Towards the Golden Age of Speech and Language Science

Mark Liberman
University of Pennsylvania

ABSTRACT:

For the sciences of speech and language, the 21st century promises to bring the kind of progress that the 17th century brought to the physical sciences.

Our telescopes and microscopes, our alembics and Pneumatical Engines, are today's vast archives of digital text and speech, along with new analysis techniques and inexpensive networked computation.

However, the scientific use of these new instruments remains mainly exploratory and potential. There are several critical problems for which we have at best partial solutions; and like our 17th-century predecessors, we need to unlearn some old ideas on the way to learning new ones.

Focusing especially on Henry Sweet's own interests in phonetics and in the history of English, this talk will discuss some of the barriers to be overcome, present some successful examples, and speculate about future directions.

According to the National Academy of Sciences:

We see that the computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe these can be aptly compared with the challenges, problems, and insights of particle physics. Certainly, language is second to no phenomenon in importance. And the tools of computational linguistics are considerably less costly than the multibillion-volt accelerators of particle physics. The new linguistics presents an attractive as well as an extremely important challenge.

There is every reason to believe that facing up to this challenge will ultimately lead to important contributions in many fields.

Language and Machines: Computers in Translation and Linguistics
Report by the Automatic Language Processing Advisory Committee (**ALPAC**), National Academy of Sciences

Two wrinkles

(1) ALPAC 's main recommendation
was to de-fund Machine Translation research.

. . . wait, what?

(2) And, the ALPAC report came out in **1966** (!)

so 44 years later,
where's the QCD of linguistics?

The plan vs. the reality

- ALPAC ‘s idea:
  1. computers → new language science
  2. language science → language engineering

- What actually happened:
  1. computers → new language engineering
  2. engineering → new language science (???)

Why 2011 is like 1611

- Telescope: invented 1608

  Galileo 1609, Kepler 1611, Newton 1668

- Microscope: invented 1590

  Hooke 1665, Leeuvenhoek 1674

Instruments that opened new worlds to view

Today

- Convenient data creation
  - acoustic perception and production
  - articulatory measurement (ultrasound, EMMA)
  - of conversation by text, voice, video
- Automated bookkeeping
- Easy search and statistical modeling
- Easy sharing and re-use

Wider horizons: Found data

- Digital networks are flooded with
  - trillions of words of text
  - millions of hours of speech
  - billion-node networks
    of social and topical connections
- With cheap, smart sensors,
  and compact, ubiquitous, networked interfaces,
  any human activity can be “instrumented”
  and added to the flood
- A more and more complete digital record
  of human social interaction

That's what they all say . . .

Progress in any science depends on a combination of improved observation, measurement, and techniques. The cheap computing of the past two decades means there has been a tremendous increase in the availability of economic data and huge strides in econometric techniques. As a result, economics stands at the verge of a golden age of discovery.

-Diane Coyle, "Economics on the Verge of a Golden Age", *The Chronicle of Higher Education*, March 12, 2010

But in fact…

- An Age of Discovery is here
  - in all of the sciences
  - for similar reasons
- Concepts, techniques and results flow across disciplinary boundaries
- The digital traces of human communication will be increasingly important
  - in the “human” sciences
  - and beyond…

In 2011 as in 1611

Science needs theory -- but

"Sometimes you can observe a lot just by watching"

-Yogi Berra

Breakfast experiments

- Our “telescope and microscope” are
  - Accessible collections of speech and text (found or created)
  - Computer algorithms for
    - analyzing speech and text
    - aligning speech and text
    - collecting, displaying, modeling
- When we point these new instruments in almost any direction, we see interesting new things
- This is so easy and fast that we can often do an “experiment” on a laptop over breakfast.

These quick looks are not a substitute for serious research.

But they illustrate the power of our new tools,
and allow us to explore interesting new directions quickly.

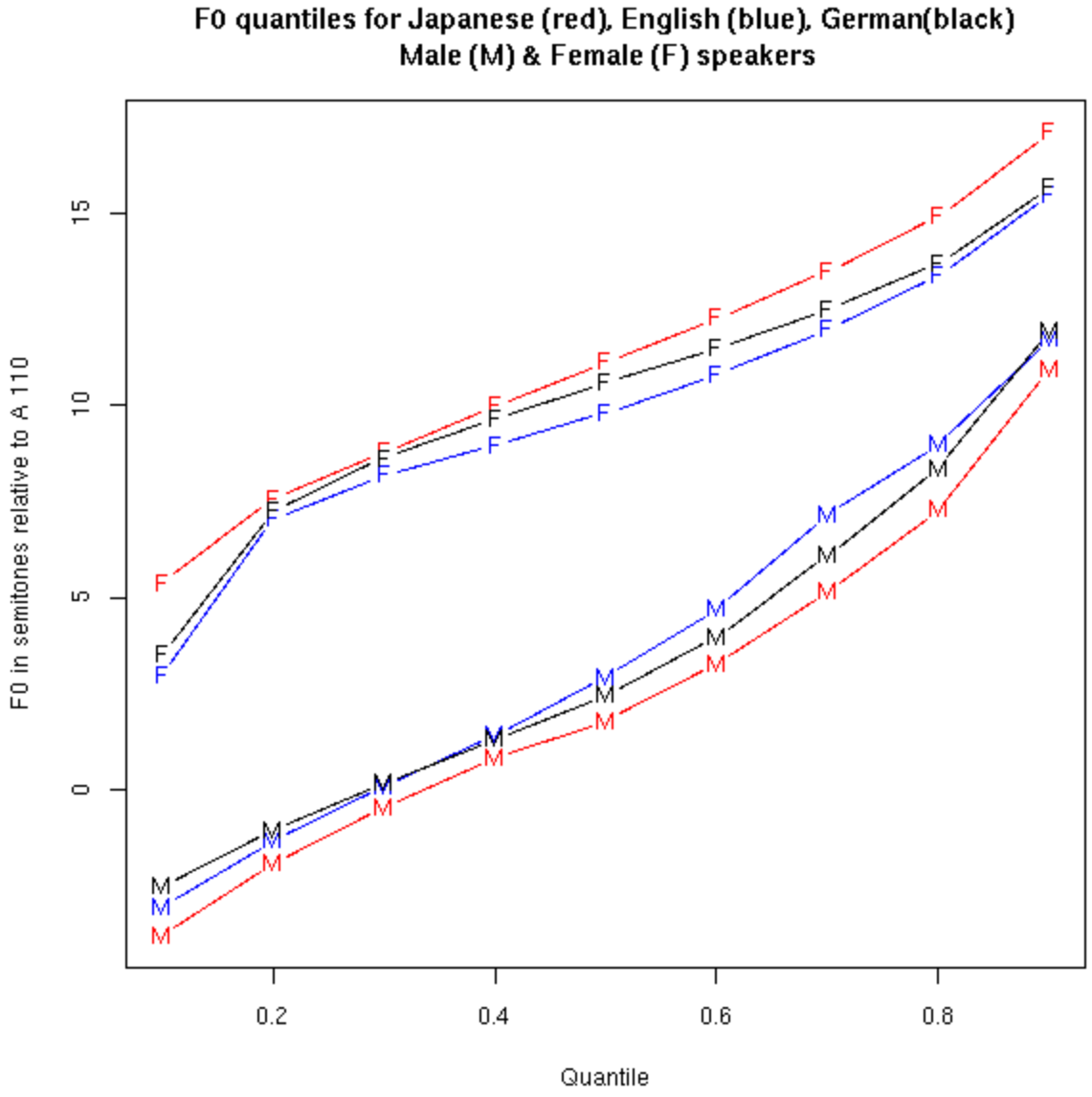
(All of the cited Breakfast Experiments™
were published in Language Log)

Seven One-Hour Explorations

- Do Japanese speakers show more gender polarization in pitch than American speakers?
- Do American women talk more (and faster) than men?
- How does word duration vary with phrase position?
- “Eigentones” for F0 analysis
- How does local speaking rate vary in the course of a conversation?
- How does disfluency vary with sex and age?
- “you know”/”I mean” ratio over the lifespan

One-hour exploration #1

- Gender polarization in conversational speech
- Question: are Japanese men and women more polarized (more different) in pitch than Americans or Europeans?
- Method:
  - Pitch-track published telephone conversations
  - LDC “Call Home” publications for Japanese, U.S. English, German
    - Collected 1995-1996 , published 1996-1997
    - about 100 conversations per language
  - Compare quantiles of pooled values
    (about 2 million numbers per sex/culture combination)
- Answer: yes, apparently so.



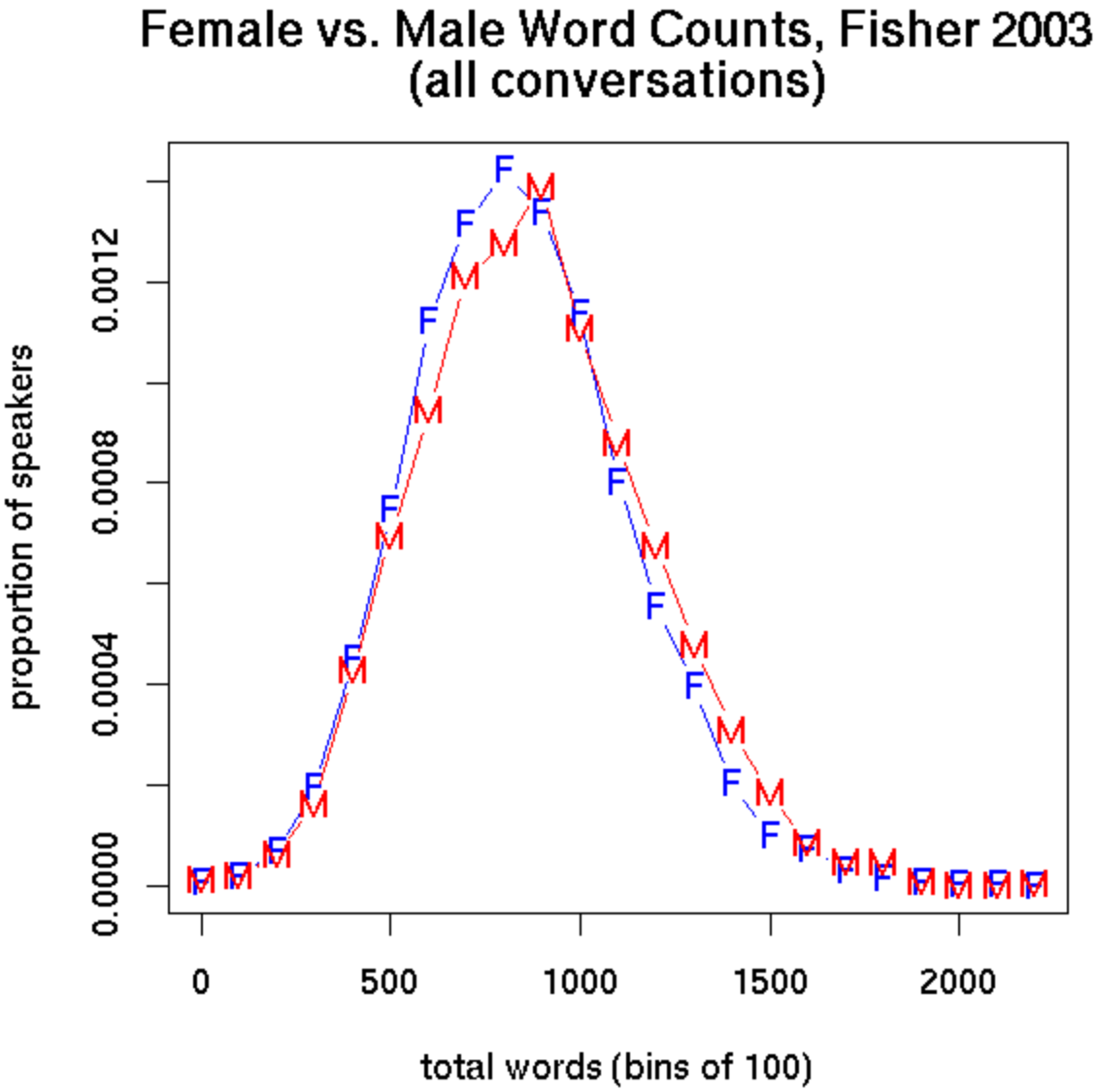
Data from CallHome M/F conversations; about 1M F0 values per category.

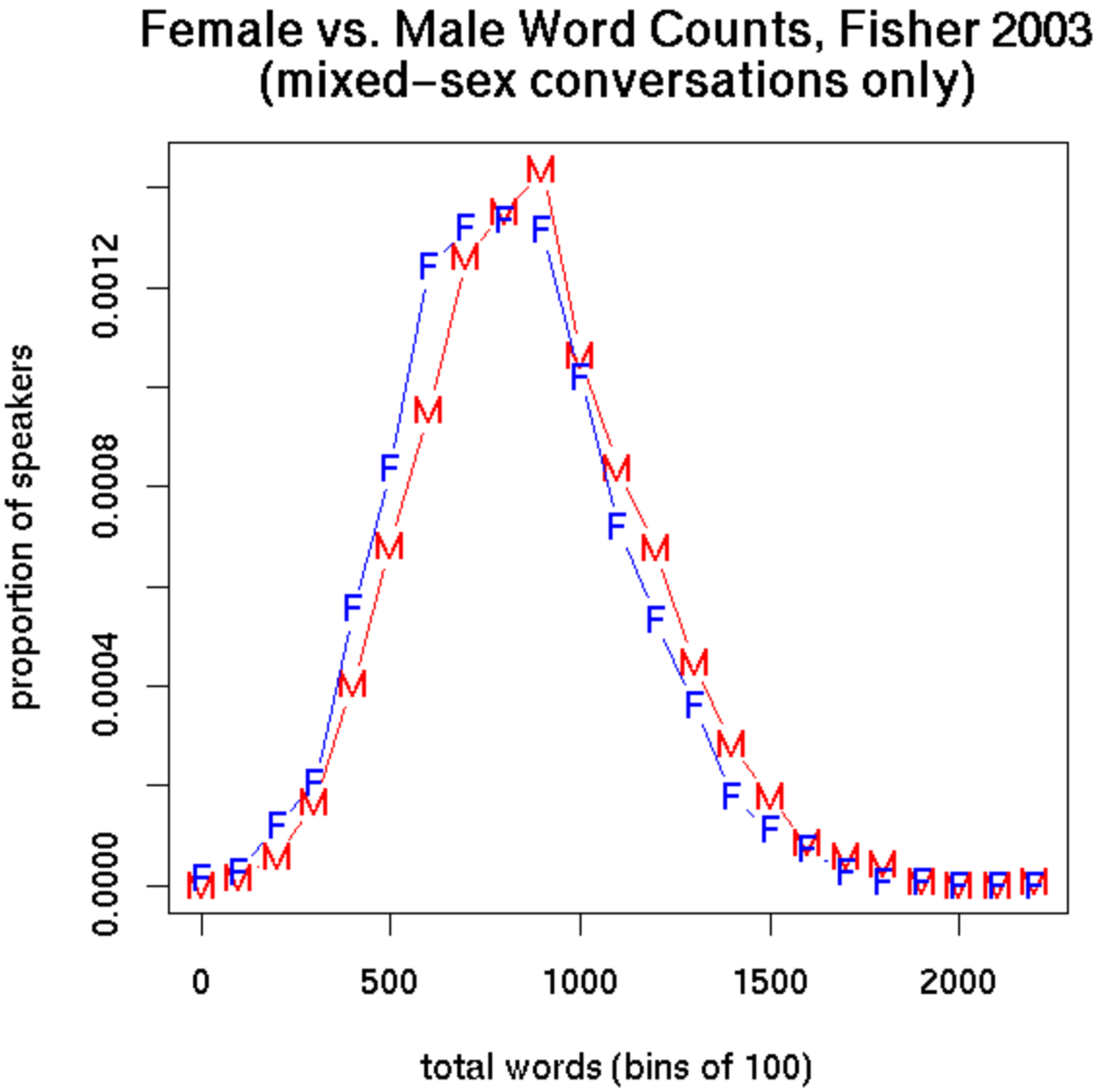
As usual, more questions:

- Other cultures and languages
- Effects of speaker's age
- Effects of relationship between speakers, nature of discussion
- Formal vs. conversational speech
- Effects of social class, region

One-hour exploration #2a

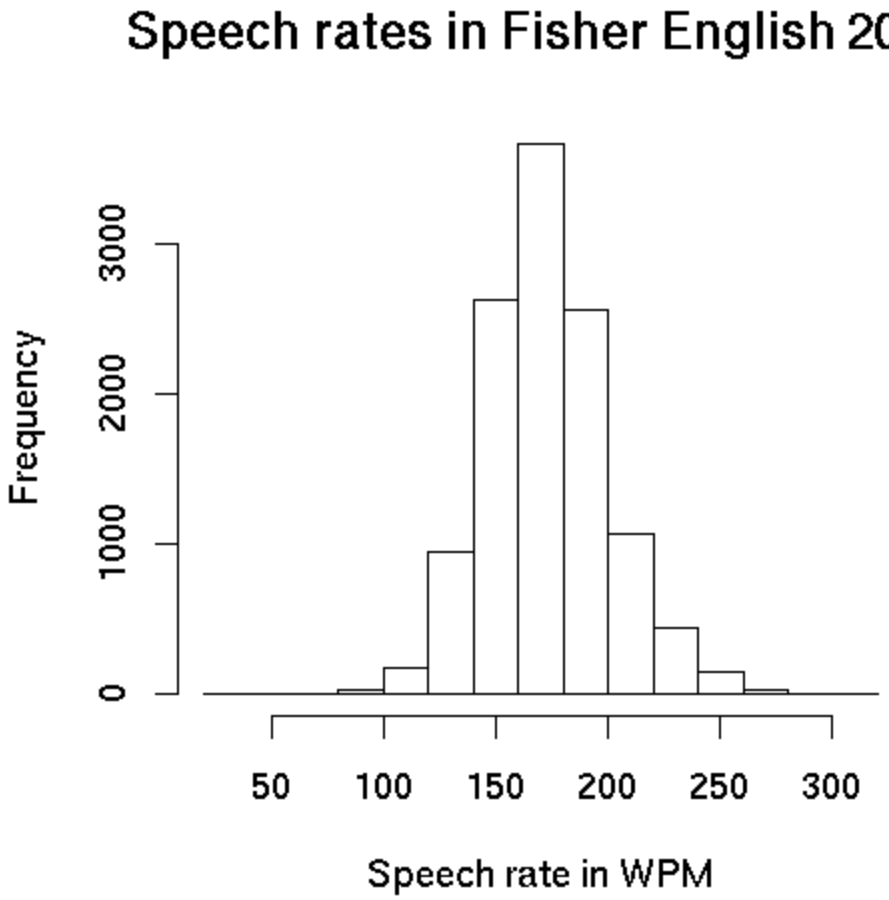
- Sex differences in conversational word counts
- Question: Do women talk more than men?
- Method: Count words in “Fisher” transcripts
  - Conversational telephone speech
    - Collected by LDC in 2003
    - 5,850 ten-minute conversations
      - 2,368 between two women
      - 1,910 one woman, one man
      - 1,572 between two men
- Answer: No.


Female vs. Male Word Counts, Fisher 2003
(all conversations)
proportion of speakers
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
total words (bins of 100)


Female vs. Male Word Counts, Fisher 2003
(mixed−sex conversations only)
proportion of speakers
0.0000
0.0004
0.0008
0.0012
0
500
1000
1500
2000
total words (bins of 100)

One-hour experiment #2b

- Sex differences in conversational speaking rates
- Question: Do women talk faster than men?
- Method: Words and speaking times in Fisher 2003
- Answer: No.



(11,700 conversational sides; mean speaking rate=173 wpm, sd=27)
(Male mean 174.3, female 172.6: difference 1.7, effect size d=0.06)

One-Hour Experiments 3a & 3b

- Phrasal modulation of speaking rate / duration
  - "final lengthening" is a well-established effect
  - first observed by Abbé J.-P. Rousselot before 1900

What's the independent variable??

- What is a "phrase"?
  - A syntactic unit?
  - A unit of information structure?
  - A unit of speech production?
- And what *are* those

  syntactic /

  information-structure /

  speech-production

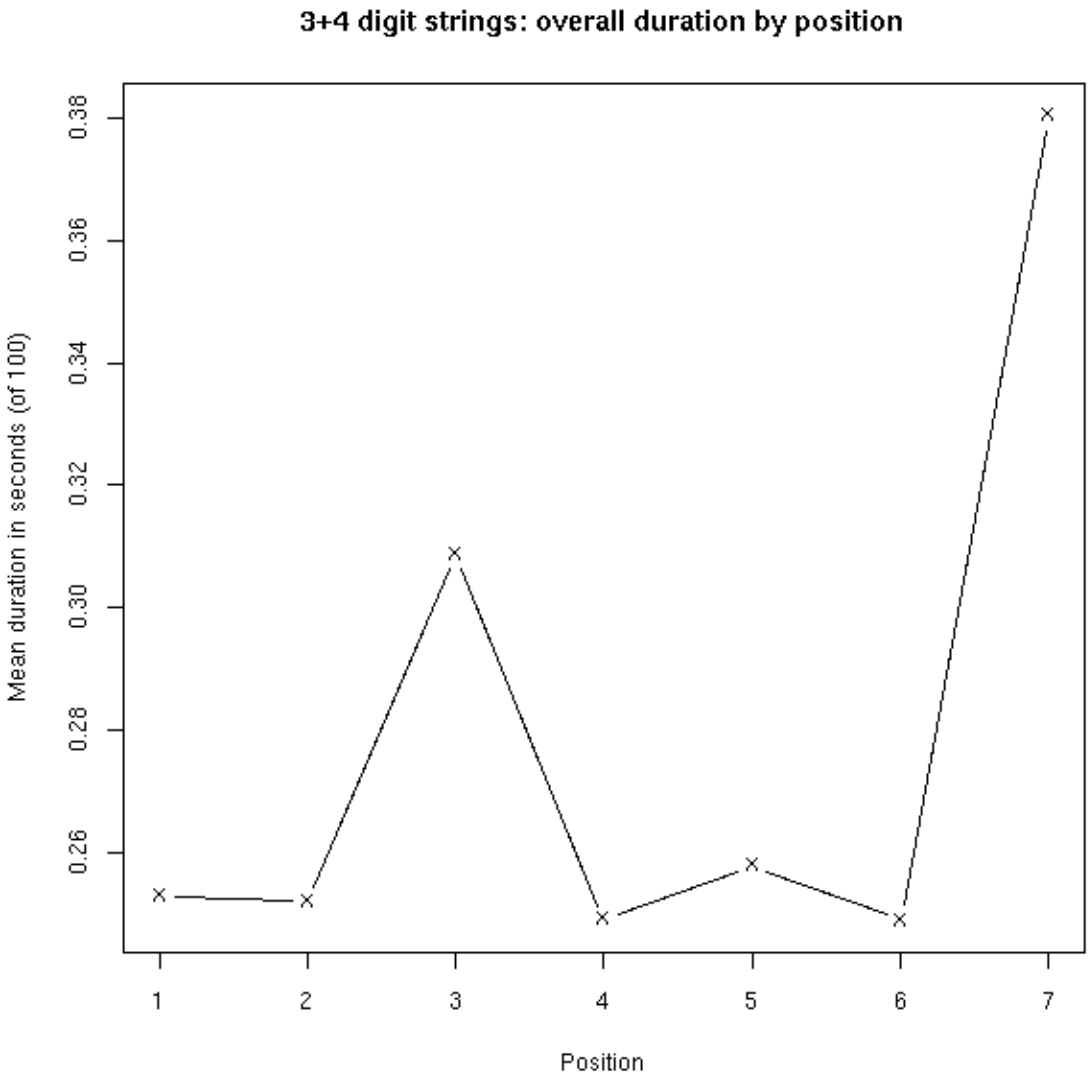
  units, anyway?

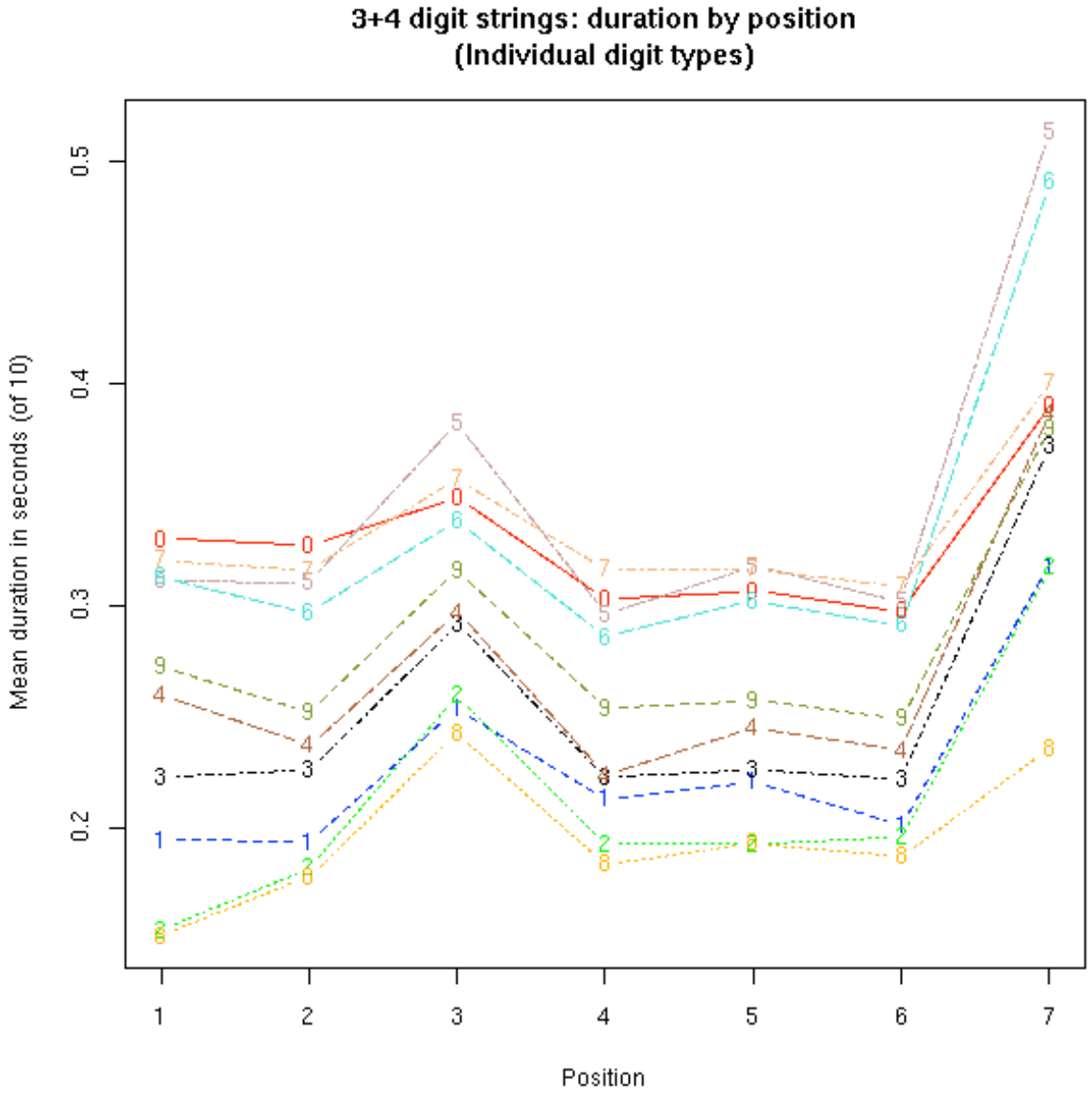
One approach…

- Random digit strings
  - With the structure implicit in real-life patterns
  - e.g. U.S. 3+4 = 7-digit phone number -- 868-6046 etc.
- We impose constraints on the data set:
  - Each digit occurs equally often in each position
  - Each pair of digits occurs equally often spanning each pair of positions
  - This requires 100 strings of whatever length
- Less than 5 minutes to record
- About 2 hours to segment by hand – (or segment automatically using forced alignment)

And this works in any language!

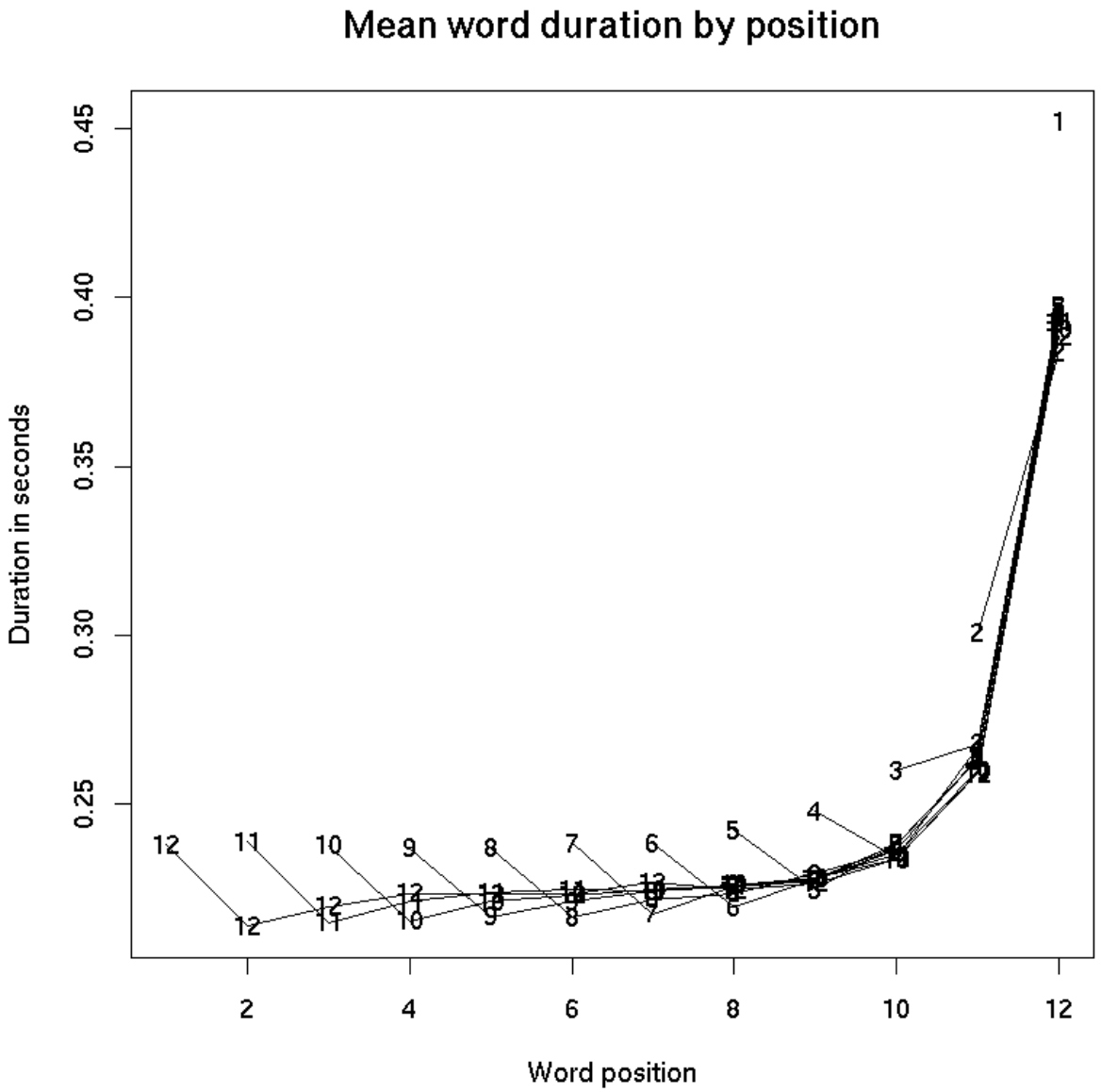
… except maybe Pirahã …


3+4 digit strings: overall duration by position
Mean duration in seconds (of 100)
0.26
0.28
0.30
0.32
0.34
0.36
0.38
1
2
3
4
5
6
7
Position

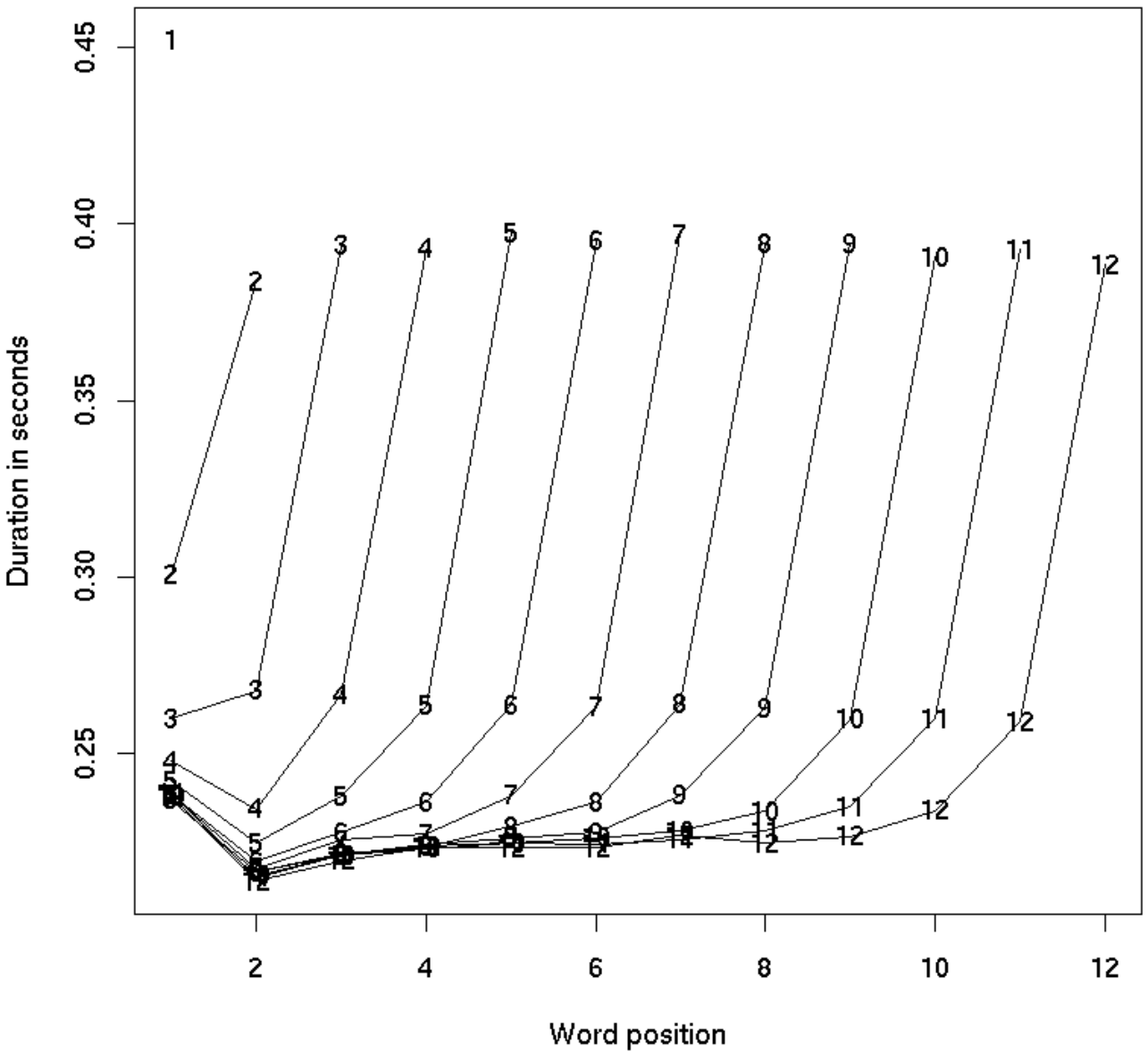

3+4 digit strings: duration by position
(Individual digit types)
Mean duration in seconds (of 10)
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
Position

Second version

- Method:
  word duration by position in “pause group”
  (defined as
  a stretch of speech
  without internal silence >100 msec)
- Data: Switchboard corpus
- Result: Amazingly regular (average) pattern



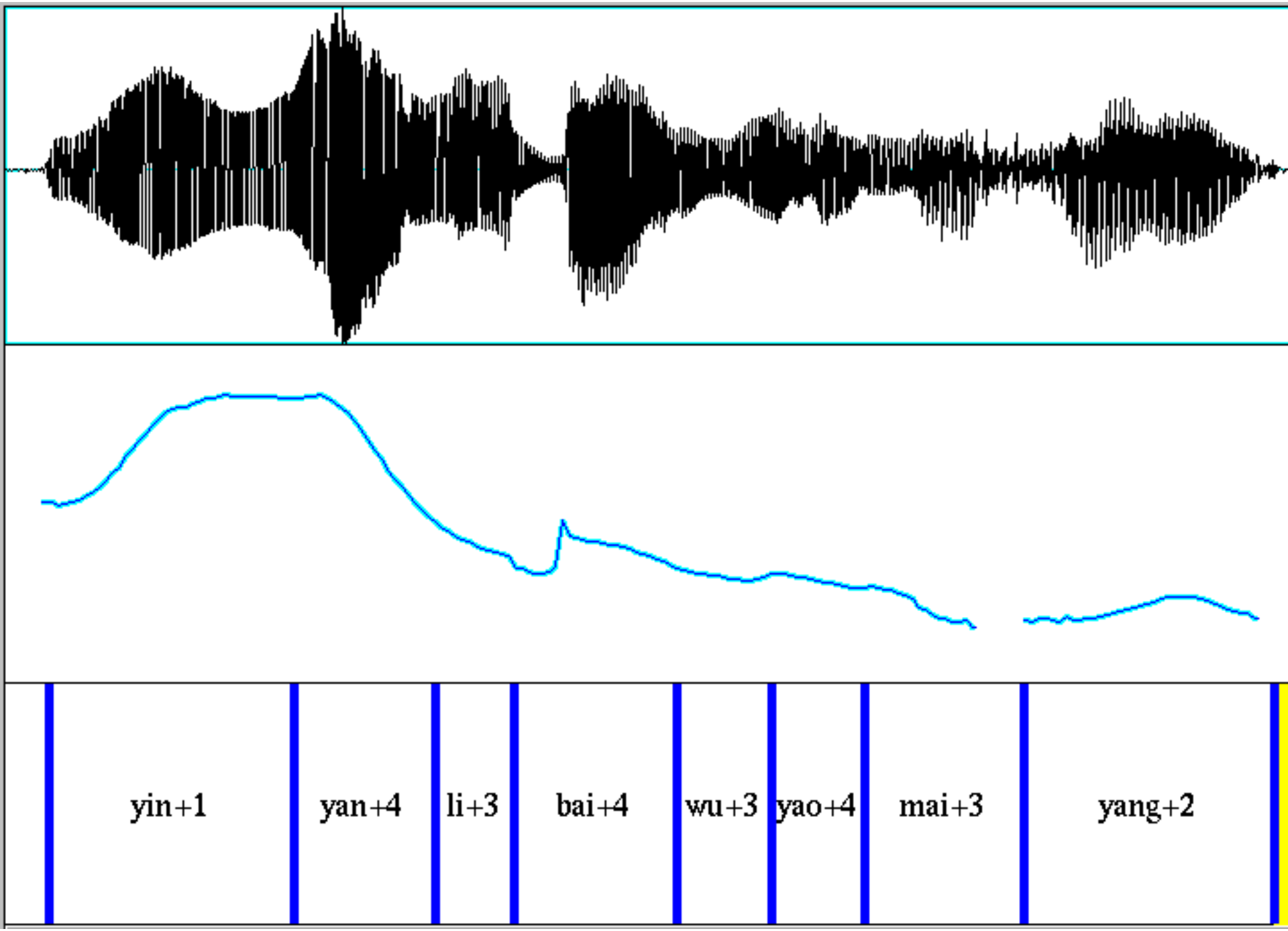
Data from Switchboard; phrases defined by silent pauses
(Yuan, Liberman & Cieri, ICSLP 2006)


Mean word duration by position
Duration in seconds
0.25
0.30
0.35
0.40
0.45
Word position
2
4
6
8
10
12

Breakfast experiment #5: Modeling Mandarin tone

Existing data from Jiahong Yuan 2004:

- 999 8-syllable utterances from 8 speakers
- Variation in tone sequence & focus
- Segmented into 7,992 syllables


yin+1
yan+4
li+3
bai+4
wu+3
yao+4
mai+3
yang+2

Approach:

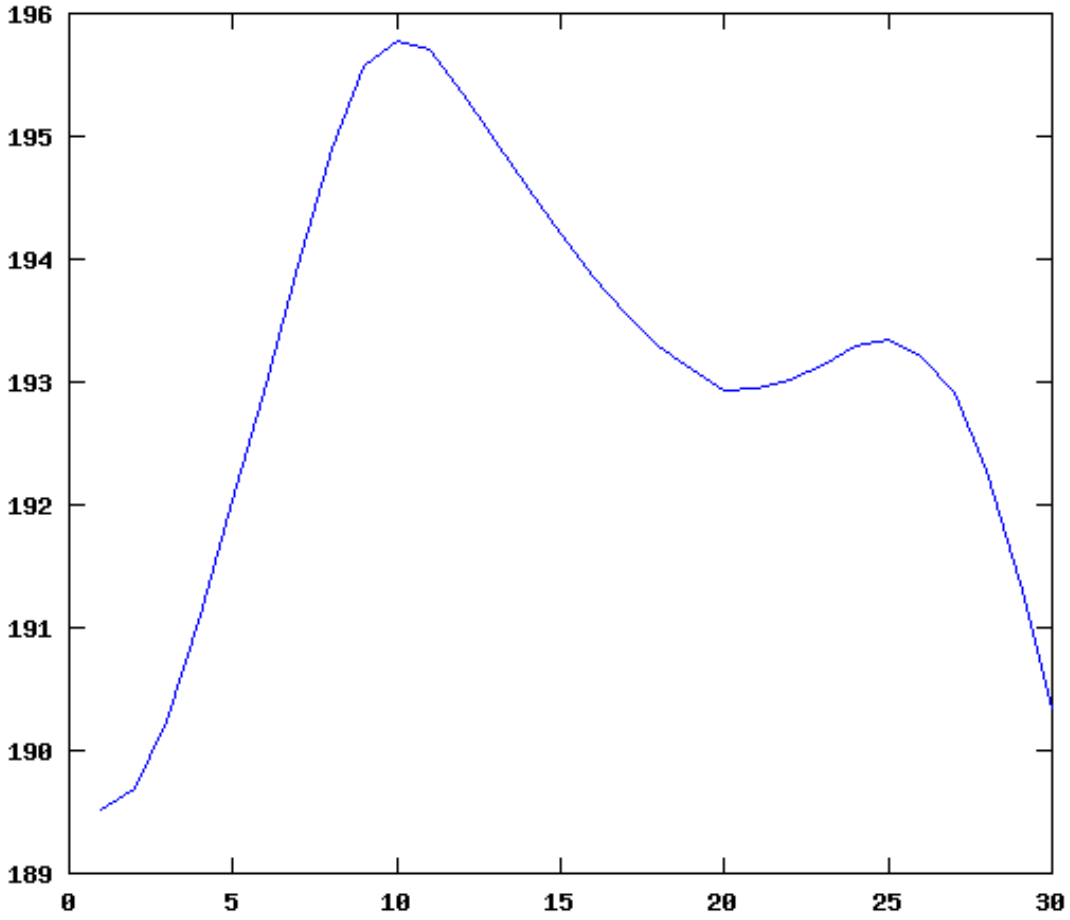
1. Time-normalize and resample F0 contours to 30 points per syllable

2. Look at mean values

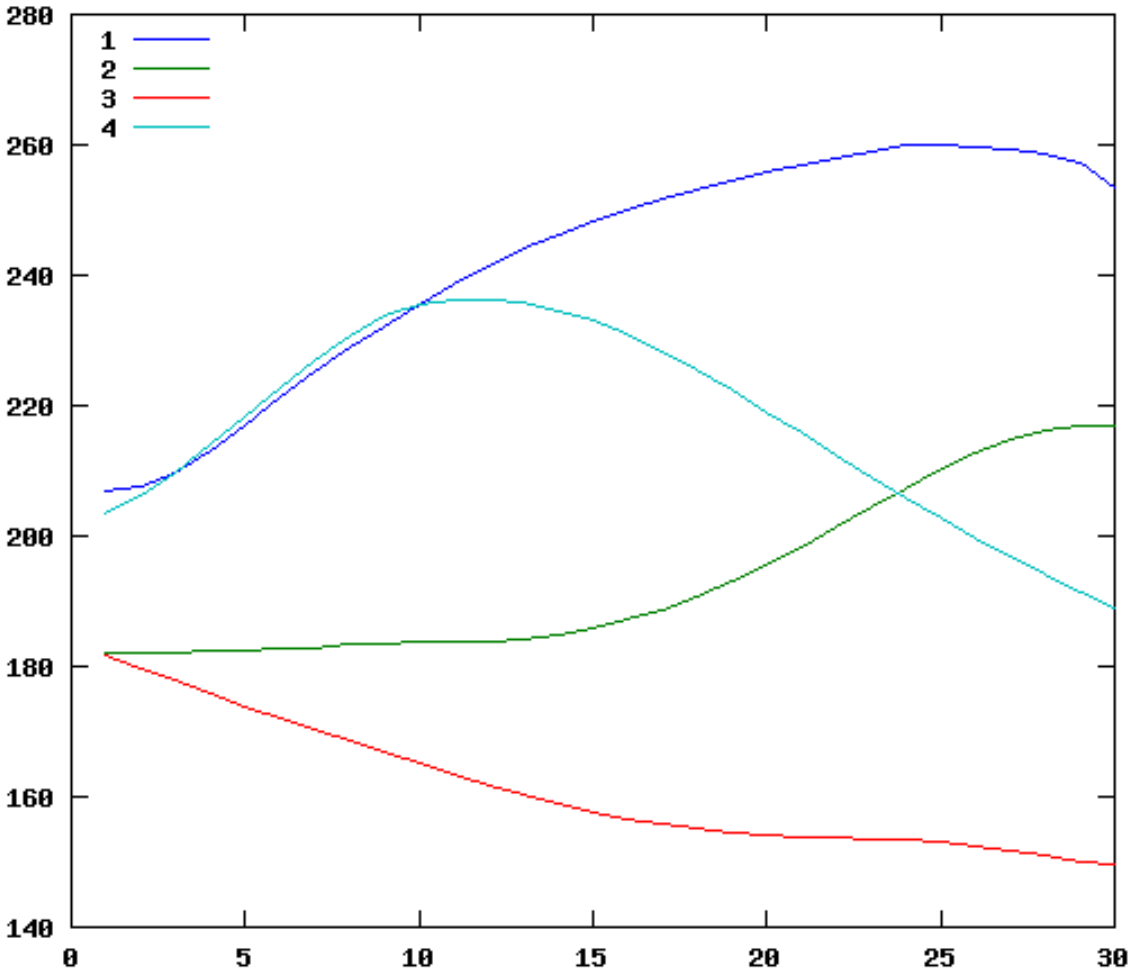
3. Try simple linear models:

e.g. Functional Principle Components Analysis

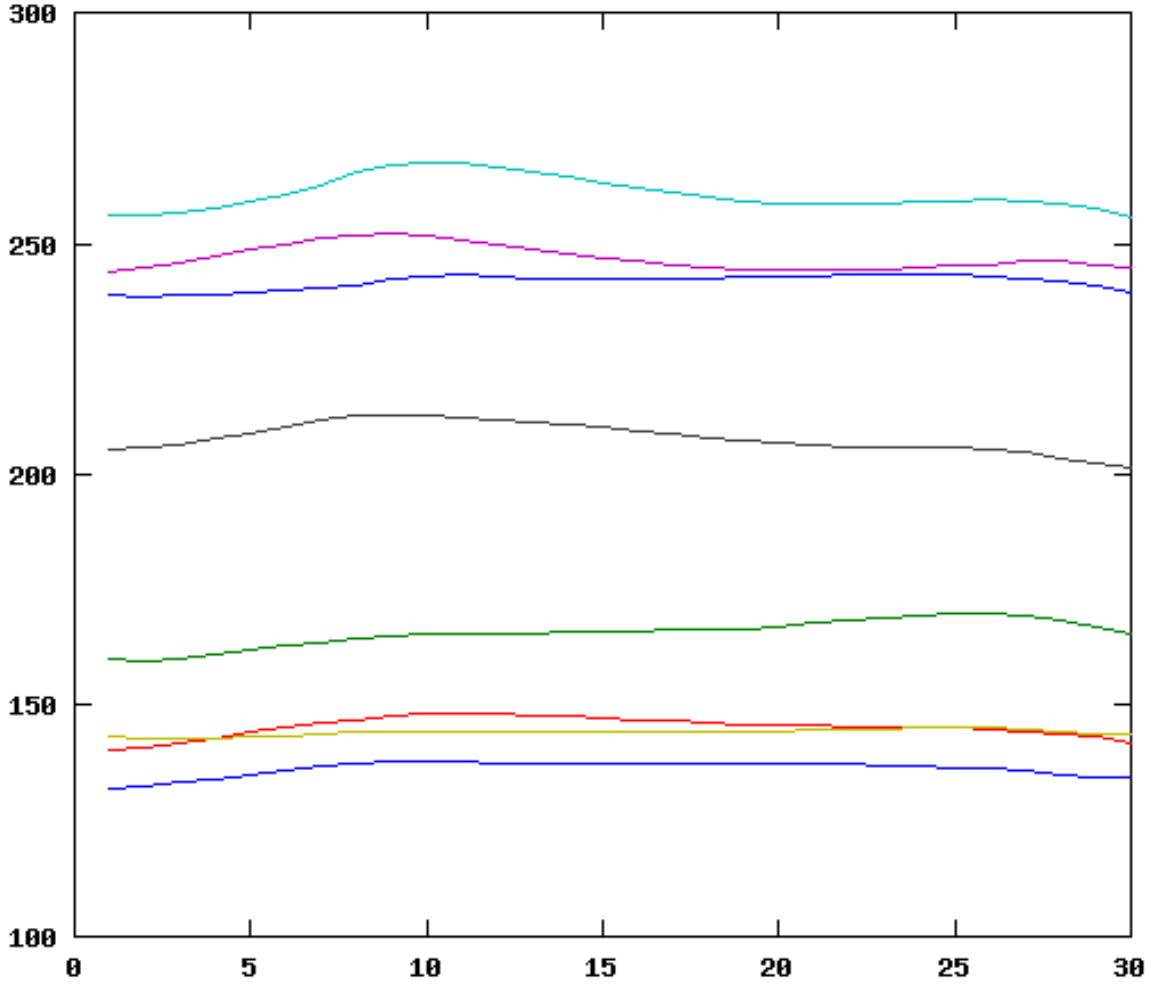
Overall mean syllabic contour:



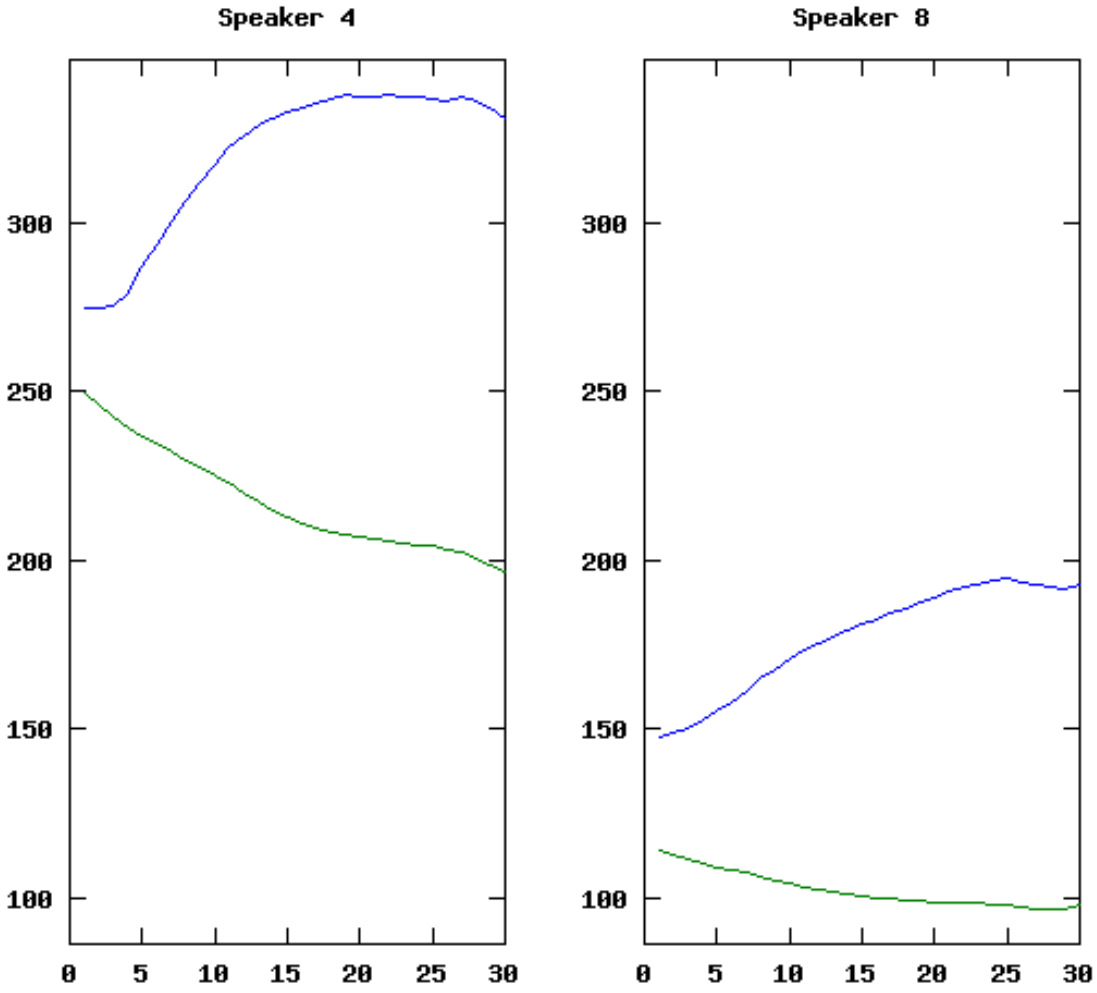
Mean contours for each of the four tones:



Speaker means:



Another view of speaker effects:

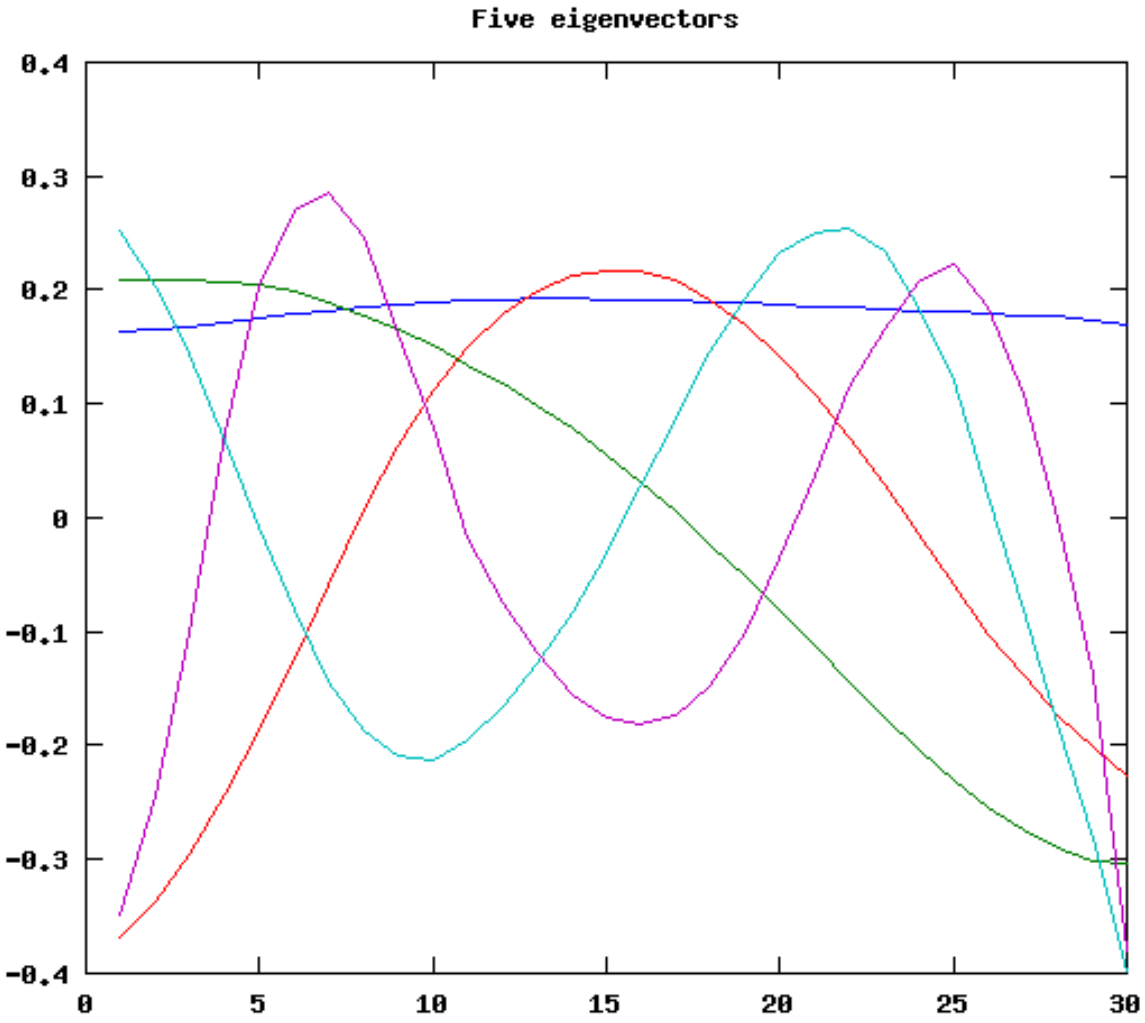


FPCA

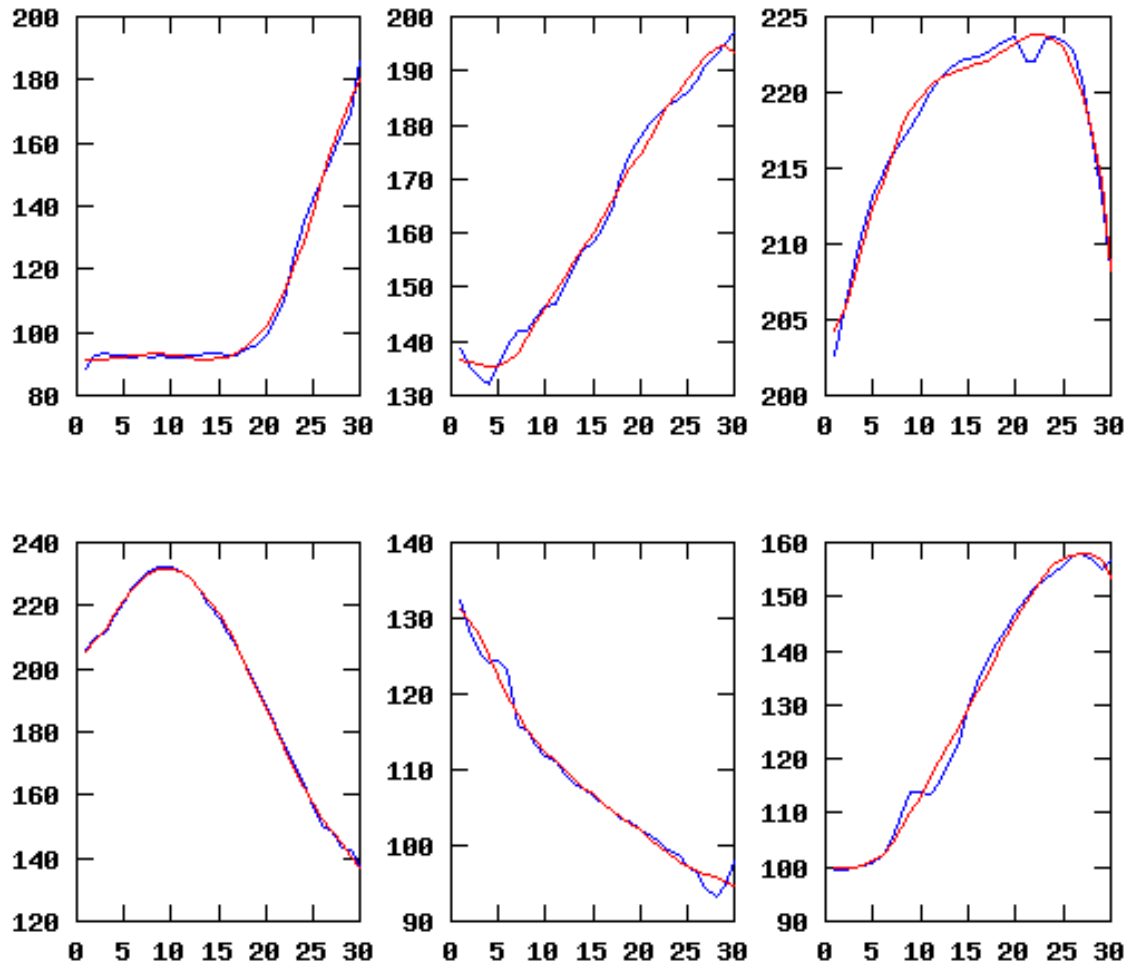
- Goal: Find new basis vectors that account for as much as possible of the variance in the F0 contours
- An old problem with a well-known solution:
  look at the eigenstructure of the covariance matrix.
- F0 data is matrix of 7,992 [syll] rows by 30 [F0] columns)
- Two lines of Matlab to do the analysis,
  two more to plot the first few FPCs:

```
f0cov = cov(f0data1);
[V L] = eig(f0cov);
% Note that eigenvectors are in columns of V
%  eigenvalues ordered from smallest to largest
basis5 = V(:,30:-1:26)';
plot(basis5'); title('Five eigenvectors')
```

FPCA basis vectors look like orthogonal polynomials . . .



And of course the “eigentones” work well . . .

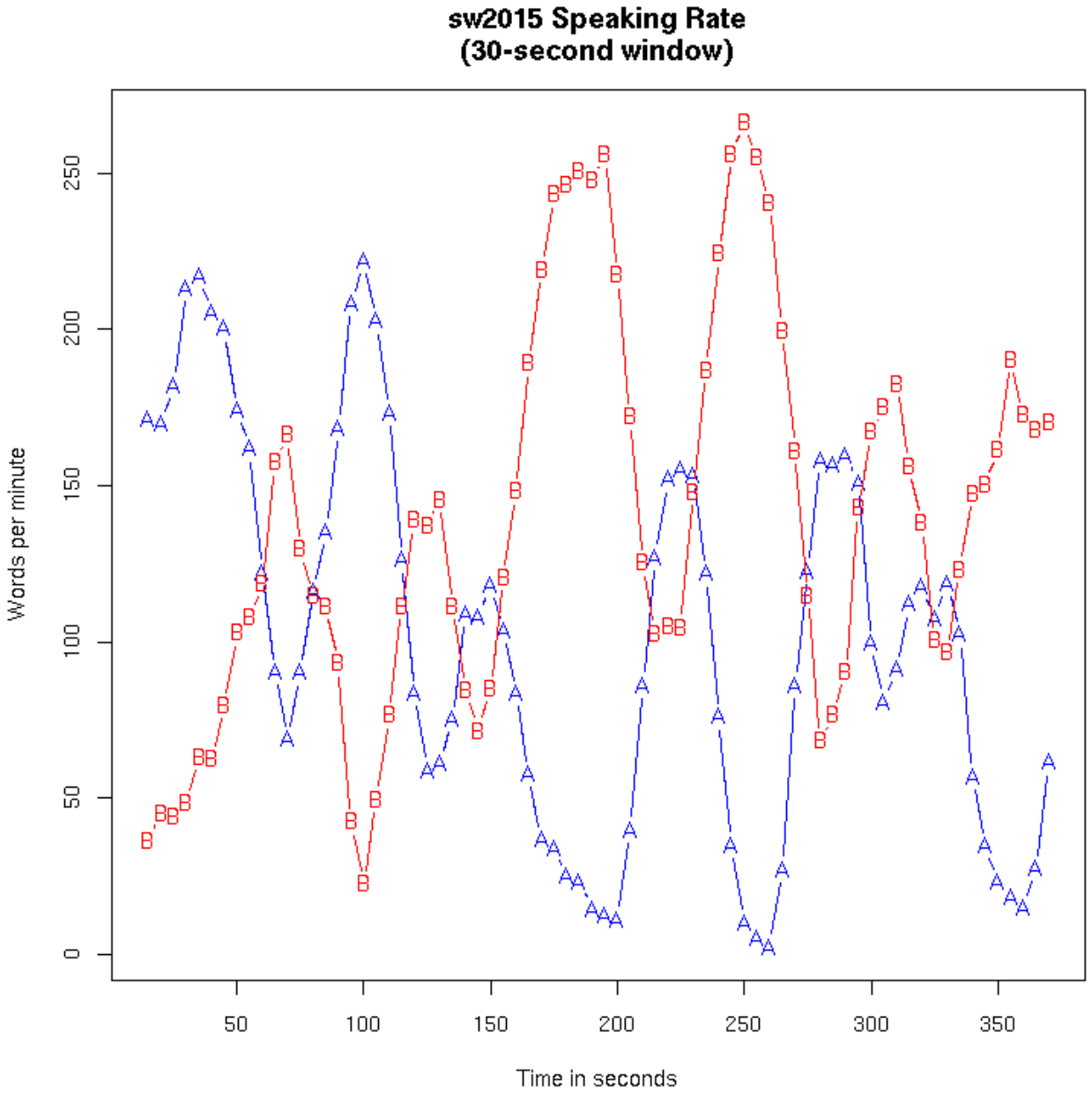


Some Questions

- Can we do the same with non-lab speech?
- How to account for time variation properly?
- Are such parameters better than alternatives?

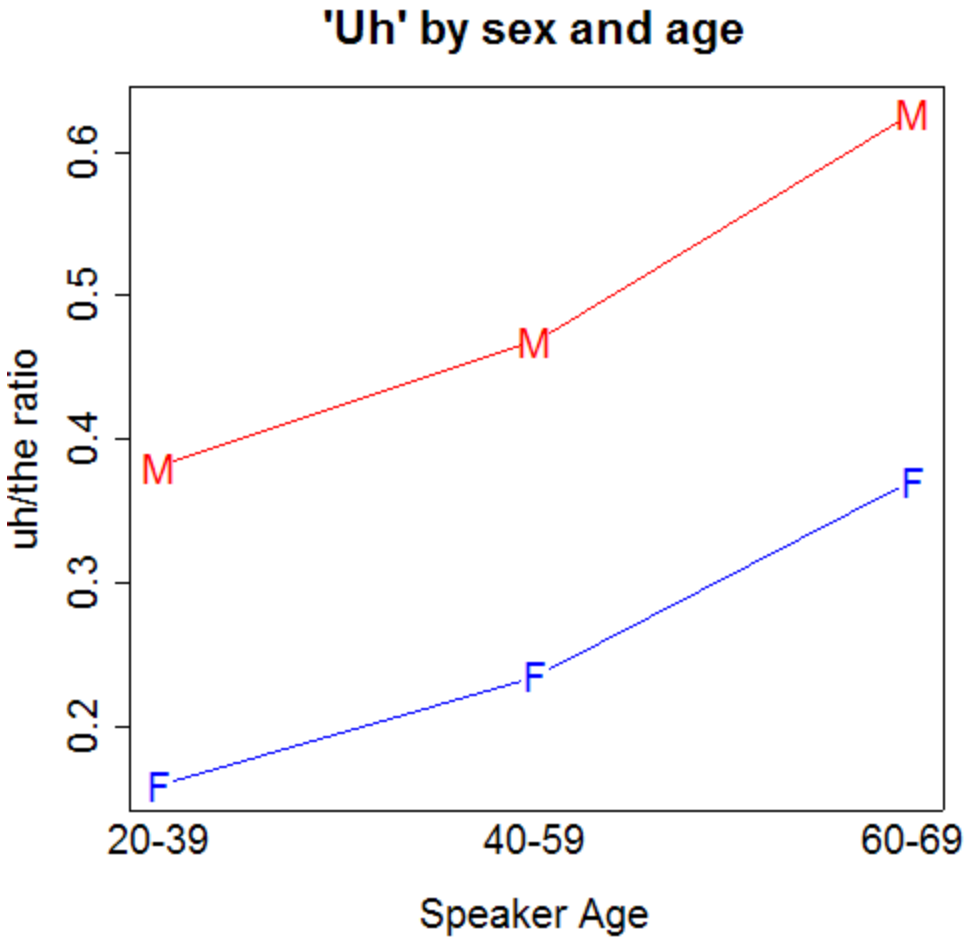
One-hour experiment #5

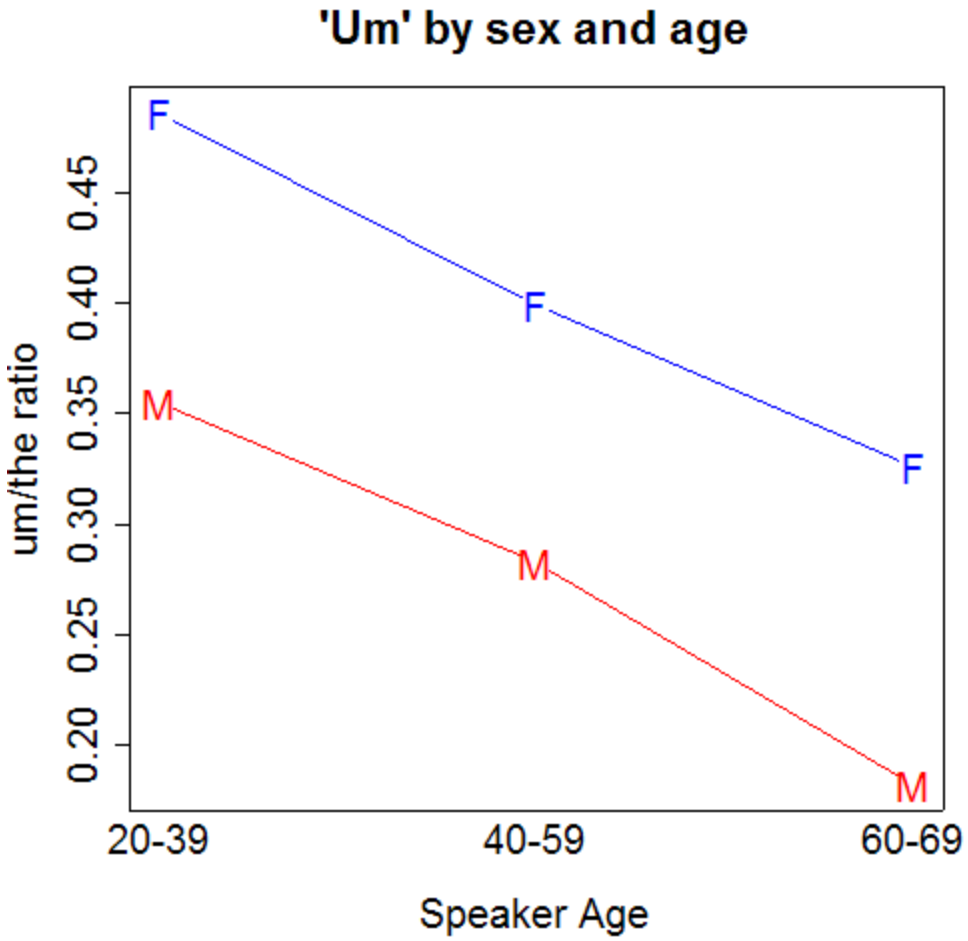
- How does speaking rate
  reflect the ebb and flow of a conversation?
- Method: word- or syllable-count
  in moving window
  over time-aligned transcripts
- Result: suggestive pictures


sw2015 Speaking Rate
(30-second window)
Words per minute
0
50
100
150
200
250
50
100
150
200
250
300
350
Time in seconds

One-hour experiment #5

- How does disfluency vary with sex and age?
- Method: count “filled pauses”
  in transcripts of U.S. English conversations
  by demographic categories of speakers
- Result: systematic but unexpected interaction


'Uh' by sex and age
uh/the ratio
0.2
0.3
0.4
0.5
0.6
M
M
M
F
F
F
20-39
40-59
60-69
Speaker Age


'Um' by sex and age
um/the ratio
0.20
0.25
0.30
0.35
0.40
0.45
F
F
F
M
M
M
20-39
40-59
60-69
Speaker Age

One-hour experiment #6

The News Editor at Psychology Today wrote to me:

> Sometimes I wonder if there are underlying personality differences between people who punctuate (litter?) their speech with "you know" versus those who use "I mean" more frequently. Any hunch on that?

I didn't have any hunches, and there didn't seem to be anything relevant in the literature. But I did have access to an indexed copy of the the 14,137 conversations (26,151,602 words) in the LDC's English-language conversational speech corpora.

(…and so do you!)

And there's demographic data for (almost) all speakers.
So I checked:

	"you know"	"I mean"	"you know"/"I mean" ratio
20-39	58,364	24,478	2.38
40-59	278,099	73,211	3.80
60+	33,477	7,518	4.45

Elapsed time: 6 queries + 3 ratio calculations = 5 minutes

What about the effect of years of education?

	"you know"	"I mean"	"you know"/"I mean" ratio
High school	2,608	408	6.39
College	191,088	51,143	3.72
Post-graduate	167,893	51,389	3.27

(Caveat:
High-school-only group was small,
and perhaps mainly older…)

Sex differences?

	"you know"	"I mean"	"you know"/"I mean" ratio
Women	198,086	51,689	3.83
Men	173,321	53,892	3.22

Elapsed time:

15 minutes for queries, 45 minutes to write it up

("I mean, you know", Language Log, 8/19/2007)

Conclusions?

Maybe greater use of "I mean" means greater involvement with self as opposed to others, and that age makes people less self-involved, but education makes them more self-involved, and men are somewhat more self-involved than women.

But this is even more tenuous than such explanations generally are, since the demographic variables in this collection of conversations are not orthogonal.

So you'd want to do some sort of hierarchical regression, and it would take a day or too to get the data and run it.

But still . . .

Serious science

- Old-style data and analysis
  is qualitatively easier and faster
- Resulting data can be shared and re-used
- ”Found” speech and language data
  opens up new universes
  on a scale 4-5 orders of magnitude
  greater than in the past

- And interesting patterns are everywhere!

Interdisciplinary opportunities

- These techniques
  will have rich applications in other fields
  - Clinical diagnosis and evaluation
  - Educational assessment
  - Social science survey methods
  - Studies of performance style
  - . . . and so on . . .
- Wherever speech and language are relevant!

Even in classical scholarship!

The early years of the twenty-first century have seen a heroic age for intellectual life. Ideas have poured across the world and new minds have joined the professionalized academics and authors in grappling with the heritage of humanity. […]

No field of study is poised to benefit more than those of us who study the ancient Greco-Roman world and especially the texts in Greek and Latin to which philologists for more than two thousand years have dedicated their lives. […]

The terms eWissenschaft and ePhilology, like their counterparts eScience and eResearch, point towards those elements that distinguish the practices of intellectual life in this emergent digital environment from print-based practices. Terms such as eWissenschaft and ePhilology do not define those differences but assert that those differences are qualitative. We cannot simply extrapolate from past practice to anticipate the future.

-- Gregory Crane et al., "Cyberinfrastructure for Classical Philology", *Digital Humanities Quarterly,* Winter 2009

Philological straws in the wind

- Just released:
  - Icelandic Parsed Historical Corpus (IcePaHC) 0.9
    - 1,002,390 words
    - From 1150 to 2008
  - Many other historical corpora, parsed and not
- Under discussion:
  - 1,000,000 English books
    - 2,500/year from 1520 to 1920
    - With good metadata
      and automatic parsing for sampling purposes
  - Other languages?

An historic opportunity:

- Take an interesting problem, and add
  - a little linguistics
  - a little psychology
  - a little signal processing
  - a little statistics and machine learning
  - a little computer science
  - your curiosity and initiative
- And the future is yours!

Thank you!