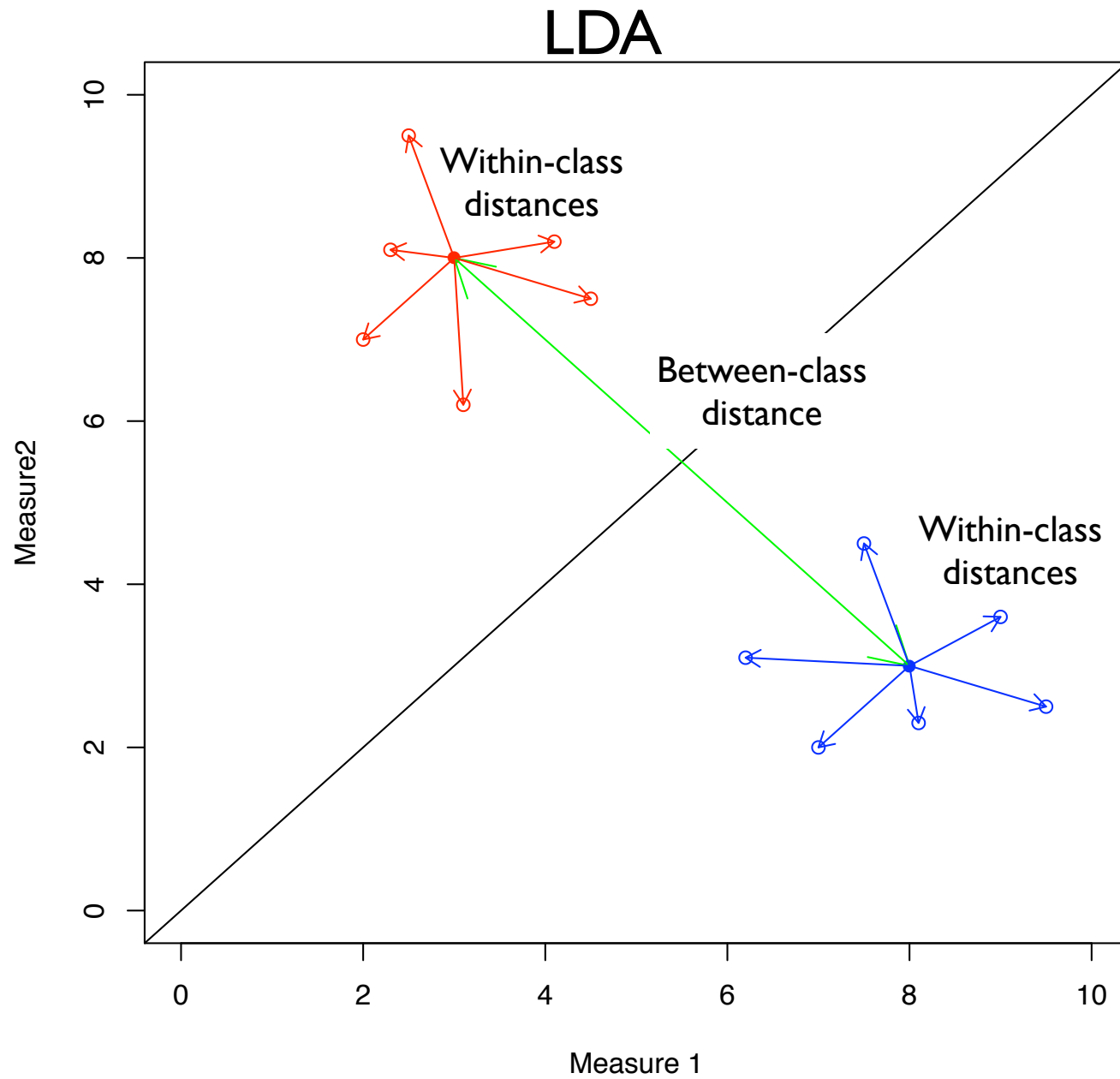
Implementations in R statistical programming environment: package MASS

# Machine Learning

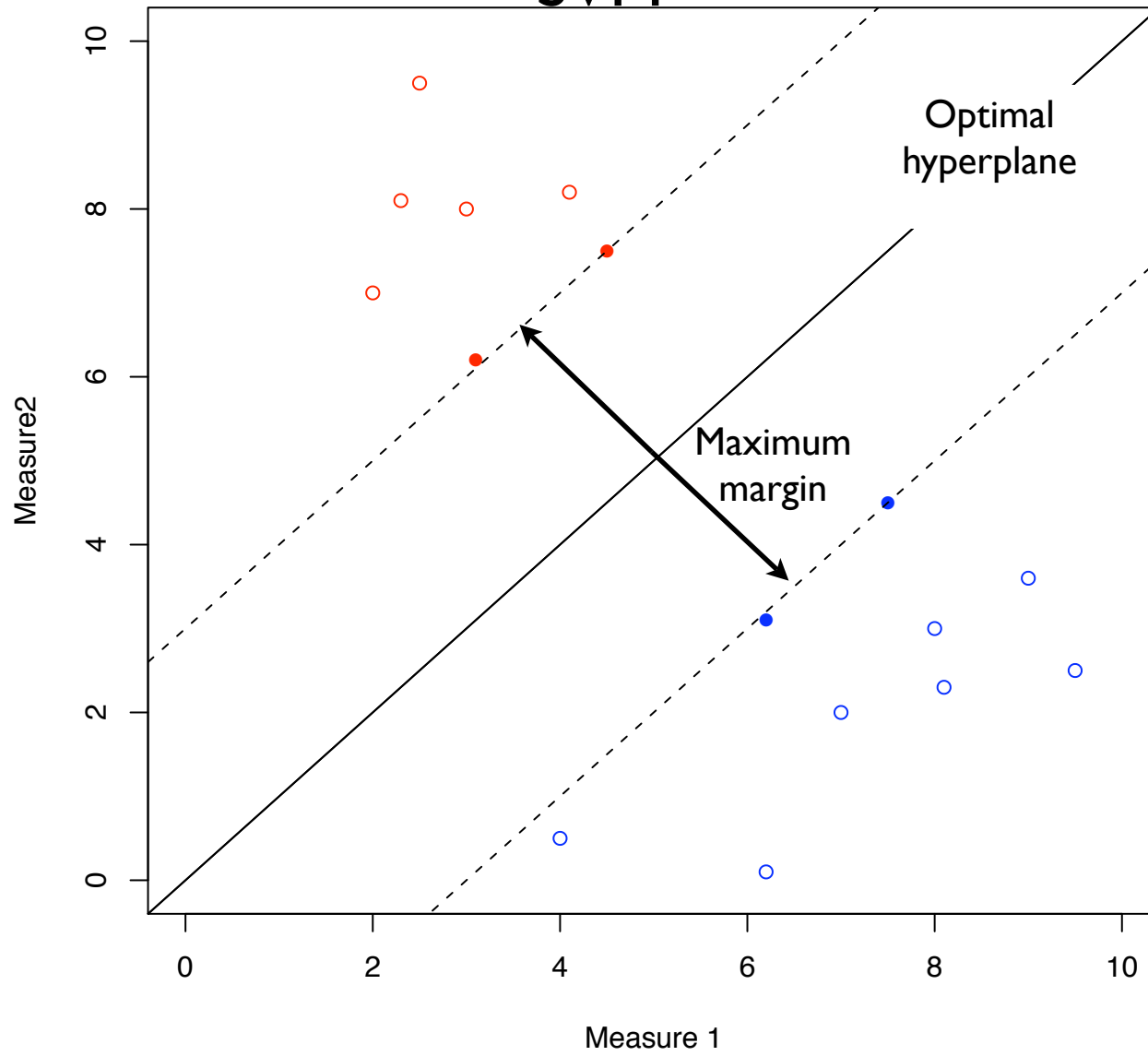
- Support Vector Machine (SVM)
  - maximize margin between classes
  - works well with large number of attributes: data mapped (using “kernels” to high-dimensional featurespace)
  - works well with sparse data: no assumptions of normal distribution and homogeneity

Implementation in R statistical programming environment: libsvm in package e1071





# SVM



# Feature Selection

- Many of the features may turn out to be redundant and/or irrelevant
  - e.g.  $f_2-f_1$  measured at 10% of vowel
  - $f_2-f_1$  measured at 20% of vowel
- Manual selection by experimenter
  - theory-informed, trail-and-error
- Automatic selection by algorithm
  - random-forest based algorithm: VarSelRF package

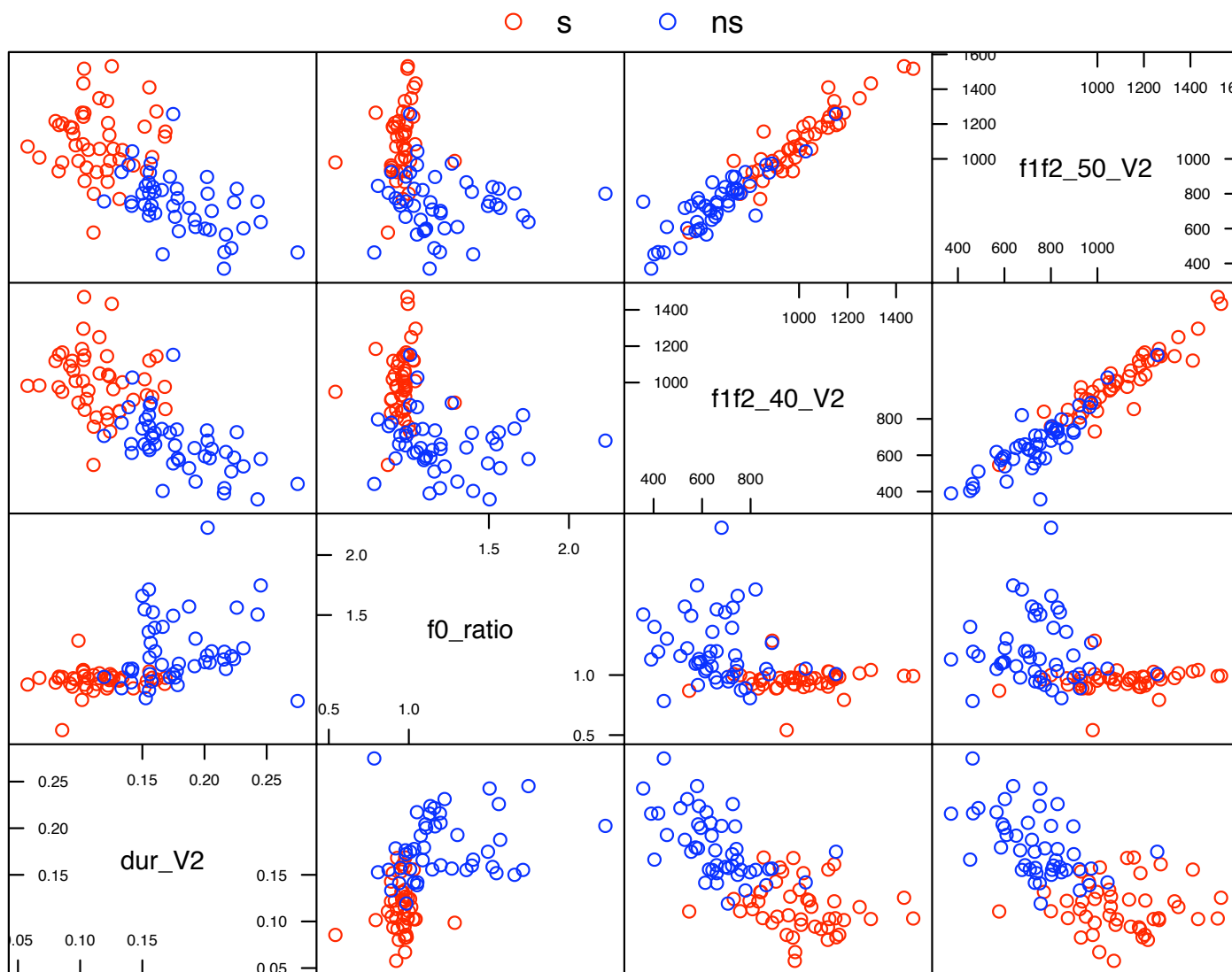


## e.g. Automated Feature Selection B

- given all 308 features, VarSelRF selected 4

|            |   |
|------------|---|
| dur_V2     | duration of <i>I</i>                          |
| f0_ratio   | ratio of f0 maxima in <i>I</i> and <i>did</i> |
| f1f2_40_V2 | f2-f1 at 40% into <i>I</i>                    |
| f1f2_50_V2 | f2-f1 at 50% into <i>I</i>                    |

# Automated Feature Selection B



Scatter Plot Matrix

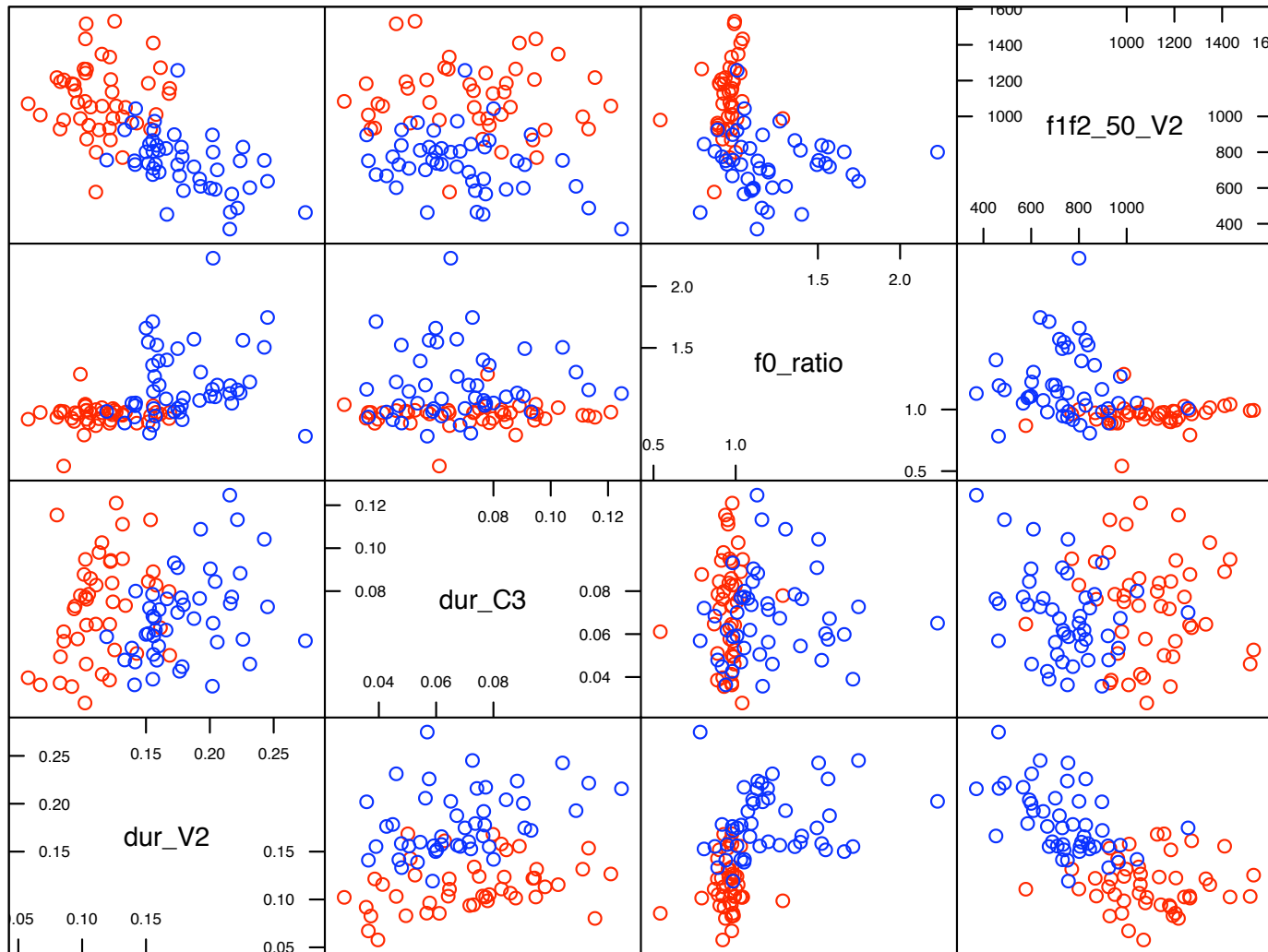
## e.g. Hand-picked A

- an experimenter-selected set that turned out to perform well

|            |  |
|------------|--|
| dur_V2     | duration of <i>l</i>                                   |
| dur_C3     | duration of 1 <sup>st</sup> stop closure in <i>did</i> |
| f0_ratio   | ratio of f0 maxima in <i>l</i> and <i>did</i>          |
| f1f2_50_V2 | f2-f1 at 50% into <i>l</i>                             |

# Hand-picked A

○ s    ○ ns



Scatter Plot Matrix

# Evaluation of classifier performance

- Baseline accuracy

- $\frac{\text{\# tokens in largest class of test set}}{\text{\# tokens in both classes in test set}}$

- Generaliation accuracy

- $\frac{\text{\# of tokens in test set accurately classified}}{\text{\# of tokens in test set}}$

- Balanced error rate

$$\frac{\text{\# incorrect "s"}}{\text{\# total "s"}} * \frac{\text{\# incorrect "ns"}}{\text{\# total "ns"}} * \frac{1}{2} * 100$$

# Classifier Performance

Training set: web1 91 tokens 46:45

Test set: web2 127 tokens 62:65

|   | Classifier performance on <i>web2</i> |                        |                           |                            |
|---|---------------------------------------|------------------------|---------------------------|----------------------------|
| Feature set   | Baseline                              | SVM<br>(radial kernel) | SVM<br>(linear<br>kernel) | LDA                        |
| 1. Full set<br>(no. features = 308)                     | 51.2                                  | 82.7 (14.3)            | 85.0 (14.2)               | n/a due to<br>collinearity |
| 2. Automated feature selection A<br>(no. features = 16) | 51.2                                  | 89.8 (10.1)            | 89.0 (10.3)               | 90.6 (9.0)                 |
| 3. Automated feature selection B<br>(no. features = 4)  | 51.2                                  | 86.6 (13.1)            | 90.6 (9.2)                | 87.4 (11.9)                |
| 4. Hand-picked A (no. features = 4)                     | 51.2                                  | <b>92.9 (6.5)</b>      | <b>92.1 (7.1)</b>         | <b>91.3 (7.7)</b>          |
| 5. Hand-picked B (no. features = 3)                     | 51.2                                  | 91.3 (7.7)             | 92.1 (7.1)                | 91.3 (7.7)                 |

# Theoretical predictions confirmed?

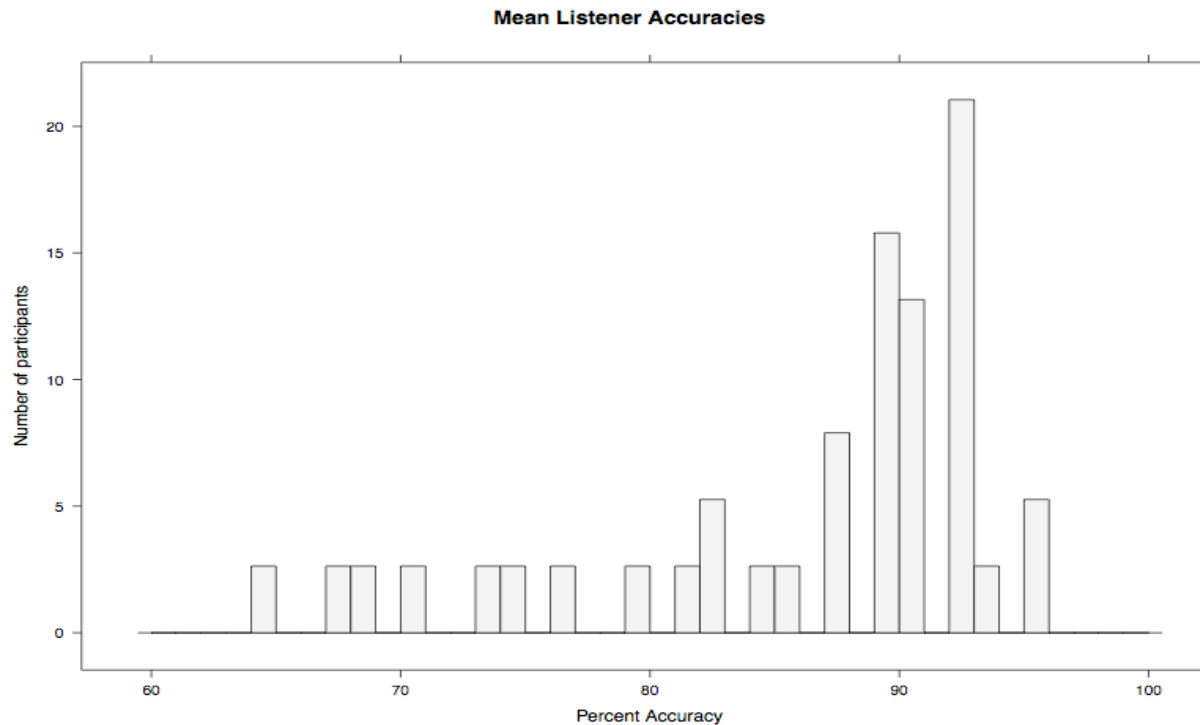
- There seems to be a very robust correlation.
- Are these accuracy and error rates reasonable?



# Human classifiers

- Perception experiment (prosodylab @ McGill)
  - 38 participants
  - subset of 64 tokens from *web2*
- *than I did* sequence extracted
- Question 1: Which is more prominent: *I* or *did*?
- Question 2: How confident are you?  
(very uncertain) 1 2 3 4 5 6 7 (very confident)

- mean accuracy: 85.9% (64.1-95.3%)
- balanced error rate: 14.1% (4.7-35.9%)
- confidence rating significant predictor of performance  
(generalized linear model:  $\sigma = 0.031$ ,  $z = -10.81$ ,  $p < 0.001$ )



# Most predictive acoustic features?

- Models with  $f_0$  information performed on par with models lacking it

Hand-picked A: dur\_V2, dur\_C3, flf2Time50\_V2, f0\_ratio

Hand-picked B: dur\_V2, dur\_C3, flf2Time50\_V2

| Feature set                            | Baseline    | SVM<br>(radial)   | SVM<br>(linear)   | LDA               |
|--|-------------|-------------------|-------------------|-------------------|
| 4. Hand-picked A<br>(no. features = 4) | <b>51.2</b> | <b>92.9 (6.5)</b> | <b>92.1 (7.1)</b> | <b>91.3 (7.7)</b> |
| 5. Hand-picked B<br>(no. features = 3) | 51.2        | 91.3 (7.7)        | 91.3 (7.7)        | 90.6 (8.3)        |

# Stress-first

- Results consistent with stress-first accounts of focus
- Note: for theories which assume alignment of pitch events, stress comes “for free”
  - e.g. pitch accents attach to strongest stress in a phrase (Pierrehumbert 1980, Selkirk 1995)

p.a.  $\{H^*, L^*, L+H^*, \dots\}$

x

x x x x x x

xxxx xxxxxx xxxxxx

- Pitch accent type has been argued to correlate with information other than just focus marking (e.g. Ward & Hirschberg 1985, Steedman 2003)

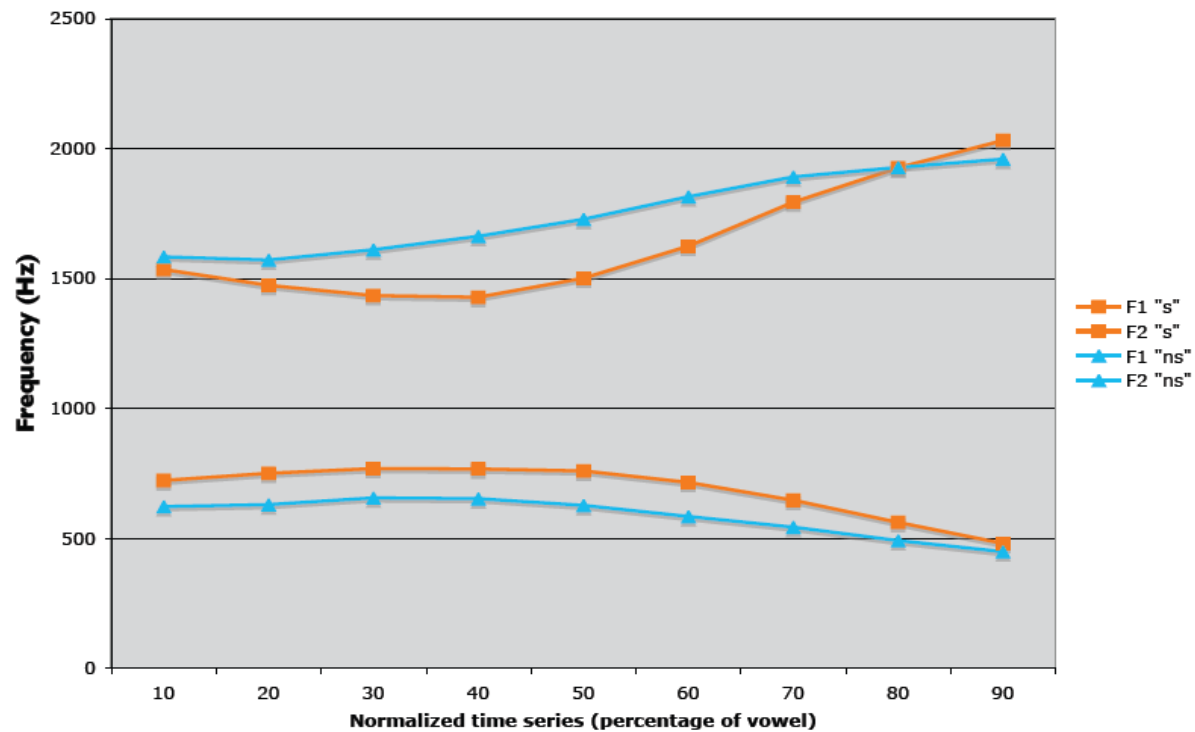
# Hyperarticulation

- From a phonetic standpoint, these results also support hyperarticulation theories of prominence (e.g. de Jong 1995, Fowler 1995, Cho 2005)
- The classifiers are good at detecting focused pronouns using local features on pronoun:
  - duration of vowel in “I”
  - distance between f1 and f2 halfway into vowel in “I”

# “than I did”

## Normalized f1 and f2 for “I” (n=91)

---



# Variability: Web vs. Lab

Will the variability in speakers, recording conditions, levels of formality, discourse conditions, etc. help or hinder classification compared to laboratory data?

# Production study

- prosodylab @ McGill University
  - 26 participants
  - 16 written stimuli, based on corpus data
  - 12 statements, 4 questions



## Classification Experiment 2: Lab-Trained, Web-Tested

|               |                    |            |         |
|---------------|--------------------|------------|---------|
| Training set: | <i>lab</i>         | 394 tokens | 193:201 |
| Test set:     | <i>web1 + web2</i> | 218 tokens | 110:108 |

|   |             | <i>web_1+2</i> : web1 & web2 (web corpus) |                    |                    |
|---|-------------|---|--------------------|--------------------|
| Feature set   | Baseline    | SVM (RBF)                                 | SVM (linear)       | LDA                |
| 1. Full set<br>(no. features = 308)                     | 50.5        | 79.8 (17.4)                               | 73.4 (24.7)        | --                 |
| 2. Automated feature selection C<br>(no. features = 43) | 50.5        | 83.9 (15.2)                               | 79.4 (20.6)        | --                 |
| 3. Automated feature selection D<br>(no. features = 18) | 50.5        | 81.7 (16.8)                               | 72.9 (27.1)        | --                 |
| 4. Hand-picked A<br>(no. features = 3)                  | <b>50.5</b> | <b>89.4 (9.8)</b>                         | <b>88.5 (10.3)</b> | <b>88.1 (10.9)</b> |
| 5. Hand-picked B<br>(no. features = 4)                  | 50.5        | 85.8 (12.9)                               | 88.5 (10.3)        | 88.1 (10.9)        |

- Performance did not vary significantly between the declarative and interrogative contexts.

# Classification Experiment 2

- Again, predictions of prominence in comparative clauses confirmed
- Again, models without f0 information performed on par with those lacking it
  - > consistent with stress-first theories of focus
- Classifiers trained on lab data performed on par with classifiers trained on web data
  - > lab data have sufficient variability to train classifiers
- Are *these* accuracy rates reasonable?
  - > perception experiment in progress

# Method for future studies

- Find common grammatical or lexical contexts that trigger representations with different prosodic realization, according to relatively well-understood and well-supported theory.
- Correlate the semantic-grammatical categories directly with the speech signal using machine learning
  - Question: How successful are classifiers that lack information about phonemic/morphemic categories (e.g. H\*, L+H\*)?

# Future directions

- Simple, interactive GUI for aggregation, downloading and processing for web harvest method
- Full comparative paradigm (e.g. *than he did, than you do*)
  - Which acoustic cues are robust across word and vowel type?
- Second occurrence focus (cf. Howell in prep.)
  - debate over the semantics of focus in contexts favoring pitch-reduction
- Constructions alleged to be inherently contrastive (e.g. in MY opinion, NEXT friday, the President himSELF)
- Constructions alleged to use a particular accent type (e.g. for ONE thing, the ONE thing)

# Distribution of datasets

Audio snippets can probably be distributed under fair use.

<http://confluence.cornell.edu/display/prosody/Prosody+Datasets>

File Edit View History Bookmarks Tools Help

cornell.edu https://confluence.cornell.edu/display/prosody/Prosody+Datasets

Prosody Datasets - Confluence

Prosody Datasets







## Prosody Datasets

Added by [Marina Tokman](#), last edited by [Mats Rooth](#) on Aug 31, 2009 ([view change](#))

Labels  
(None)

We are harvesting datasets bearing on the form and meaning of prosody from web sources. This Wiki will make our methods and results available to the research and education community. The datasets are relevant to Linguistics, the Philosophy of Language, Psycholinguistics, and Artificial Intelligence. This is our prototype *than I did* dataset.

### "than I did"

| filenumber | class | soundfile   | transcription   | source  |
|------------|-------|---|---|---|
| 001        | X     |    | That's the whole beastiality angle, right? / Sort of. It's like cuddly kitties. / Well, it's more <b>than I did</b> , in that I loved animals and Disney's Robin Hood as a kid.       | no longer available: <a href="http://www.channelchaos.com/">http://www.channelchaos.com/</a>  |
| 002        | ns    |   | I made more money than you for a while. / I think I've made more total money than you. / But, you made a very small amount more <b>than I did</b> . Now I make much more than you do. | <a href="http://media.libsyn.com/media/geeknights/20070614.mp3">http://media.libsyn.com/media/geeknights/20070614.mp3</a>                     |
| 003        | s     |  | The aquarium got more attention <b>than I did</b> . / But, you made a very small amount more <b>than I did</b> . Now I make much more than you do.                                    | <a href="http://podcast.wckq.com/wckq1/524088.mp3">http://podcast.wckq.com/wckq1/524088.mp3</a>   |
| 004        | s     |  | Tom actually said it a lot better <b>than I did</b> .   | <a href="http://www.patriots.com/StreamFile.cfm?f=20070613_neplor.mp3">http://www.patriots.com/StreamFile.cfm?f=20070613_neplor.mp3</a>       |
| 005        | s     |  | There were a lot of photographers who would shoot more <b>than I did</b> .  | <a href="http://www.wfhh.org/audio/download/5922/Interchange20070612.mp3">http://www.wfhh.org/audio/download/5922/Interchange20070612.mp3</a> |
| 006        | ns    |  | I should have liked that song I   | no longer available: <a href="http://podcast.971freem.com/klx1/520789.mp3">http://podcast.971freem.com/klx1/520789.mp3</a>                    |