Acoustic Classification of Focus in a Web Corpus of Comparatives

Jonathan Howell, Mats Rooth, Michael Wagner
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]₂
    B: Yes. In fact, [the WHOLE TEAM₇ 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]$_2$
  than [ l$_F$ stayed d long ] $\sim$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

<table>
<thead>
<tr>
<th>Datasource</th>
<th>Web</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>search.everyzing.com (now defunct)</td>
<td>web1</td>
<td>91 true tokens</td>
</tr>
<tr>
<td>multimedia.play.it (now defunct)</td>
<td>web2</td>
<td>127 true tokens</td>
</tr>
<tr>
<td>mediasearch.wnyc.org</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ramp (formerly Everyzing) is a commercial audio indexing service using ASR
Content includes radio and tv programs, podcasts
WAMU: The Diane Rehm Show Podcast - Hans Blix: "Why Nuclear Disarmament Matters" (Boston Review)

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

Your Search Term: "than I did"
Key Terms: "US senate" "Julie Andrews" "Iran Iraq" "Iraq war"

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 -- ...
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 you 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 38:27

"...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 okay 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

- **Target** $w_1 + \ldots + w_n$
- **Number of hits** $N$

- **Experiment**
  - **awk program**
  - **script calling curl**
    - **run**
  - **N/10 html pages with 10 hits each**
    - **run**
  - **N html files**
    - **awk**
  - **URLs of mp3**
    - **awk**
  - **Praat textgrid**

- **Short mp3s**
  - **time offsets**
    - **awk**
  - **cutmp3**
  - **Praat script**
    - **for each hit**
      - 10 second wav file, Praat textgrid with Everyzing’s starts for $w_1, \ldots w_n$. 
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) He stayed longer than I_F did</td>
<td>s class</td>
</tr>
<tr>
<td>antecedent: he stayed x long</td>
<td></td>
</tr>
<tr>
<td>(2) I should have liked that song more than I did_F</td>
<td>ns class</td>
</tr>
<tr>
<td>antecedent: I should have liked that song x much.</td>
<td></td>
</tr>
<tr>
<td>(3) I understand even less than I did before_F</td>
<td>ns class</td>
</tr>
<tr>
<td>antecedent: I understand even x little</td>
<td></td>
</tr>
</tbody>
</table>
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
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
Separating hyperplanes
SVM

Maximum margin

Optimal hyperplane

Measure 1

Measure 2
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dur_V2</strong></td>
<td>duration of <em>l</em></td>
</tr>
<tr>
<td><strong>f0_ratio</strong></td>
<td>ratio of f0 maxima in <em>l</em> and <em>did</em></td>
</tr>
<tr>
<td><strong>f1f2_40_V2</strong></td>
<td>f2-f1 at 40% into <em>l</em></td>
</tr>
<tr>
<td><strong>f1f2_50_V2</strong></td>
<td>f2-f1 at 50% into <em>l</em></td>
</tr>
</tbody>
</table>
Automated Feature Selection B

Scatter Plot Matrix
e.g. Hand-picked A

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dur_V2</td>
<td>duration of /</td>
</tr>
<tr>
<td>dur_C3</td>
<td>duration of 1st stop closure in did</td>
</tr>
<tr>
<td>f0_ratio</td>
<td>ratio of f0 maxima in / and did</td>
</tr>
<tr>
<td>f1f2_50_V2</td>
<td>f2-f1 at 50% into /</td>
</tr>
</tbody>
</table>
Evaluation of classifier performance

• Baseline accuracy
  • # tokens in largest class of test set
    # tokens in both classes in test set

• Generaliation accuracy
  • # of tokens in test set accurately classified
    # of tokens in test set

• Balanced error rate
  \[
  \frac{\# \text{incorrect “s”}}{\# \text{total “s”}} \times \frac{\# \text{incorrect “ns”}}{\# \text{total “ns”}} \times \frac{1}{2} \times 100
  \]
Classifier Performance

Training set: web1  91 tokens  46:45
Test set: web2  127 tokens  62:65
<table>
<thead>
<tr>
<th>Feature set</th>
<th>Baseline</th>
<th>SVM (radial kernel)</th>
<th>SVM (linear kernel)</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Full set (no. features = 308)</td>
<td>51.2</td>
<td>82.7 (14.3)</td>
<td>85.0 (14.2)</td>
<td>n/a due to collinearity</td>
</tr>
<tr>
<td>2. Automated feature selection A (no. features = 16)</td>
<td>51.2</td>
<td>89.8 (10.1)</td>
<td>89.0 (10.3)</td>
<td>90.6 (9.0)</td>
</tr>
<tr>
<td>3. Automated feature selection B (no. features = 4)</td>
<td>51.2</td>
<td>86.6 (13.1)</td>
<td>90.6 (9.2)</td>
<td>87.4 (11.9)</td>
</tr>
<tr>
<td>4. Hand-picked A (no. features = 4)</td>
<td>51.2</td>
<td>92.9 (6.5)</td>
<td>92.1 (7.1)</td>
<td>91.3 (7.7)</td>
</tr>
<tr>
<td>5. Hand-picked B (no. features = 3)</td>
<td>51.2</td>
<td>91.3 (7.7)</td>
<td>92.1 (7.1)</td>
<td>91.3 (7.7)</td>
</tr>
</tbody>
</table>
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 f0 information performed on par with models lacking it

Hand-picked A: dur_V2, dur_C3, f1f2Time50_V2, f0_ratio
Hand-picked B: dur_V2, dur_C3, f1f2Time50_V2

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Baseline</th>
<th>SVM (radial)</th>
<th>SVM (linear)</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Hand-picked A</td>
<td>51.2</td>
<td>92.9 (6.5)</td>
<td>92.1 (7.1)</td>
<td>91.3 (7.7)</td>
</tr>
<tr>
<td>(no. features = 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Hand-picked B</td>
<td>51.2</td>
<td>91.3 (7.7)</td>
<td>91.3 (7.7)</td>
<td>90.6 (8.3)</td>
</tr>
<tr>
<td>(no. features = 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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)

  \[
  \begin{array}{cccccccc}
  \text{p.a.} & \{H^*, L^*, L+H^*, \ldots\} \\
  x & x & x & x & x & x & x & x \\
  x & x & x & x & x & x & x & x \\
  x & x & x & x & x & x & x & x \\
  \end{array}
  \]

• 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: \textit{lab} \hspace{1cm} 394 tokens \hspace{1cm} 193:201

Test set: \textit{web1+web2} \hspace{1cm} 218 tokens \hspace{1cm} 110:108
<table>
<thead>
<tr>
<th>Feature set</th>
<th>Baseline</th>
<th>SVM (RBF)</th>
<th>SVM (linear)</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Full set (no. features = 308)</td>
<td>50.5</td>
<td>79.8 (17.4)</td>
<td>73.4 (24.7)</td>
<td>--</td>
</tr>
<tr>
<td>2. Automated feature selection C (no. features = 43)</td>
<td>50.5</td>
<td>83.9 (15.2)</td>
<td>79.4 (20.6)</td>
<td>--</td>
</tr>
<tr>
<td>3. Automated feature selection D (no. features = 18)</td>
<td>50.5</td>
<td>81.7 (16.8)</td>
<td>72.9 (27.1)</td>
<td>--</td>
</tr>
<tr>
<td>4. Hand-picked A (no. features = 3)</td>
<td><strong>50.5</strong></td>
<td><strong>89.4 (9.8)</strong></td>
<td><strong>88.5 (10.3)</strong></td>
<td><strong>88.1 (10.9)</strong></td>
</tr>
<tr>
<td>5. Hand-picked B (no. features = 4)</td>
<td>50.5</td>
<td>85.8 (12.9)</td>
<td>88.5 (10.3)</td>
<td>88.1 (10.9)</td>
</tr>
</tbody>
</table>

- 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
Prosody Datasets

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"

<table>
<thead>
<tr>
<th>filename</th>
<th>class</th>
<th>soundfile</th>
<th>transcription</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>X</td>
<td></td>
<td>That's the whole beastiality angle, right? / Sort of, it's like cuddly kittens. / Well, it's more than I did, in that I loved animals and Disney's Robin Hood as a kid.</td>
<td>no longer available: <a href="http://www.channelchaos.com/">http://www.channelchaos.com/</a></td>
</tr>
<tr>
<td>002</td>
<td>ns</td>
<td></td>
<td>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 than I did. Now I make much more than you do.</td>
<td><a href="http://media.librsyn.com/media/geeknights/20070614.mp3">http://media.librsyn.com/media/geeknights/20070614.mp3</a></td>
</tr>
<tr>
<td>003</td>
<td>s</td>
<td></td>
<td>The aquarium got more attention than I did. / But, you made a very small amount more than I did. Now I make much more than you do.</td>
<td><a href="http://podcast.wrcq.com/wrcq1/524088.mp3">http://podcast.wrcq.com/wrcq1/524088.mp3</a></td>
</tr>
<tr>
<td>004</td>
<td>s</td>
<td></td>
<td>Tom actually said it a lot better than I did.</td>
<td><a href="http://www.patriots.com/StreamFile.cfm?f=20070613_neplor.mp3">http://www.patriots.com/StreamFile.cfm?f=20070613_neplor.mp3</a></td>
</tr>
<tr>
<td>005</td>
<td>s</td>
<td></td>
<td>There were a lot of photographers who would shoot more than I did.</td>
<td><a href="http://www.wfhb.org/audio/download/5922/Interchange20070612.mp3">http://www.wfhb.org/audio/download/5922/Interchange20070612.mp3</a></td>
</tr>
<tr>
<td>006</td>
<td>ns</td>
<td></td>
<td>I should have liked that song I</td>
<td>no longer available: <a href="http://podcast.974freefm.com/klsa1/520789.mp3">http://podcast.974freefm.com/klsa1/520789.mp3</a></td>
</tr>
</tbody>
</table>