(Clinical) Human Language Technology -- and Science

Mark Liberman

University of Pennsylvania

http://ling.upenn.edu/~myl

We infer a lot from the way someone talks: personal characteristics like age, gender, background, personality; contextual characteristics like mood and attitude towards the interaction; physiological characteristics like fatigue or intoxication. Many clinical diagnostic categories have symptoms that are partly or entirely manifest in spoken interaction: autism spectrum disorder, neurodegenerative disorders, schizophrenia, and so on.

The development of modern speech and language technology makes it possible to create automated methods for diagnostic screening or monitoring. Even more important is the fact that these diagnostic categories are phenotypically diverse, representing (sometimes apparently discontinuous) regions of complex multidimensional behavioral spaces. We can hope that automated analysis of large relevant datasets will allow us to do better science, and learn what the true latent dimensions of those behavioral spaces are.

In this talk, I'll present some suggestive preliminary results, and discuss future research opportunities as well as the existing barriers to progress.

The context:

The past 50 years
have seen enormous quantitative changes
in the efficiency and reproducibility
of speech and language research,
thanks to advances in digital technology.

The near future will bring even larger changes –
not only quantitative changes in productivity and scale ,
but also qualitative changes in the nature of our research,
enabled by new (semi-)automatic methods.

New sources of data
and new methods of automated analysis
are opening up vast new territories of linguistic research.

We can easily acquire and manage new sources of linguistic data
that are several orders of magnitude bigger than old ones.

Because new methods can do old tasks several orders of magnitude more efficiently,
it's increasingly easy to explore these new datasets in old ways.

We can also easily experiment with completely new approaches to analysis and modeling.

And these new methodologies are rapidly spreading
into all the fields that study speech, language, and communicative interaction,
from poetics, sociology, and politics to psychology and neuroscience.

A trivial example:

In June 2014, I participated in a workshop discussion of *tonogenesis*
(A historical change in Chinese, Vietnamese, Thai etc.
where consonant manner distinctions turn into tone differences)

The anatomy, physiology, and physics of voicing distinctions in speech
naturally produce differences in f0.
This has been observed in isolated examples,
but there seemed to be no systematic study in the literature.

So over breakfast the next morning,
I checked out syllable-initial consonants in the TIMIT corpus.

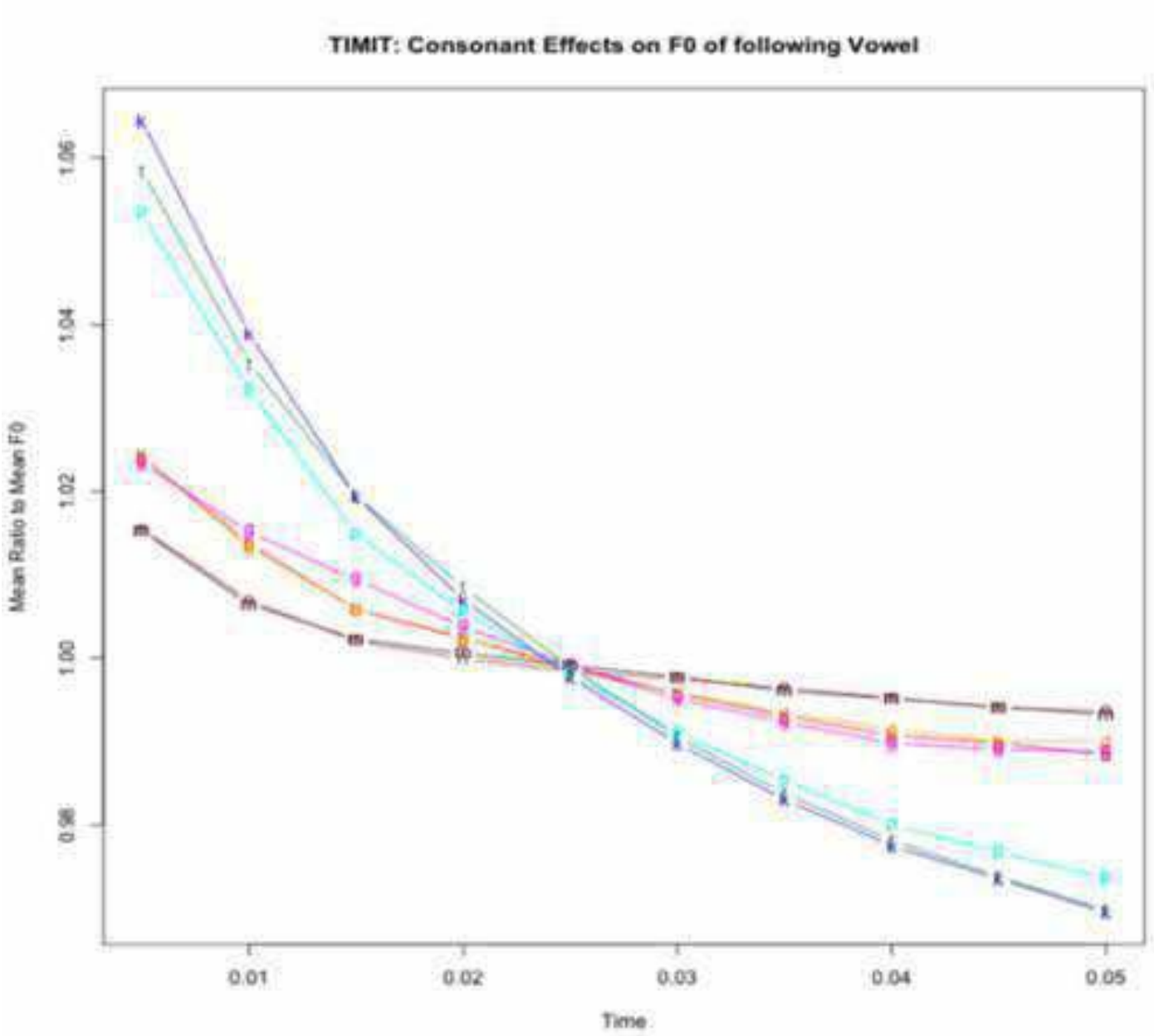
Method – write a script to

1. Pitch track all 6300 TIMIT sentences, creating f0 estimates every 5 msec
2. Select all syllables beginning with p,t,k b,d,g or m,n
3. Pull out the first 10 f0 estimates (first 50 msec of voiced speech per syllable)
4. Reject all sequences containing discontinuities
5. Normalize the f0 vectors as the ratio relative to each vector's mean value

$$y = x/mean(x)$$

6. Plot the mean of all normalized vectors for each initial consonant

("Consonant Effects on F0 of Following Vowels")


TIMIT: Consonant Effects on F0 of following Vowel
Mean Ratio to Mean F0
1.06
1.04
1.02
1.00
0.98
0.01
0.02
0.03
0.04
0.05
Time

In the old days, this would have been several years of work
(which is presumably why no one did it…)

In 2014, I could do it in an hour or so,
while consuming a bowl of cereal and several cups of coffee,
using a laptop computer and a page or so of code.

In later Breakfast Experiments,
I replicated the results for Spanish and Chinese.

But major challenges remain.

In some cases, the revolution in data and algorithms is simply incomplete:
There are kinds of data that are not generally available,
or not available at all.
And long before we get to the hypothetical automatic linguist,
there are many simple tasks where the state of the art is shockingly bad.

At the same time, as Human Language Technology gets better and better,
commercial success risks destroying the engines of research progress.

And paradoxically, there some are issues intrinsic to our new research methods
that create new problems at the same time that they solve old ones.

Challenges:

Important types of speech and language data are missing,
and filling some of the gaps requires careful coordinated effort

Unsupervised automatic language learning hardly works at all,
and (partly) supervised automatic language learning
doesn't work well enough (yet):

Commercial success may risk research failure.

And real-world language is not an orthogonally controlled experiment . . .

What about the opportunities?

They turn out to be pretty much the same as the challenges….

Challenge 1:
Hard-to-get data

Where does linguistic “big data” come from?

A digital shadow universe

increasingly mirrors real life

in flows and stores of bits.

Society is mostly about communication.

And most communication is text

(or talk, which is just text in fancy calligraphy)

. . . more and more often in digital form.

Simple properties of text

(like the words that make it up,
and the ways that they're performed)

are a good proxy for content.

Better than anything else we have, anyhow…

Bigger faster cheaper digital everything

(and better programming languages, and . . .)

make it easier and easier

to pull content out of the flows of text

in that digital shadow universe.

So in that new evolutionary niche:

a host of newly-evolving life forms
have got means, motive, and opportunity
to live off of these flows and stores of text

. . . while adding their digestion products
to the ecosystem.

From that digital ecosystem,
many kinds of text and speech
are easily collected and distributed.

In some cases,
there are intellectual property rights to be licensed,
but this is generally not hard to do.

In contrast, there are some kinds of datasets
where privacy and confidentiality
pose difficult ethical and legal problems,
especially for data sharing across sites.

For example:

Recordings of clinical interviews,
neuropsychological tests,
and similar things*.

There are policies, laws, and ethical concerns
that require such recordings to be treated in a special way,
and are widely believed to make cross-site sharing impossible.

** e.g. job interviews, educational testing sessions, . . .*

Why do we want such recordings for research,
and why do we want to share them?

Because speech and language are often a key behavioral marker,
cheaper and less invasive
than brain imaging, blood tests, or genomic tests,
but also often diagnostically more useful.

And more important, many (most?) relevant problems
are “phenotypically diverse”, in ways that matter –
meaning that we really don’t understand them very well.

Sample heterogeneity + small samples + poor measurement

=

non-reproducible scientific results

With enough data and enough research,
we can hope to find the true latent dimensions
of the relevant behavioral space(s).

But a single site rarely has enough data,
and no single research team is likely to find the answers.

We need to pool data across sites,
and we need a community of researchers
working together to understand it.

Example: “Autism Spectrum Disorder”

It’s clear that Autism is not a “spectrum”, i.e. a single dimension,
but rather a space, with many dimensions –

It’s a space that we all live in,
with some corners that have been medicalized
because they cause serious life problems.

Is there suitable digital data Out There?

Yes –
for instance, the Autism Diagnostic Observation Schedule (ADOS)
is a standard diagnostic tool,
consisting of a multi-part structured interview
which is video recorded and scored from the video,
with a half a dozen scoring rubrics for of the ~12 segments.

For diagnosis, the multiple scores are added up and thresholded.

O(1,000,000) ADOS recordings are Out There.

An ADOS recording DVD is stored in the patient's folder,
along with many other tests.

We've begun a collaboration
with the Center For Autism Research
at Children's Hospital of Philadelphia,
which has O(3000) such recordings.

We selected an initial set of ~100 interviews,
including interviews with neurotypical controls
and with adolescents with other diagnoses such as ADHD.

We did some preliminary work
to persuade the hospital's Institutional Review Board
that it was both possible and worthwhile
to share 20-minute ADOS audio segments for research purposes

-- with appropriate safeguards.

The CAR clinicians contacted the parents and children involved
to get informed consent for sharing
anonymized audio and transcripts with other researchers,
where “anonymized” means that
personal names, addresses, institutional names etc.
are bleeped from the audio
and replaced by generic placeholders in the transcripts.

Nearly everyone who could be reached agreed –
we ended up with 50 20-minute segments,
which should be published this year by the LDC.

Preliminary research on this small pilot corpus (~33 hours)
suggests that every sensible linguistic measurement
shows some interesting signal.

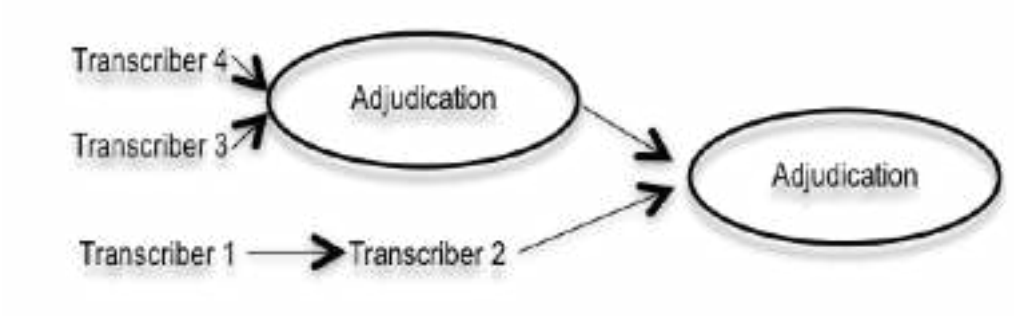
We hope to persuade other clinical centers
to join us in creating a much large collection.

As Bob Schultz, CAR's director, said:

"With ten thousand interviews,
maybe we can figure out what's really going on."

Transcription specification similar to those used for conversational speech

4 transcribers and 2 adjudicators from LDC and CAR produced a “gold standard” transcript for analysis and for evaluation/training of future transcriptionists:



Simple comparison of word level identity between CAR’s adjudicated transcripts and LDC’s transcripts: 93.22% overlap on average, before a third adjudication resolved differences between the two.

Transcripts force-aligned to audio.

Participants:

- Pilot sample
- N=100
- Mean age=10-11 years
- Primarily male
- 65 ASD, 18 TD, 17 Non-ASD mixed clinical
- Average full scale IQ, verbal IQ, nonverbal IQ

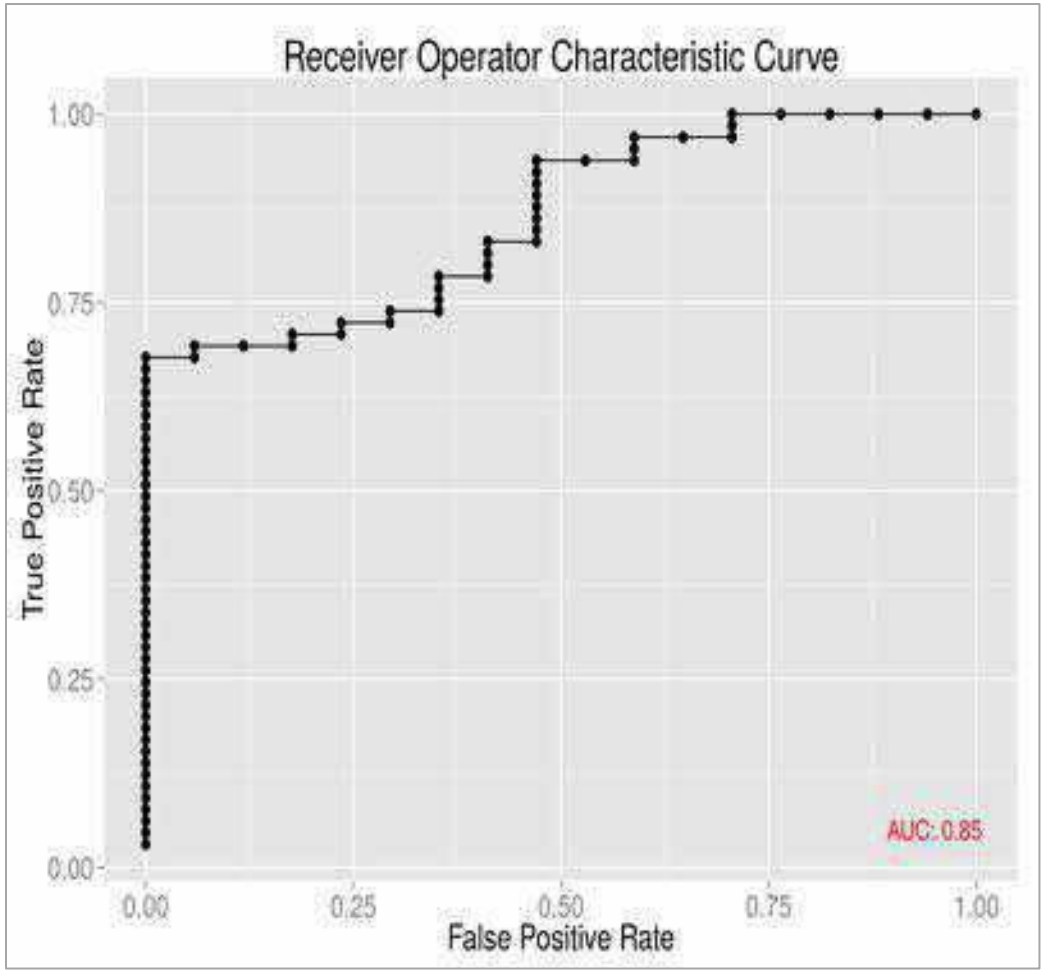
Simple bag-of-words classification – worked better than expected, as usual:

Naïve Bayes, weighted log-odds ratios

Leave-one-out cross validation

Correctly classified
68% of ASD participants
and 100% of typical participants

AUC=85%



20 most “ASD-like” words:

{nsv}, know, he, a, now ,no , uh, well, is, actually, mhm, w-, years, eh, right, first, year, once, saw, was

{nsv} stands for “non-speech vocalization”,
“uh” appears in this list, as does “w-”, a stuttering-like disfluency.

20 least “ASD-like” words:

like, um, and, hundred, so, basketball, something, dishes, go, york, or, if, them, {laugh}, wrong, be, pay, when, friends.

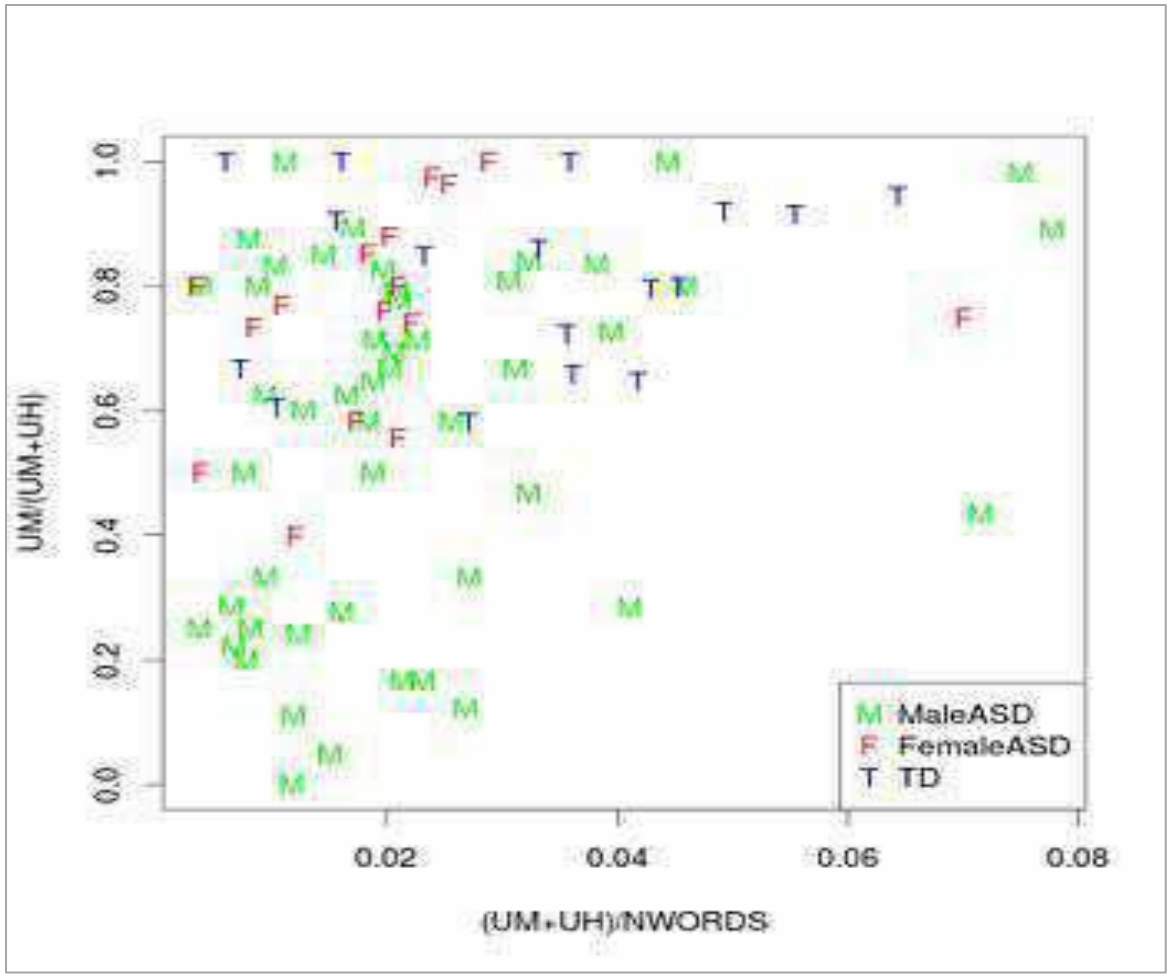
“um” appears, as does laughter and the word “friends”

Rates of *UM* production in ASD and TD groups: ***um/(um+uh))***

ASD group: *UM* was 61% of their filled pauses (CI: 54%-68%)

TD group: *UM* was 82% of their filled pauses (CI: 75%-88%)

Minimum value for the TD group was 58.1%, and 23 of 65 participants in the ASD group fell below that value.



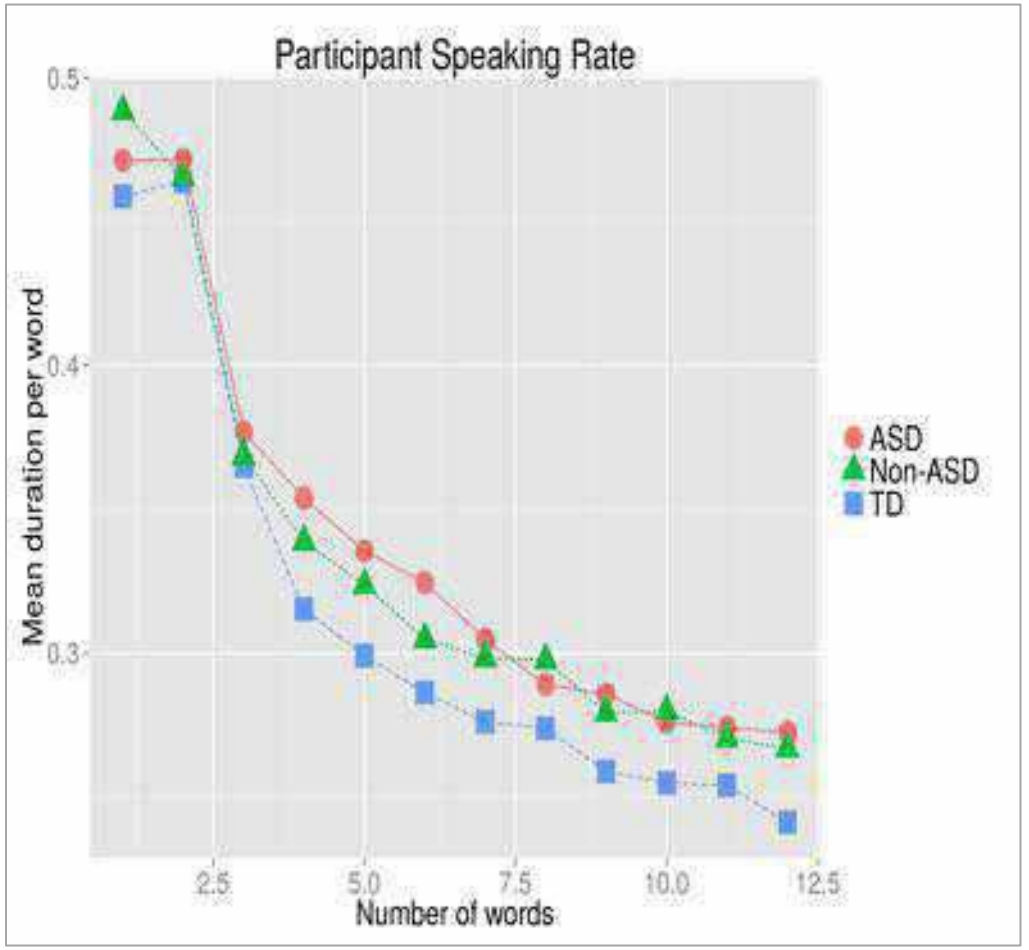
And every other feature we looked at
also showed some diagnostically relevant signal . . .

Mean word duration as a function of phrase length:

TD participants spoke the fastest
(overall mean word duration of 376 ms, CI 369-382,
calculated from 6,891 phrases)

Followed by the non-ASD mixed clinical group:
(mean=395 ms; CI 388-401,
calculated from 6,640 phrases)

Followed by the ASD group:
(mean=402 ms; CI: 398-405,
calculated from 24,276 phrases)

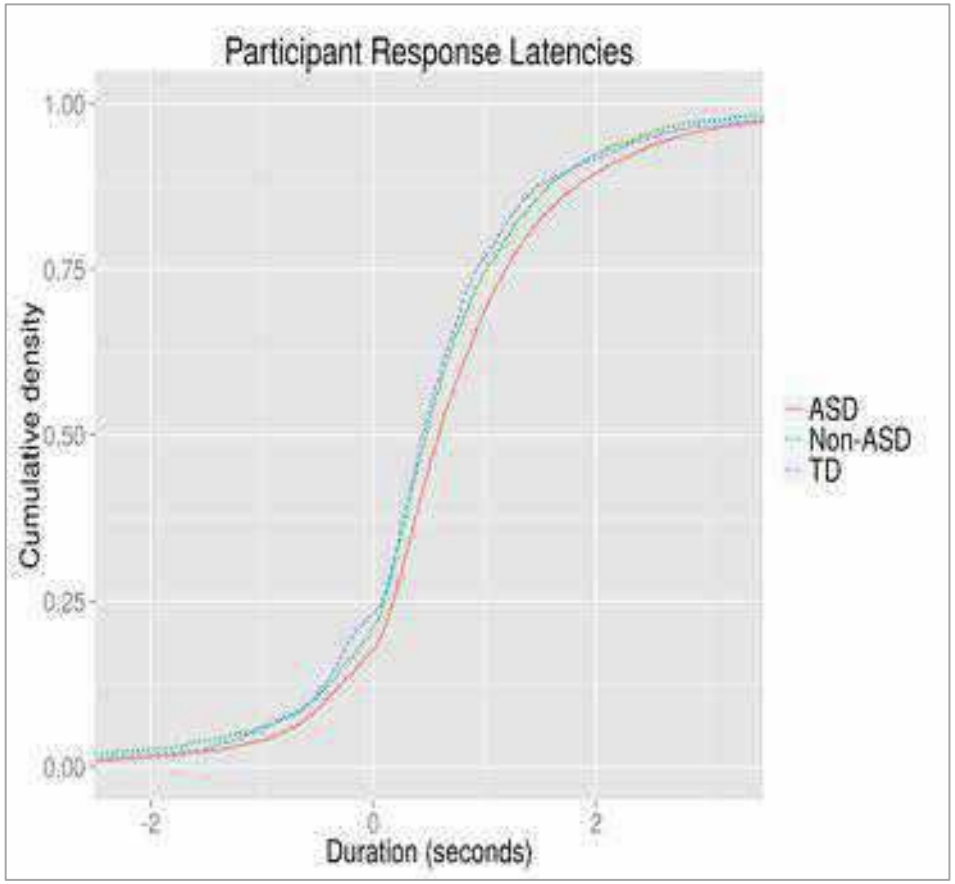


Latency to respond:

Too short = interrupting
speaking over a conversational partner

Too long = awkward silences
interfere with smooth social exchanges

ASD slower than TD

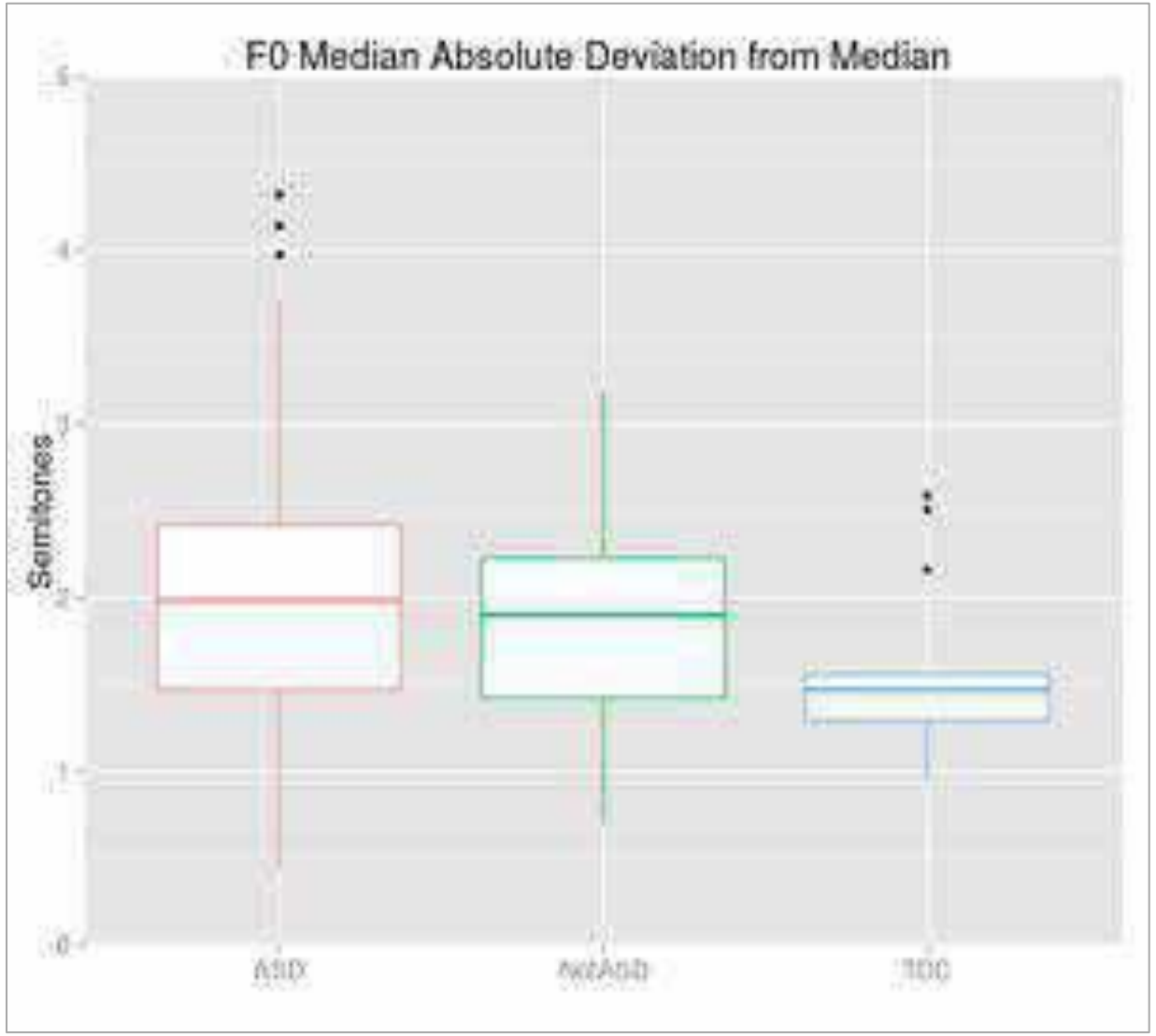


F0 Variation:

Median absolute deviation
from the median (MAD)
Calculated in semitones
relative to speaker's 5th percentile

MAD values are both higher
and more variable
in the ASD
and non-ASD mixed clinical group
compared to the TD group:

ASD: median 1.99 IQR: 0.95
Non-ASD: median 1.95 IQR 0.80
TD: median 1.47 IQR 0.26



. . . and so on . . .

Next steps for ADOS analysis

Expand sample size, enlarge age range, improve specificity

- Multi-site collaboration?
- Downward extension to infancy
- Chart growth to identify points of divergence/targets for intervention

New measures

- New textual and acoustic-phonetic features
- Integration of textual & phonetic features
  - (e.g. dysfluency & pause locations)
- Gesture, gaze, face, posture during conversation
- Other phenotypic data
- Neuroimaging
- Genetics

BUT…

ADOS requires expensive, time-consuming, in-person expert collection --

We (also) need scalable, inexpensive methods
to collect large, diverse samples.

New ASD Data Collection Initiatives:

Phone bank

Inexpensive student worker asks ADOS-like questions
Child and parent language samples, questionnaires, online IQ
Nationally representative cohort

Computerized Social Affective Language Task (C-SALT)

Portable self-contained app
Records language and social affect in schools, clinics, homes
Controlled recording is conducive to automated approaches
(reduces need for transcription)

Goals and applications:

Support clinical decision-making and improve access

- Low-cost, remote screening
- Direct behavioral observation: record in clinics, integrate into EHR
- Inform identification efforts and assist in differential diagnosis

Identify behavioral markers of underlying (treatable) pathobiology

- Profiles of individual strengths and weaknesses, link to biology
- Personalized treatment planning and improved outcomes
- ****Monitoring and measuring response to interventions****

Give participants and families more information about themselves

- Online feedback
- Monitor growth trajectories

There are many other kinds of datasets
relevant for ASD research –

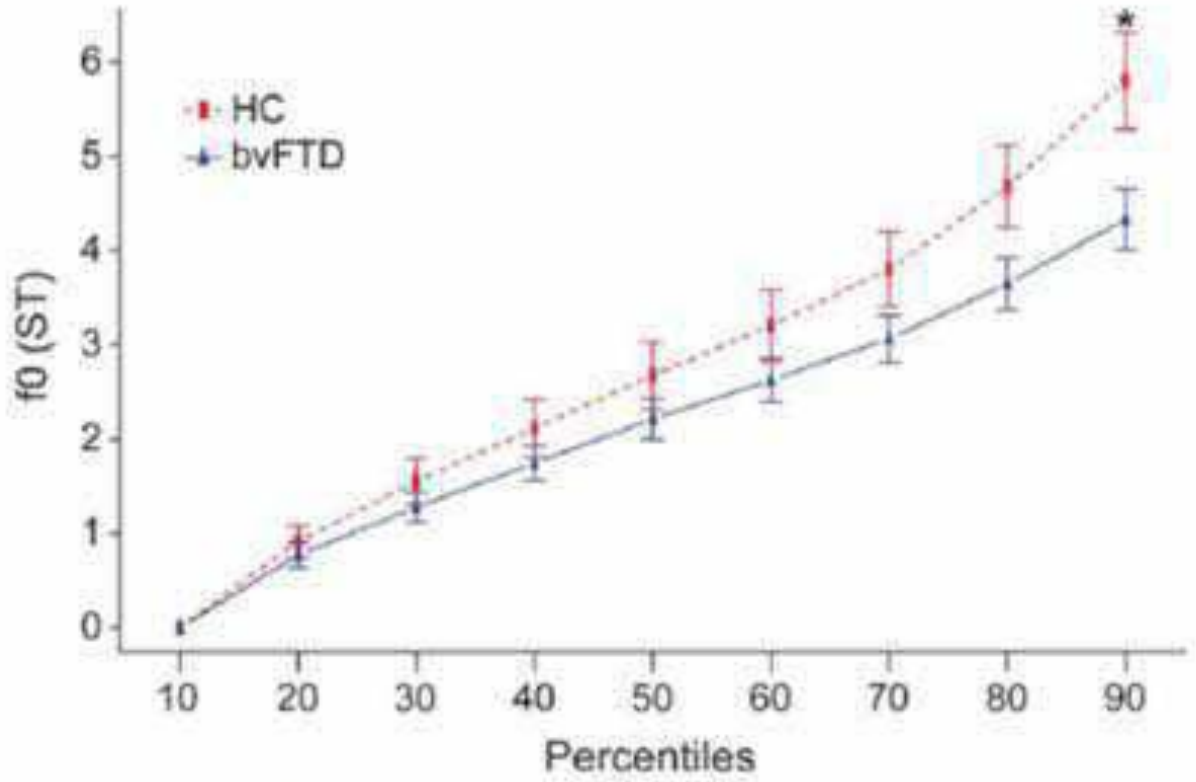
And many other possible targets for similar research,

for example, the many diverse varieties
of neurodegenerative disorders,
such as Frontotemporal Degeneration,
Parkinsonism, and Alzheimers.

We're working with Penn's Frontotemporal Dementia Center
on a dataset of picture-description recordings
from ~1000 patients and elderly controls.

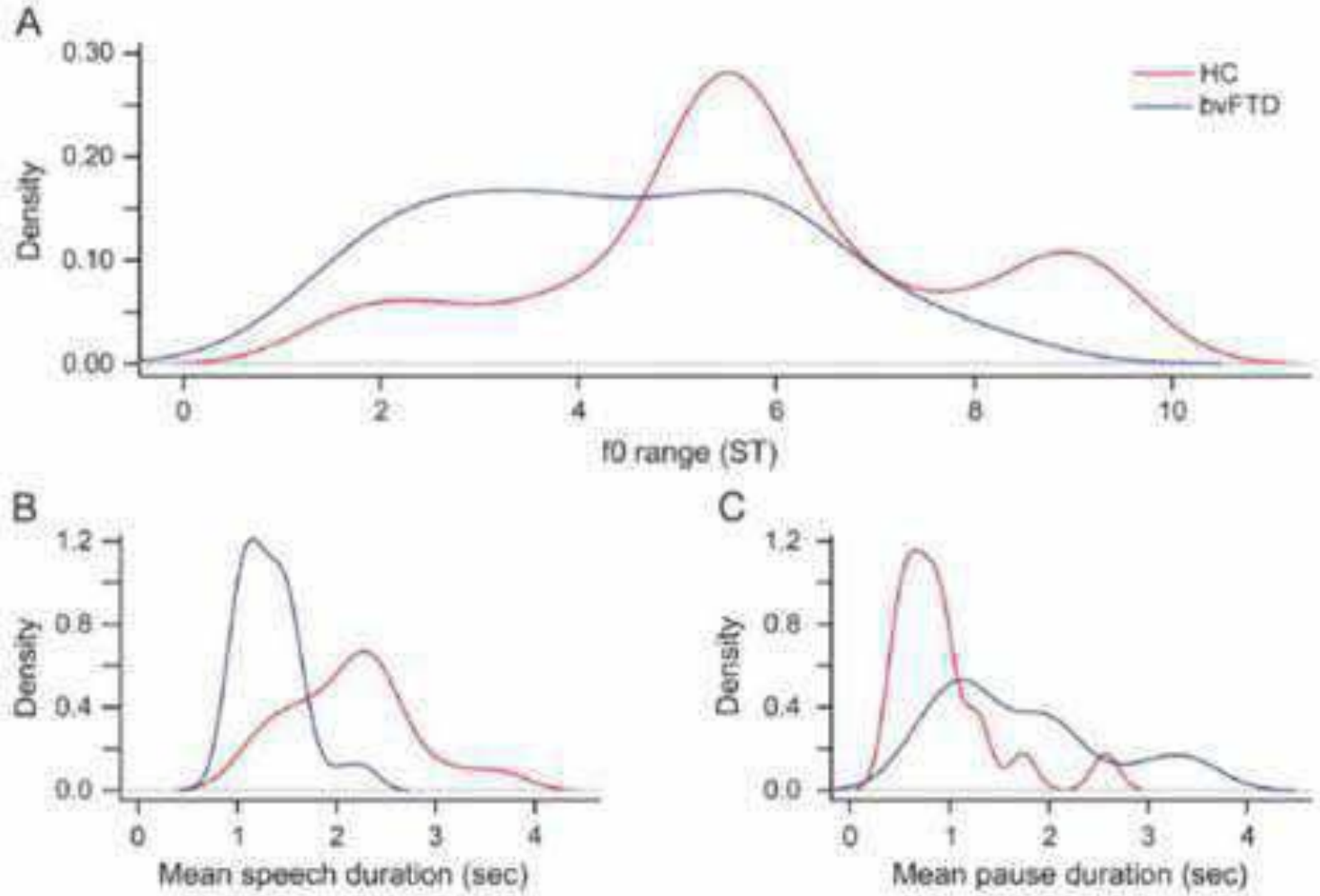
Early results show diagnostic value
in simple acoustic-phonetic measures.

Figure 1 Fundamental frequency percentiles per group

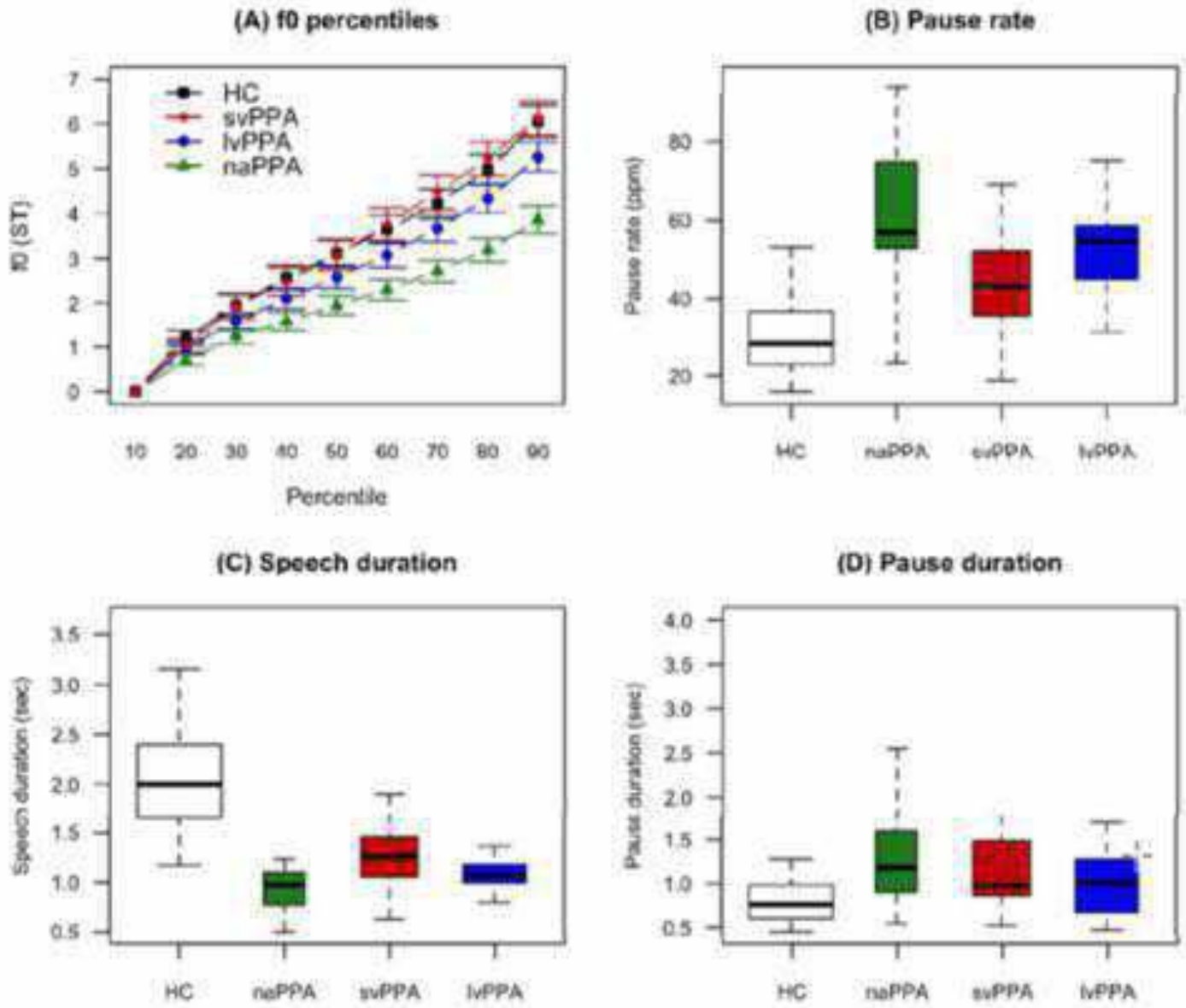


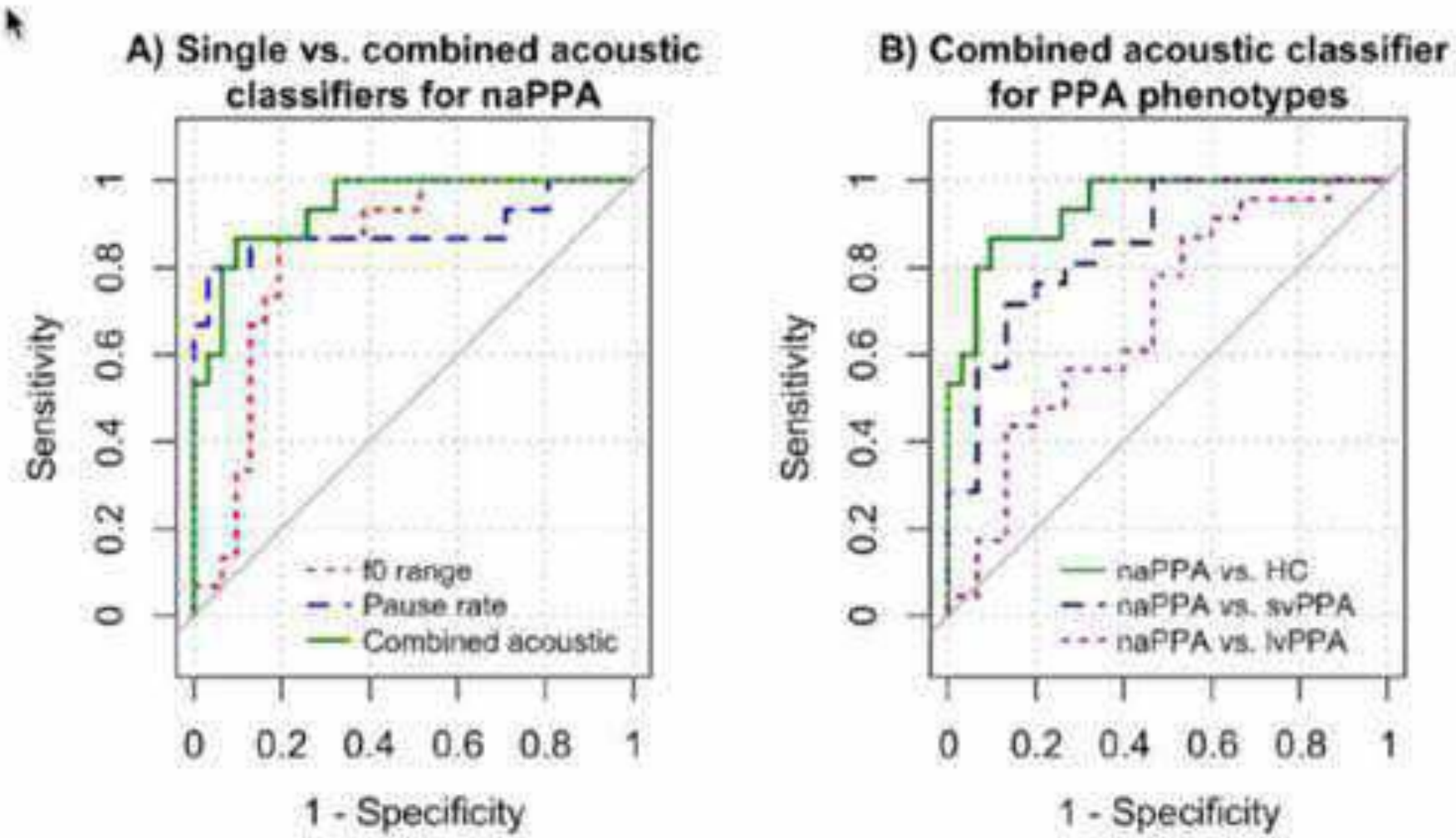
Fundamental frequency (f0) estimates in 10th percentile bins for healthy controls (HC) (n = 17) and behavioral variant of frontotemporal dementia (bvFTD) patient group (n = 32) with standard error bars. The f0 range is represented by the 90th percentile and is limited to 4.3 ± 1.8 semitones (ST) for the patient group compared to HC (5.8 ± 2.1 ST). *p = 0.03.

Figure 3 Speech measures distributions



Kernel density plots for fundamental frequency (f0) range (A), speech segment (B), and pause segment (C) durations for patients with behavioral variant of frontotemporal dementia (bvFTD) vs healthy controls (HC). ST = semitones.


(A) f0 percentiles
HC
svPPA
lvPPA
naPPA
f0 (ST)
Percentile
(B) Pause rate
Pause rate (spm)
HC
naPPA
svPPA
lvPPA
(C) Speech duration
Speech duration (sec)
HC
naPPA
svPPA
lvPPA
(D) Pause duration
Pause duration (sec)
HC
naPPA
svPPA
lvPPA


A) Single vs. combined acoustic classifiers for naPPA
Sensitivity
0
0.2
0.4
0.6
0.8
1
f0 range
Pause rate
Combined acoustic
0
0.2
0.4
0.6
0.8
1
1 - Specificity
B) Combined acoustic classifier for PPA phenotypes
Sensitivity
0
0.2
0.4
0.6
0.8
1
naPPA vs. HC
naPPA vs. svPPA
naPPA vs. lvPPA
0
0.2
0.4
0.6
0.8
1
1 - Specificity

We're working with the Framingham Heart Study
on ~9,000 1-2 hour recordings of neuropsychological testing
from 5,267 subjects, with about 100 new sessions/month.

There are currently 122 participants diagnosed with MCI
and 212 with dementia

As the website for the Framingham Heart Study explains:

> Since our beginning in 1948, the Framingham Heart Study, under the direction of the National Heart, Lung and Blood Institute (NHLBI), formerly known as the National Heart Institute, has been committed to identifying the common factors or characteristics that contribute to cardiovascular disease (CVD). We have followed CVD development over a long period of time in three generations of participants.
>
> Our Study began in 1948 by recruiting an Original Cohort of 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts, who had not yet developed overt symptoms of cardiovascular disease or suffered a heart attack or stroke. Since that time the Study has added an Offspring Cohort in 1971, the Omni Cohort in 1994, a Third Generation Cohort in 2002, a New Offspring Spouse Cohort in 2003, and a Second Generation Omni Cohort in 2003.
>
> Over the years, careful monitoring of the Framingham Study population has led to the identification of major CVD risk factors, as well as valuable information on the effects of these factors such as blood pressure, blood triglyceride and cholesterol levels, age, gender, and psychosocial issues. Risk factors for other physiological conditions such as dementia have been and continue to be investigated. In addition, the relationships between physical traits and genetic patterns are being studied.

The FHS began doing neuropsychological testing of its participants in 1999, and has been making digital recordings of the testing interviews since 2005. We have recently begun a collaboration with Rhoda Au, the Director of Neuropsychology for the FHS, to explore the possibility that a more detailed analysis of these recordings might produce interesting results.

The testing sessions involve 23-29 segments, including paired-associate learning, logical memory (= story retelling), digit span, picture naming, verbal fluency, etc.

There are > 220 autopsy cases, and 576 participants who are currently enrolled in a brain donation program, the majority of whom are older.

Formal aims of the FHS collaboration (from the proposal to the FHS Exec):

1: Generate "gold standard" transcripts of 8000+ existing voice recordings as well as those being acquired through on-going neuropsych testing of all FHS cohorts.
2: Analysis of the 8000+ existing voice recordings obtained between 2005-present using existing voice recognition and voice analysis software.
3: Build additional software to analyze the digital voice signals and generate novel cognitive metrics from latency and other behavioral characteristics.
4: For each neuropsychological test as well as across tests, identify normative values for e-voice metrics stratified by age, education, sex, both individually and in combination (e.g., age x education; age x sex; age x education x sex)
5: Conduct factor and cluster analysis of e-cognitive metrics across neuropsychological tests to identify domain specific measures.
6: Determine e-voice metrics/profiles that differentiate between those with and without known AD risk factors, including but not limited to, ApoE, family history of dementia/AD, homocysteine, vascular risk factors (including metabolic), inflammatory markers.
7: Determine whether neuroimaging biomarkers are related to e-voice metrics/profiles.
8: Determine whether incident change in neuroimaging biomarkers and neuropsychological tests are related to e-cognitive profiles.
9: Determine whether e-voice metrics/profiles can differentiate participants who are low to high risk for dementia/AD.
10: Conduct data driven analyses to identify e-voice metrics/profiles predictive of AD endophenotypes and risk for dementia/AD, in isolation and in combination with other health, lifestyle, biomarkers and genetic risk factors.

Pilot project goals – summer 2018:

1. Recruit and train transcription crew
2. Create and test transcription guidelines
3. Adapt transcription software
4. Design workflow and create workflow management system
5. Transcribe, align, and analyze ~200 selected interviews
6. Try simple forms of automation, reducing labor from 15X to 3X
    a) Speech activity detection
    b) Diarization
    c) Interview segmentation
7. Experiment with ASR first pass
8. Get funding

Neither the FTD nor the FHS dataset can easily be shared at present,
though we have hopes for the FHS collection.

In general, there needs to be a painful cultural shift
in the biomedical research community,
where we can hope that the pressing need for reproducibility
will overcome researchers' proprietary attitudes towards data.

Challenge 2:

Inadequate Algorithms

Unsupervised (computational) language learning doesn't work at all.

As a result, most of the worlds languages and and language varieties are resource-poor.

We can build decent acoustic models with easily-collected speech data
(…though multiplied by the number of languages and local varieties it's still a big job,
and unsupervised methods don't work yet here either).

But getting adequate transcribed data for a language model
(much less enough annotated data for syntactic, semantic, and discourse analysis)
remains a massively labor-intensive process.

Even supervised automatic methods don't work well enough:

ASR transcripts are not yet good enough
for many kinds of linguistic research.

General phonetic annotation given a orthographic transcript
is not good enough.

And automatic diarization (who spoke when)
is shockingly bad.

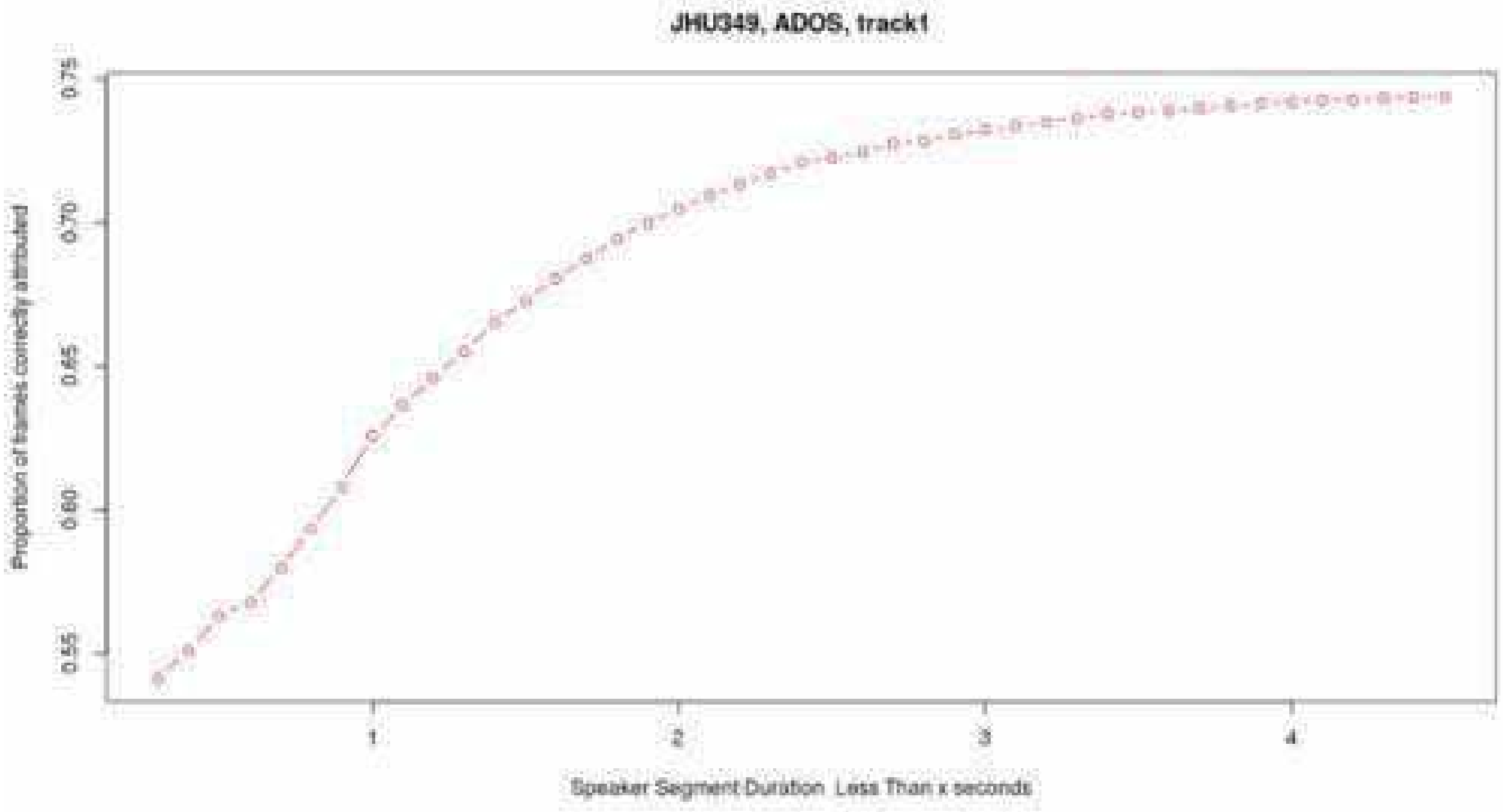
Some examples from the first DIHARD Diarization Challenge
(to be presented at Interspeech 2018) --

In DIHARD Track 1, systems are given ”gold” speech segment boundaries.

But the best system’s frame-wise speaker assignments in ADOS were

1. No better than chance for short segments
2. Not terrific for longer segments

[Discussion here]


JHU349, ADOS, track1
Proportion of trials correctly attributed
0.55
0.60
0.65
0.70
0.75
1
2
3
4
Speaker Segment Duration Less Than x seconds

At this point we can probably leave the general ASR problem
to the big companies.

But accurate phonetic annotation
of orthographically-transcribed audio
is not a problem of much current interest to those companies,

And smaller research groups are making progress on it.

Similarly for diarization – for more on this,
see the reports from JSALT 2017 at Interspeech 2018.

Challenge 3:

Commercial success

Paradoxically, the commercial success of HLT
can threaten research progress –

-- especially on problems
where commercial cost/benefit analysis
doesn't motivate research investment,
or where current engineering orthodoxy
points in the wrong direction.

Some sources of public funding are starting to take the view that IBM, Microsoft, Apple, Google, Facebook, Amazon, & Baidu have solved all the problems of Human Language Technology, or are about to do so.

This risks removing public funding from the process of building "common task" research communities, and the difficult problem of spreading that methodology to new areas like clinical research.

Structure of the “Common Task” method

- A detailed task definition and “evaluation plan”
  developed in consultation with researchers
  and published as the first step in the project.
- Automatic evaluation software
  written and maintained by a neutral third party
  and published at the start of the project.
- **Shared data:**
  Training and “dev(elopment) test” data
  is published at start of project;
  “eval(uation) test” data is withheld
  for periodic public evaluations

This method was originally developed
for Human Language Technology,
where it's been strikingly successful –

And it's spread widely to other areas of engineering.

But it's still rare in science,
especially in clinical areas
where the situation is in some ways similar
to HLT research.

Even in when clinical data sharing has been mandated, other part of the structure are missing

e.g. the **Alzheimer's Disease Neuroimaging Initiative** (ADNI) organized in 2004 by NIH

A slide from a presentation by Neil Buckholtz (National Institute on Aging)

"Transforming Research through Open Access to Discovery Inputs and Outputs"

at the Berlin 9 Open Access Conference, November 2011:

GOALS OF THE ADNI: LONGITUDINAL MULTI-SITE OBSERVATIONAL STUDY

- Major goal is collection of data and samples to establish a brain imaging, biomarker, and clinical database in order to identify the best markers for following disease progression and monitoring treatment response
- Determine the optimum methods for acquiring, processing, and distributing images and biomarkers in conjunction with clinical and neuropsychological data in a multi-site context
- "Validate" imaging and biomarker data by correlating with neuropsychological and clinical data.
- Rapid public access of *all* data and access to samples

BUT ADNI has

- No speech/language data
  (not in 2004, and not added later)
- No well-defined versioning of datasets
- No quantitative evaluation metric
- No focused workshops

Predicting the time course of Alzheimer's Disease
is exactly the kind of problem
("algorithmic analysis of the natural world")
for which the Common Task method has worked in the past.

We should apply such methods
to the large class of similar biomedical problems
(including those where speech and language are centrally involved)

Many scientists will be horrified
(just as speech and NLP engineers were in 1987-1992).
But in the face of the reproducibility crisis,
public funders need to force it to happen.

Challenge 4:

Real-world speech and language

Traditional instrumental phonetics

was usually based on recordings created by “subjects”

reading lists of artificial material in a laboratory setting

(often isolated words, words in a carrier phrase,

or somewhat strange sentences like

“She had my dark suit in greasy wash water all year”).

Advantage: a controlled setting, and to some extent

controlled variation of relevant factors.

Disadvantage: material is not “ecologically valid”,

and many factors are simply absent.

If somewhat more natural material was used
in traditional instrumental phonetics, like a read passage,
the size was necessarily limited.

Now, due to large amounts of available real-world data,
whether conversational or read,
and effective (semi-)automatic processing methods,
we can work with more natural material
at a scale three to six orders of magnitude larger.

But data quantity is not just for bragging rights!

Many linguistic dimensions interact:

- Language + Regional, social, and individual variation
- Register and style
  - Read speech vs. conversation; formality
  - Cultural patterns
- Syllabic and segmental context
- Phrasal structure and position
- Focus, emphasis, redundancy
- Emotion – arousal and valence
- . . .

And therefore quantity potentially turns into quality.

In order to understand the linguistic correlates
of phenotypically diverse clinical disorders –
especially in early stages
or to quantify trajectories over time,
as in the effects of treatments
or developmental patterns –

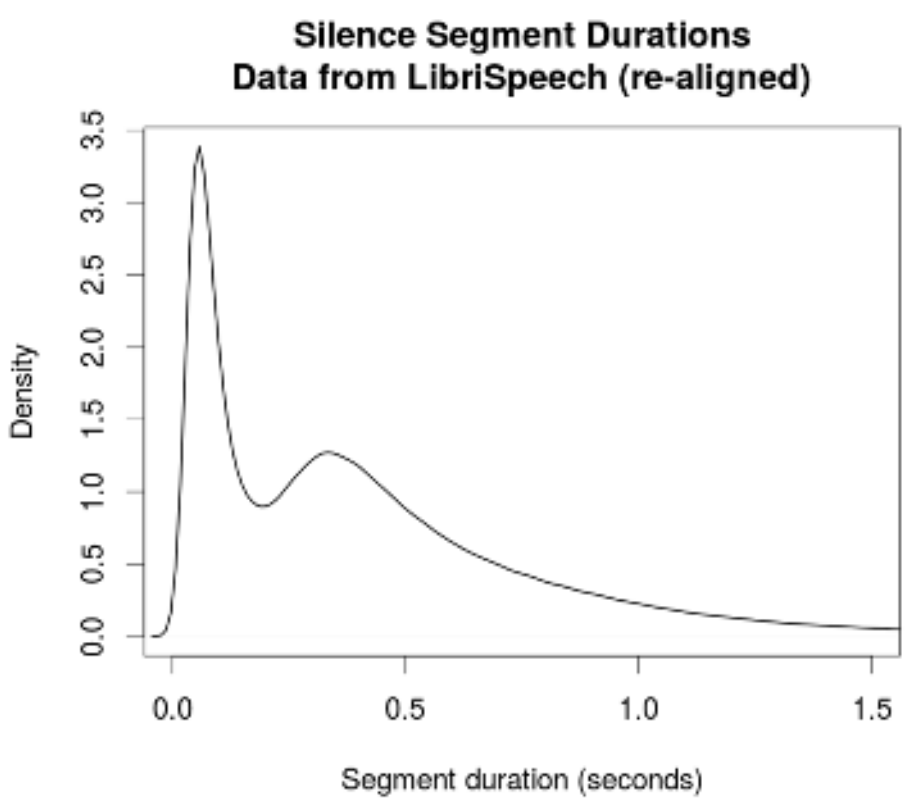
we need to understand the landscape of normal variation
across all of the many relevant dimensions.

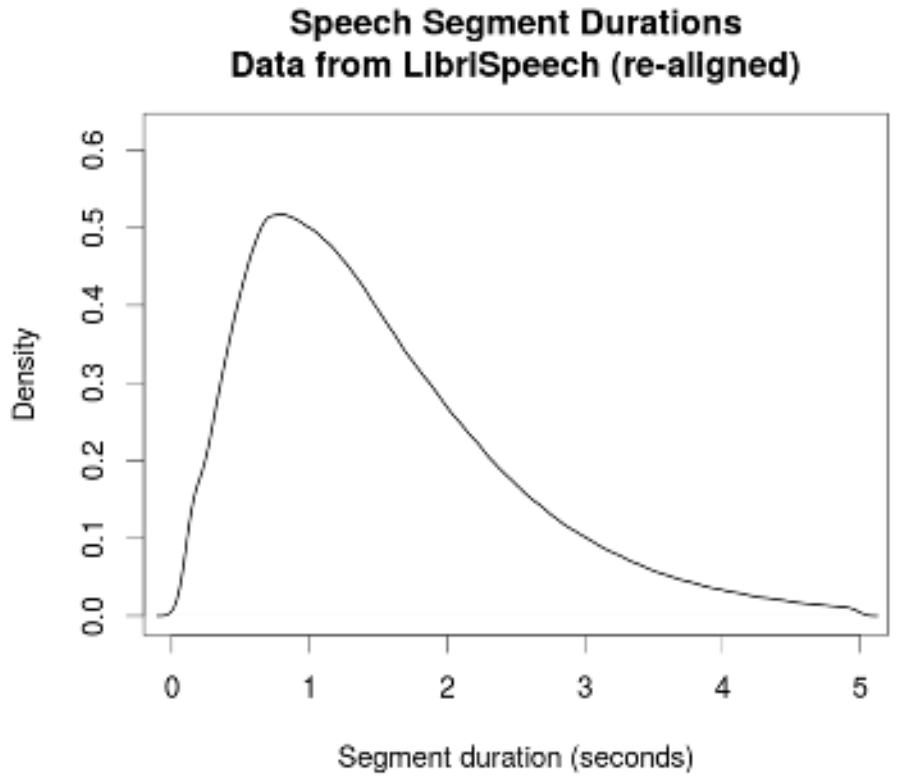
A simple example:

Some of the prosodic differences
between reading and spontaneous speech

The LibriSpeech dataset consists of 5,832 English-language audiobook chapters
read by 2,484 speakers, with a total duration of of nearly 1,600 hours.
[The LibriVox collection as a whole now has more than 60k hours of English]

The overall distributions of silence-segment and speech-segment durations look like this:

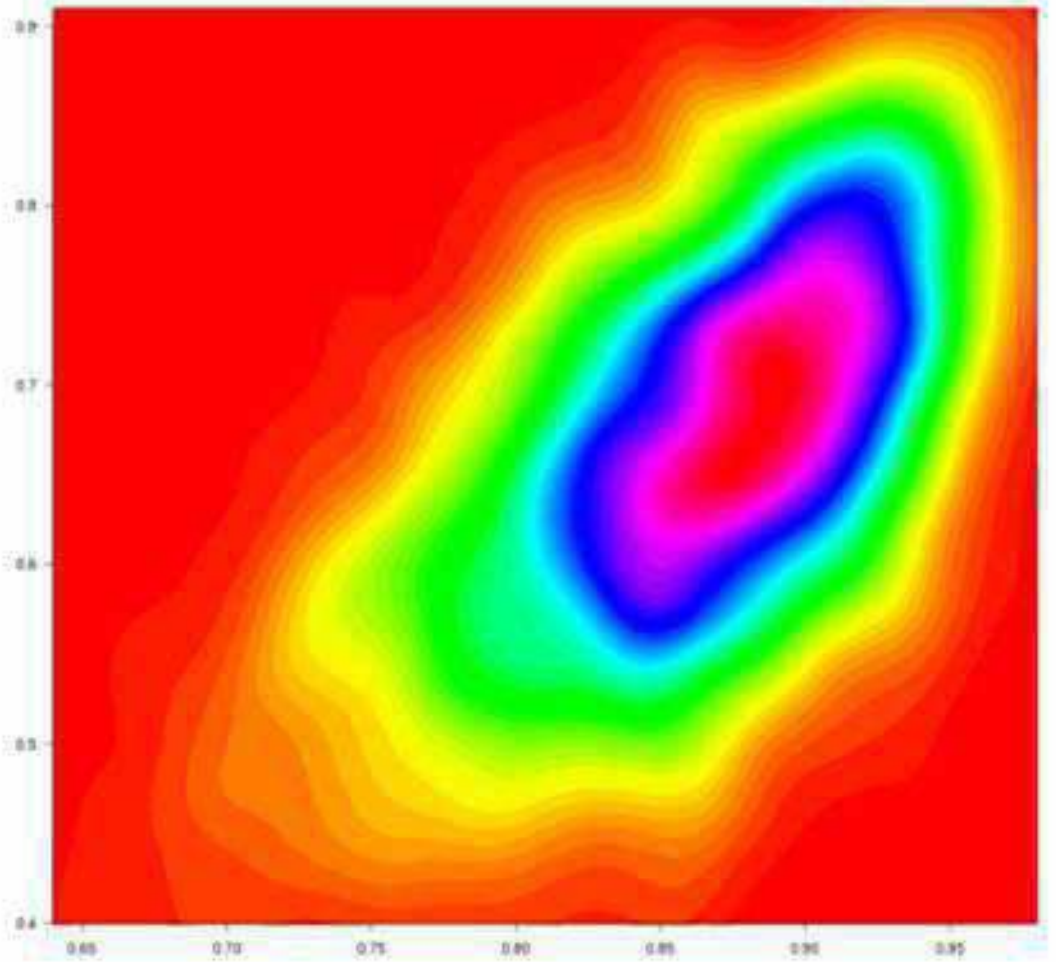




How to model variation among readers?

One simple way to map individual differences
is to plot the proportion of silence segments greater than X1, and the proportion of speech segments greater than X2.

With X1 = 200 msec. & X2 = 600 msec.,
the resulting 2D density plot looks like this:



The x-axis is
the proportion of silence segments > 200 ms.

The y-axis is
the proportion of speech segments > 600 ms.

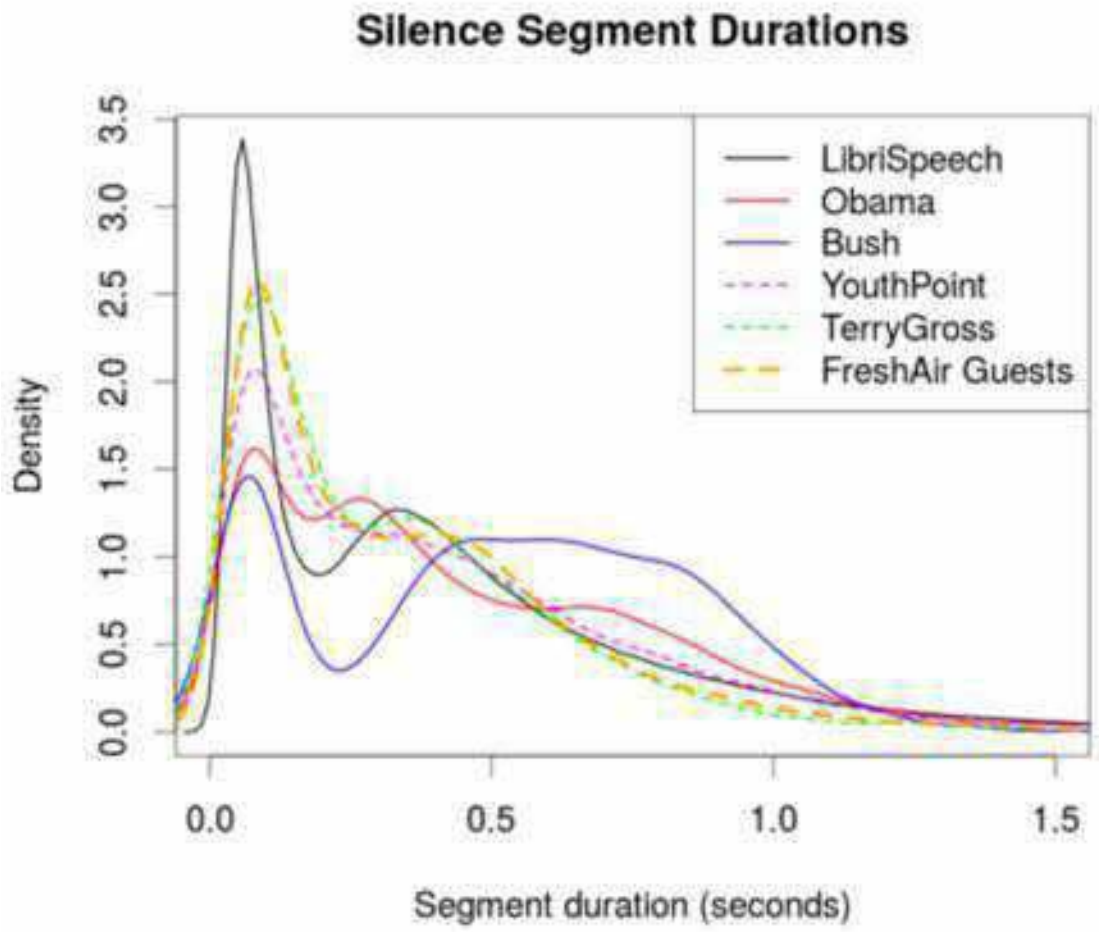
Does this distribution of speaker characteristics mean anything?

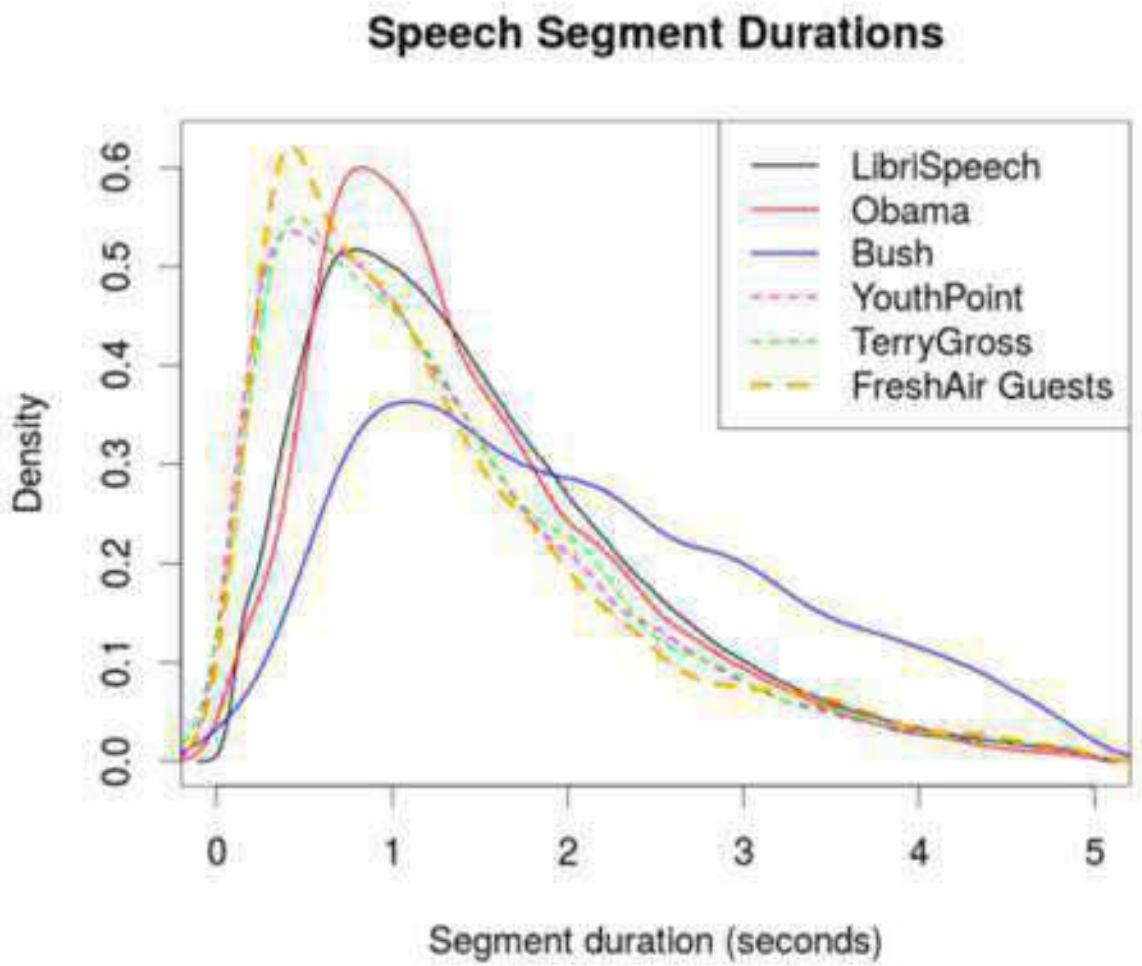
Let's compare some other sources of spontaneous and read speech.

Spontaneous: Fourteen *Fresh Air* radio interviews, involving public figures ranging from Lena Dunham to Stephen King to Gloria Steinem. The host Terry Gross is treated separately from the interviewees.

Spontaneous: A radio program produced by students at the University of Pennsylvania in the late 1970s. Our data set includes a subset of 50 sessions with 57 interviewees, including Ann Landers, Mario Andretti, Francesco Scavullo, Mark Hamill, Annie Potts, Chuck Norris, Buckminster Fuller, Erica Jong, Chaim Potok, Isaac Asimov, Ed Muskie and Joe Biden.

Read: 50 weekly radio addresses given by George W. Bush during 2008, and 127 weekly addresses and prepared statements given by Barak Obama between 2009 and 2011.





The distribution of read speech segment durations
seem to shift towards longer segments
compared to the spontaneous speech segment durations.

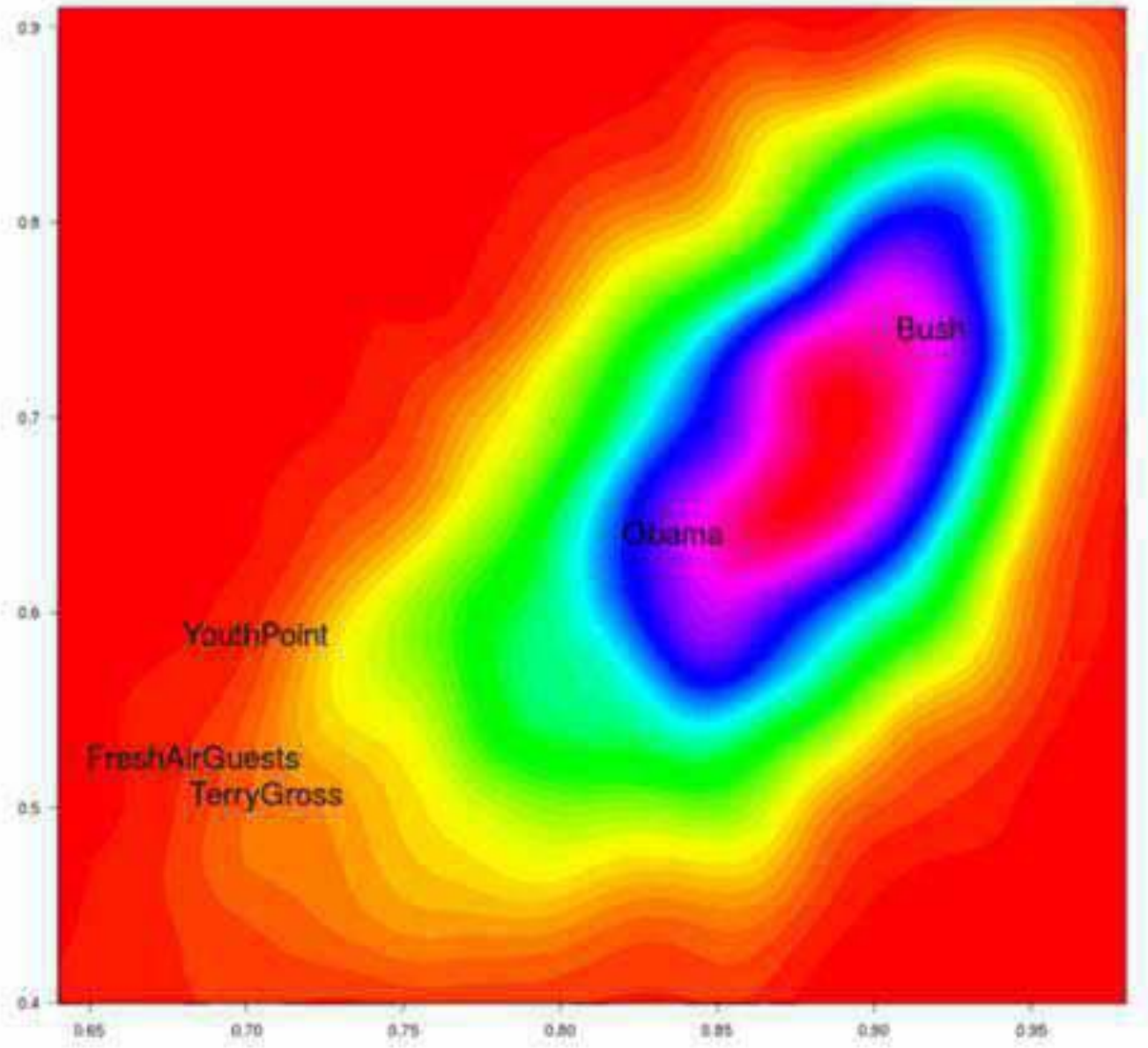
And the 2D density plot
shows this effect clearly:

Obama and Bush
are on opposite sides
of the modal region of readers.

All of the conversational speakers
are down in the tail.

Having background data from
2,484 readers is essential
to establishing this pattern.

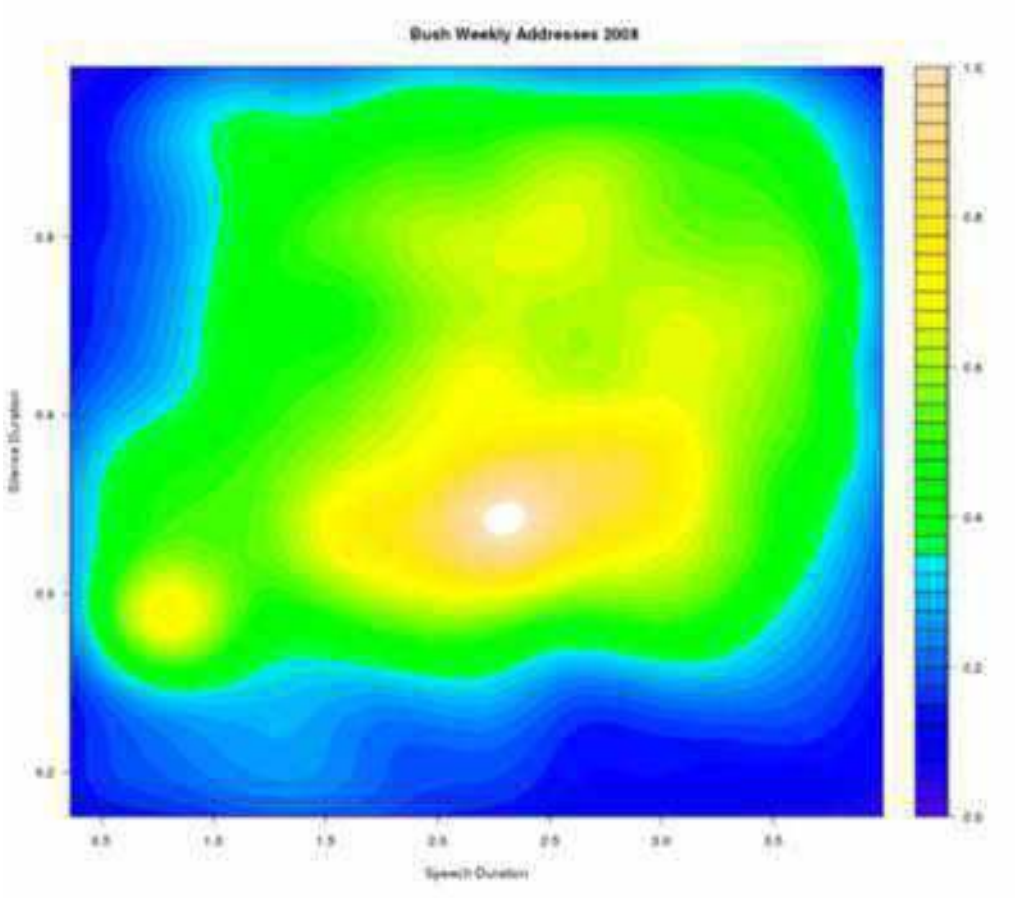
[From Ryant & Liberman, IS 2016]

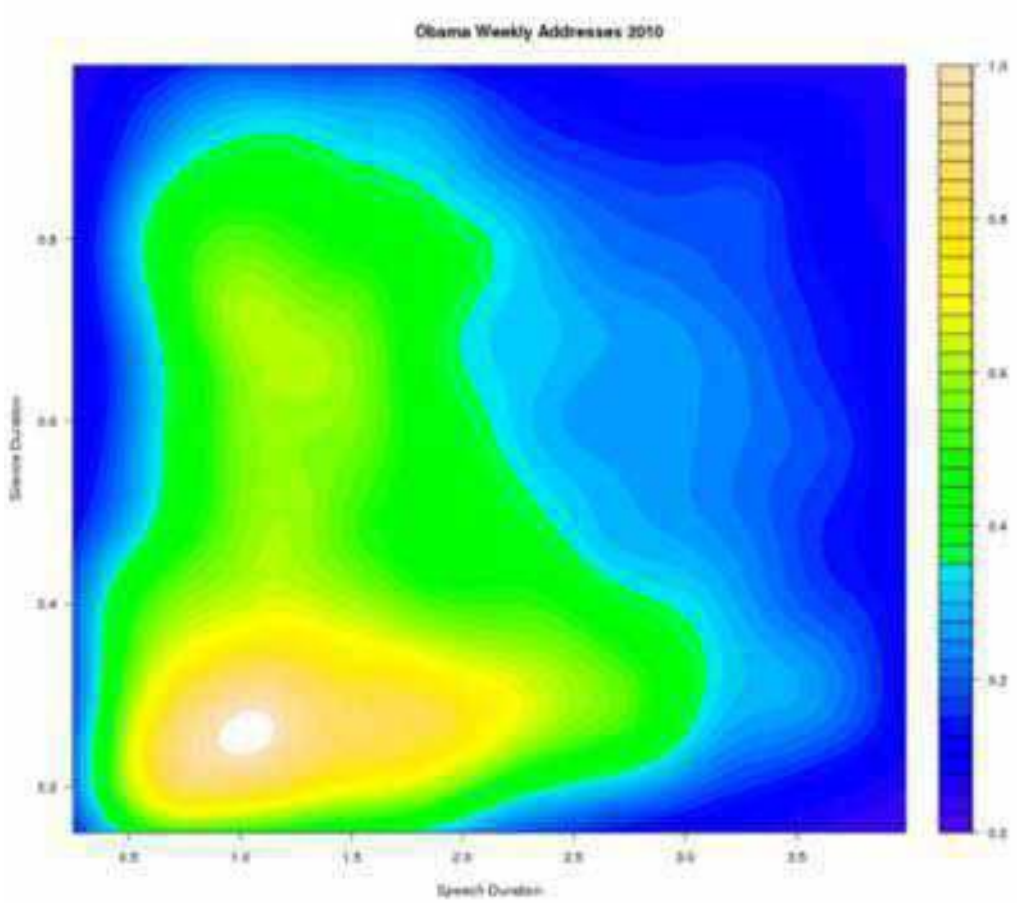


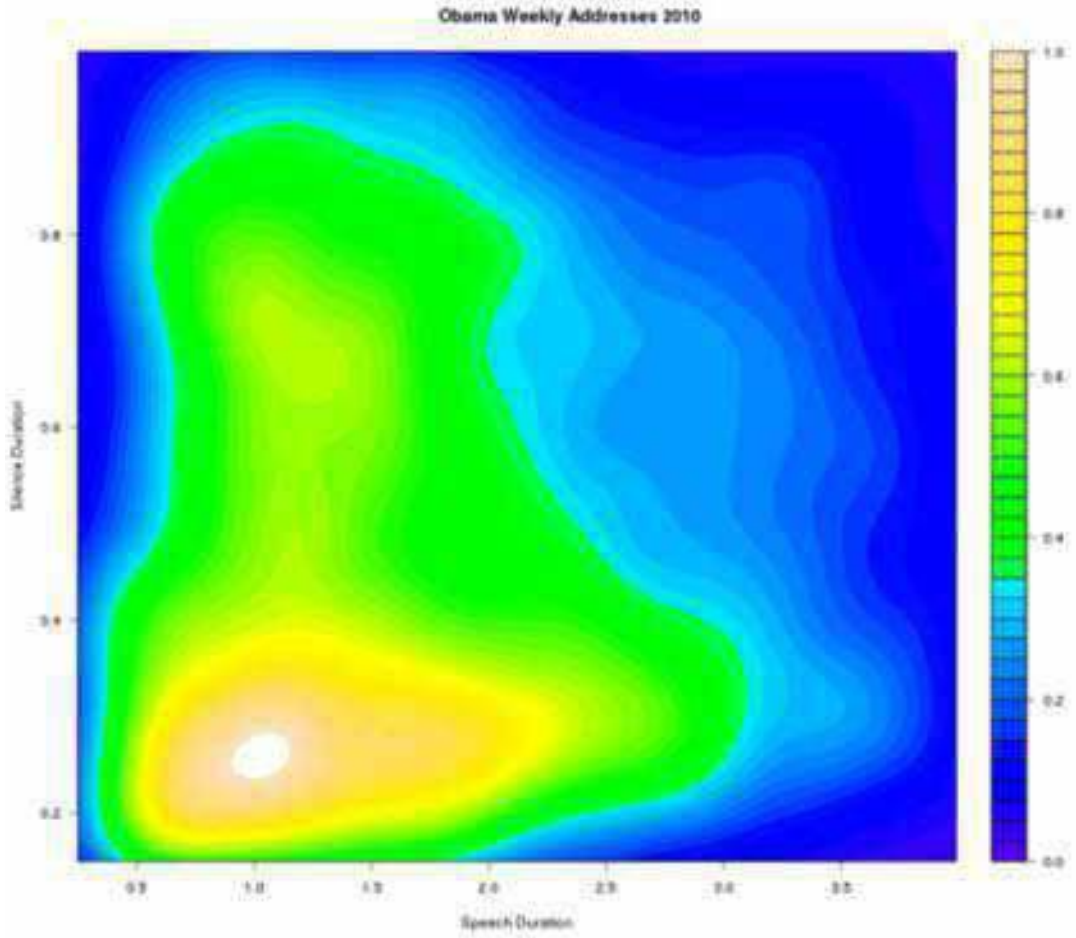
There's more structure to explore even in this trivial speech/non-speech data –

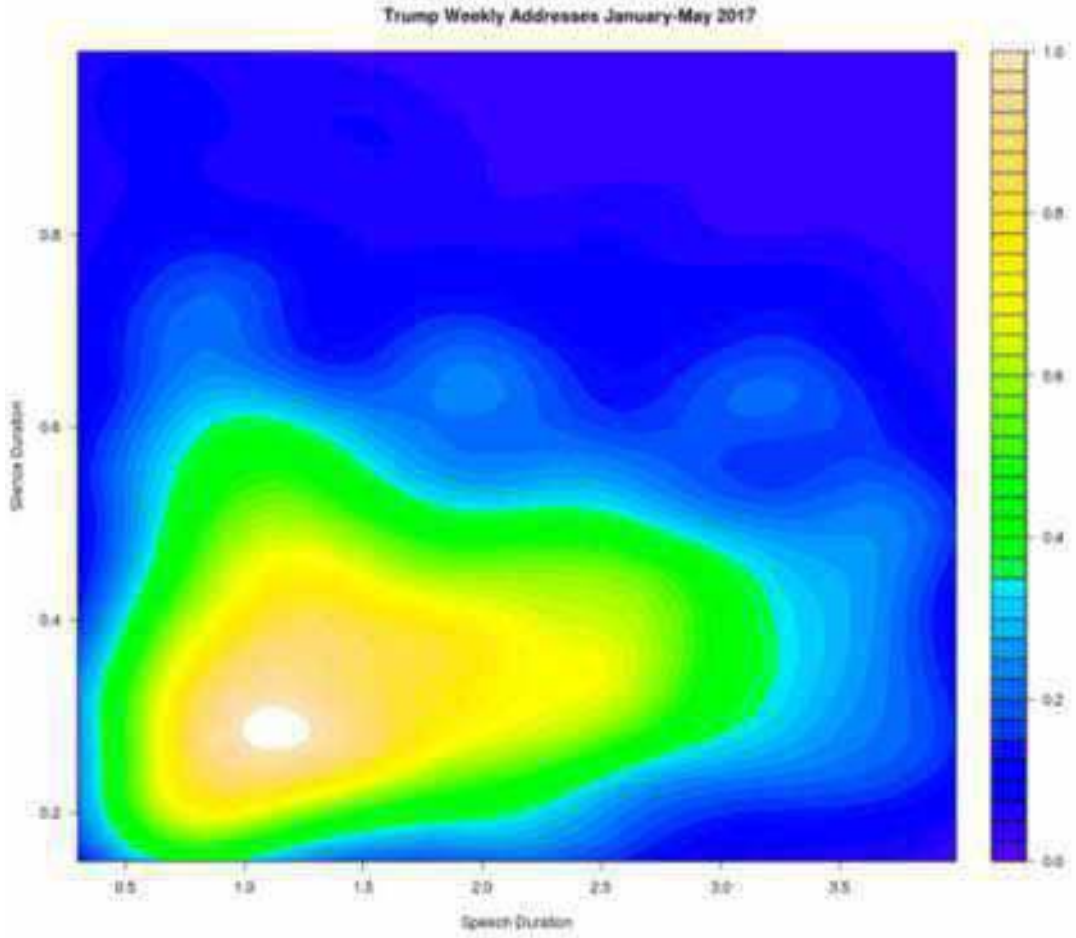
Here are 2-D distribution of speech segment durations

and immediately following silence segment durations:






Obama Weekly Addresses 2010
Silence Duration
Speech Duration


Trump Weekly Addresses January-May 2017
Silence Duration
Speech Duration

The good news:

Human Language Technology
brings to scientific or scholarly investigations

- easy re-use of existing digital datasets;
- analysis and tabulation with orders of magnitude less human labor.

We can work on a scale several orders of magnitude larger than before.

We can test hypotheses in minutes, hours, or days,
rather than weeks, months, or years.

And we can explore large and complex datasets interactively,
to see patterns and generate descriptions.

An optimistic picture of speech research in 2030:

1. Much more acoustic data
2. A much larger number of more sophisticated researchers
3. Strong requirements for publication of data and code via “notebooks” like Jupyter
4. Standards + easy access to well-designed tools and distributed datasets
5. Significant amounts of open articulatory, physiological, perceptual data
6. More complete automation
    a. Accurate universal pronunciation modeling in forced alignment
        i. First for major languages
        ii. Then for less-resourced languages
    b. Accurate language-independent phonetic analysis
        i. Phonetic feature recognition
        ii. Forced alignment “trained” by listing phonetic features of phonological categories in context
    c. ASR good enough for automated phonetic analysis without a transcript

Let’s make it happen!

?