(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 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. 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. And we can hope for convenient, inexpensive, and psychometrically reliable ways to estimate the efficacy of treatments.

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.

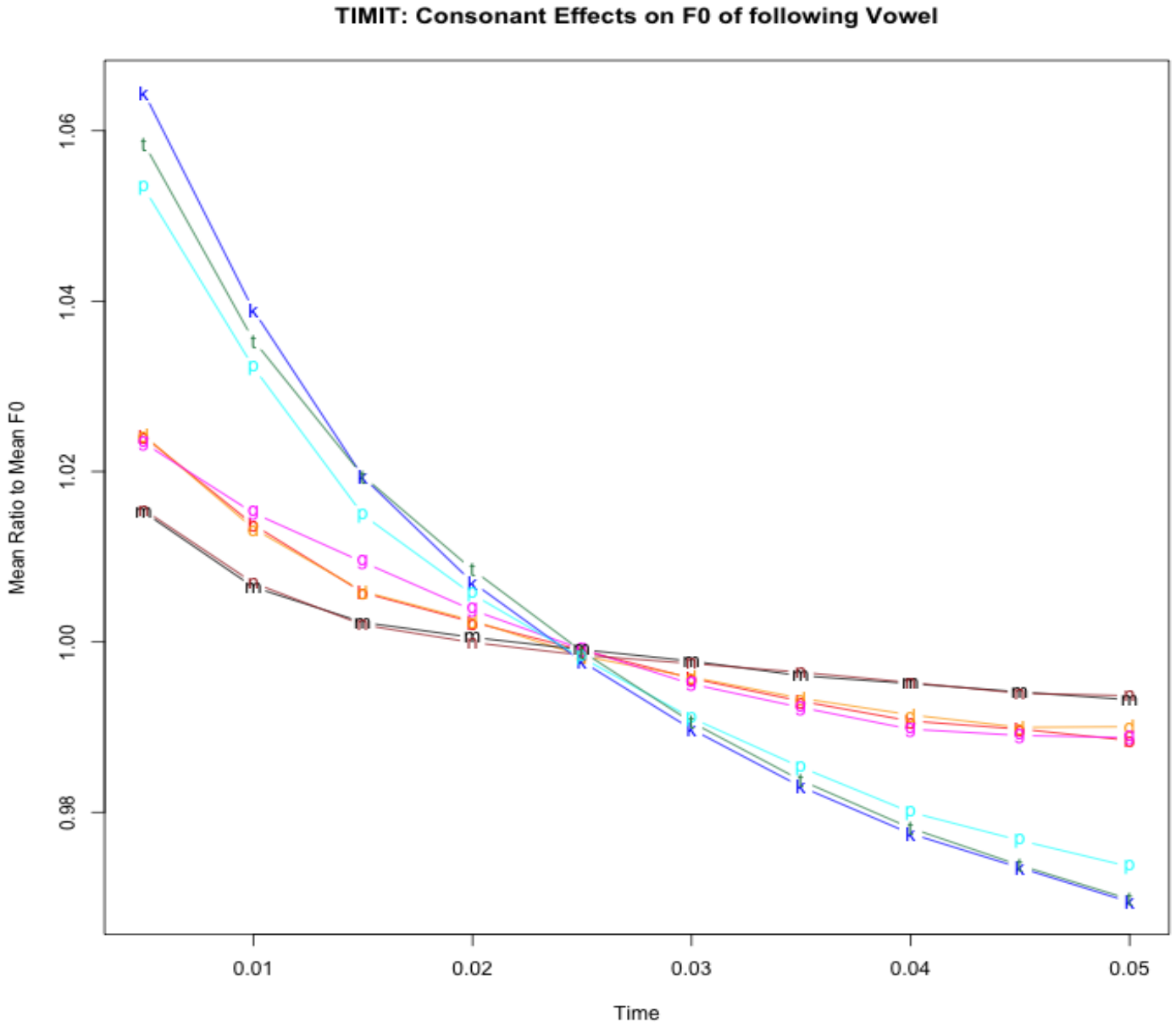
So over breakfast the next morning,
I checked out syllable-initial consonants in the TIMIT corpus.

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 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

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

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


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.

But major challenges remain.

Many annotation problems remain substantially unsolved, such as diarization of real-world conversations.

And there are kinds of data that are not generally available, or not available at all.

In this talk, I'll focus an important area of inadequate data:

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 (but falsely) believed
to make cross-site sharing impossible.

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

Because speech and language are provide key behavioral markers,
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.

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.

As exemplified in the this afternoon's talks,
even small and limited datasets
yield promising results from simple techniques.

This motivates a serious effort
to find ways to share clinical speech and language data
in consistent and accessible ways
on a large scale.

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.

Analysis of 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 larger collection.

As Bob Schultz, CAR's director, said:

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

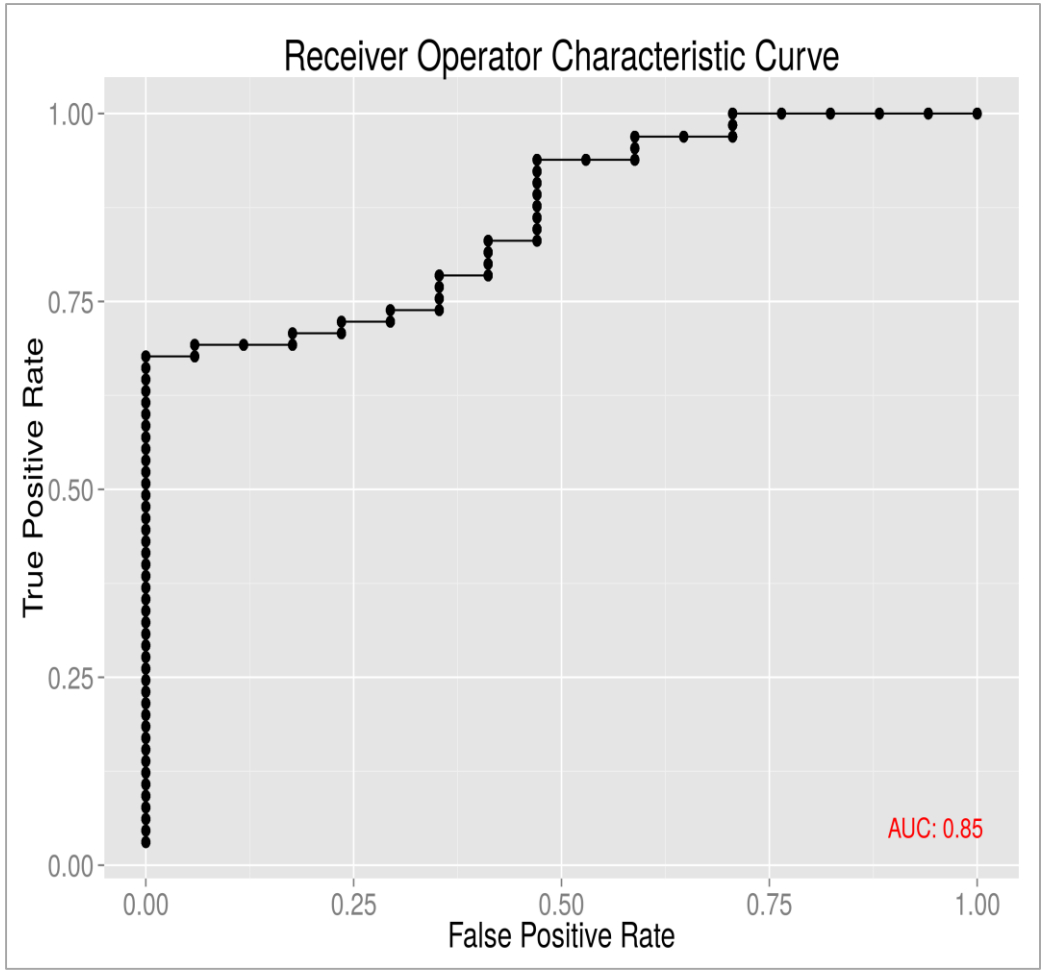
Simple bag-of-words classification – worked better than expected, as usual:

Naïve Bayes, weighted log-odds ratios

Leave-one-out cross validation

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

AUC=85%

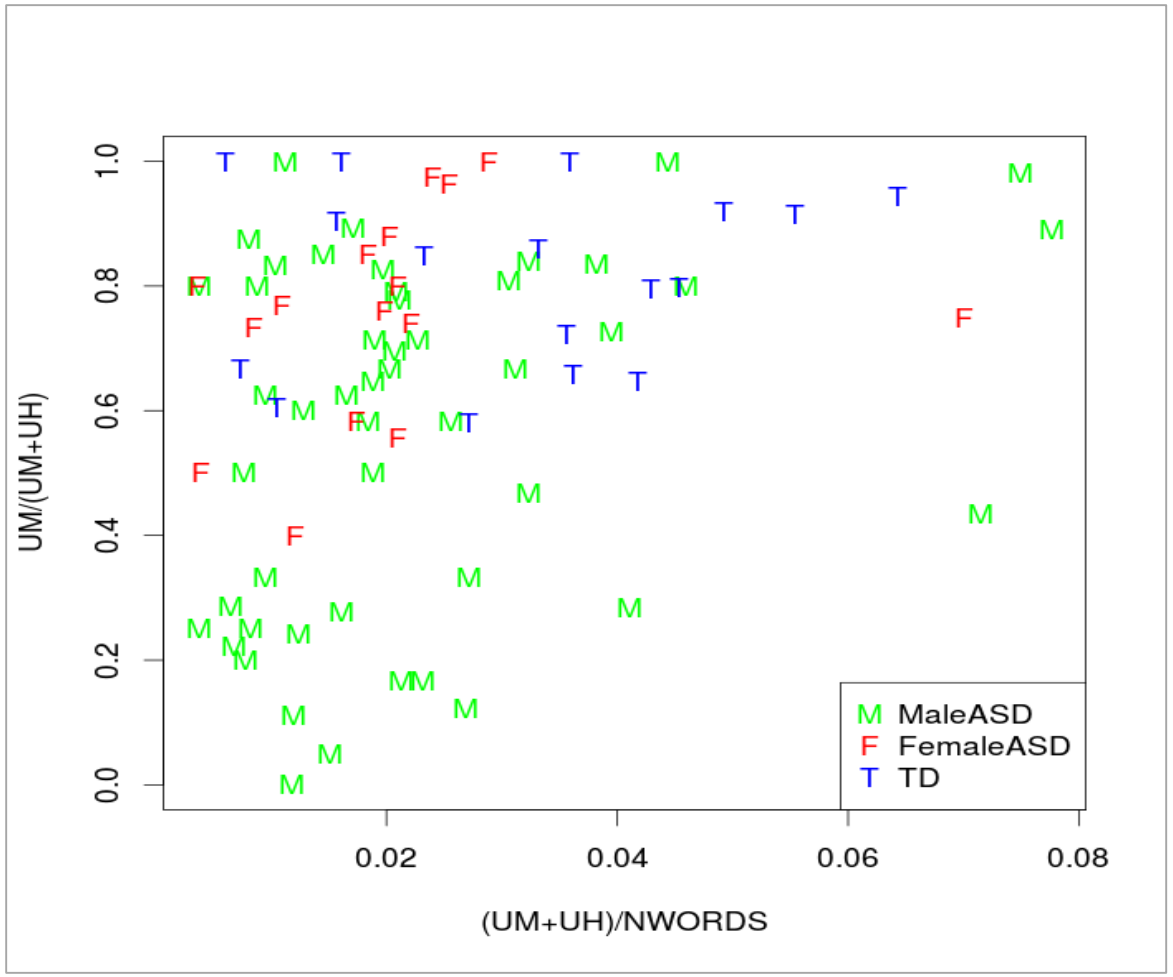


Rates of *UM* production in ASD and TD groups: ***um/(um+uh))***

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.

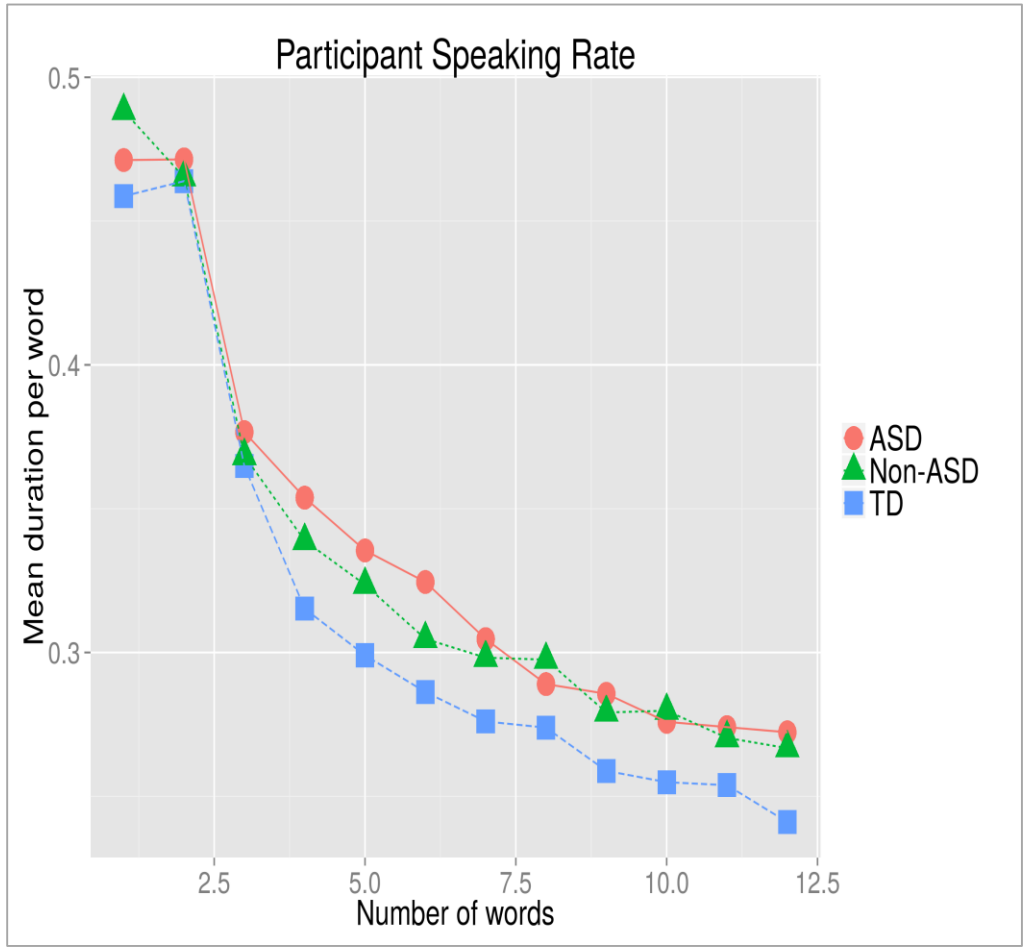


Mean word duration as a function of phrase length:

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)

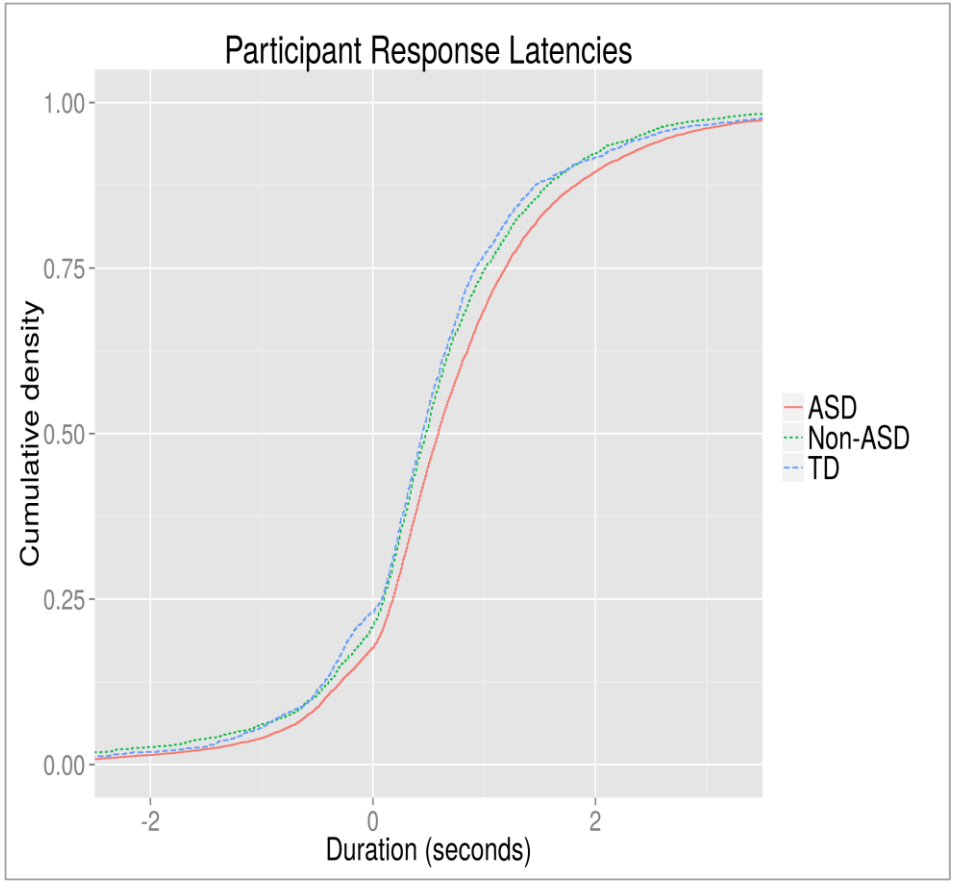


Latency to respond:

Too short = interrupting
speaking over a conversational partner

Too long = awkward silences
interfere with smooth social exchanges

ASD slower than TD

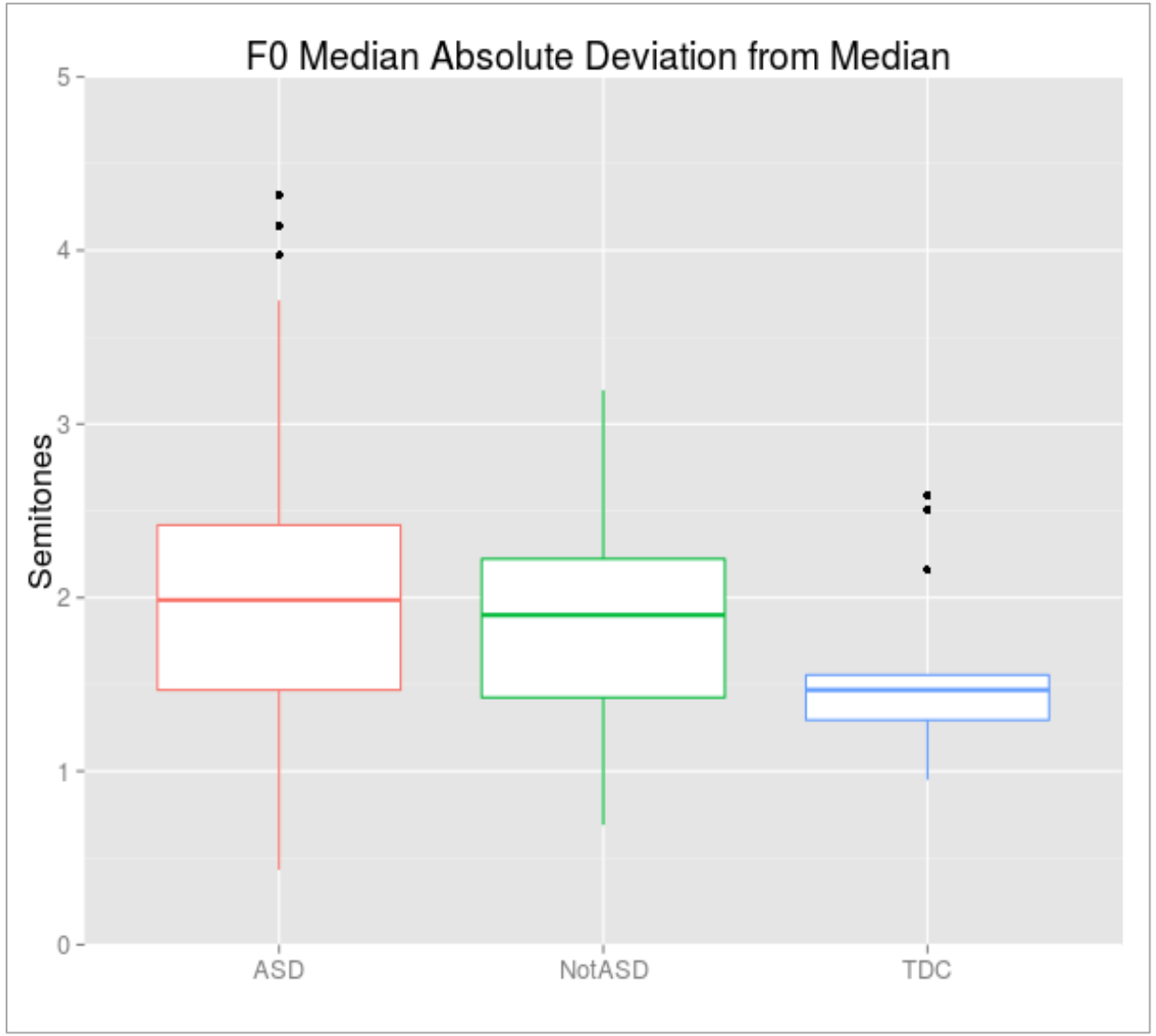


F0 Variation:

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 in-person expert collection --

We (also) need scalable automated methods
to collect large and 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,

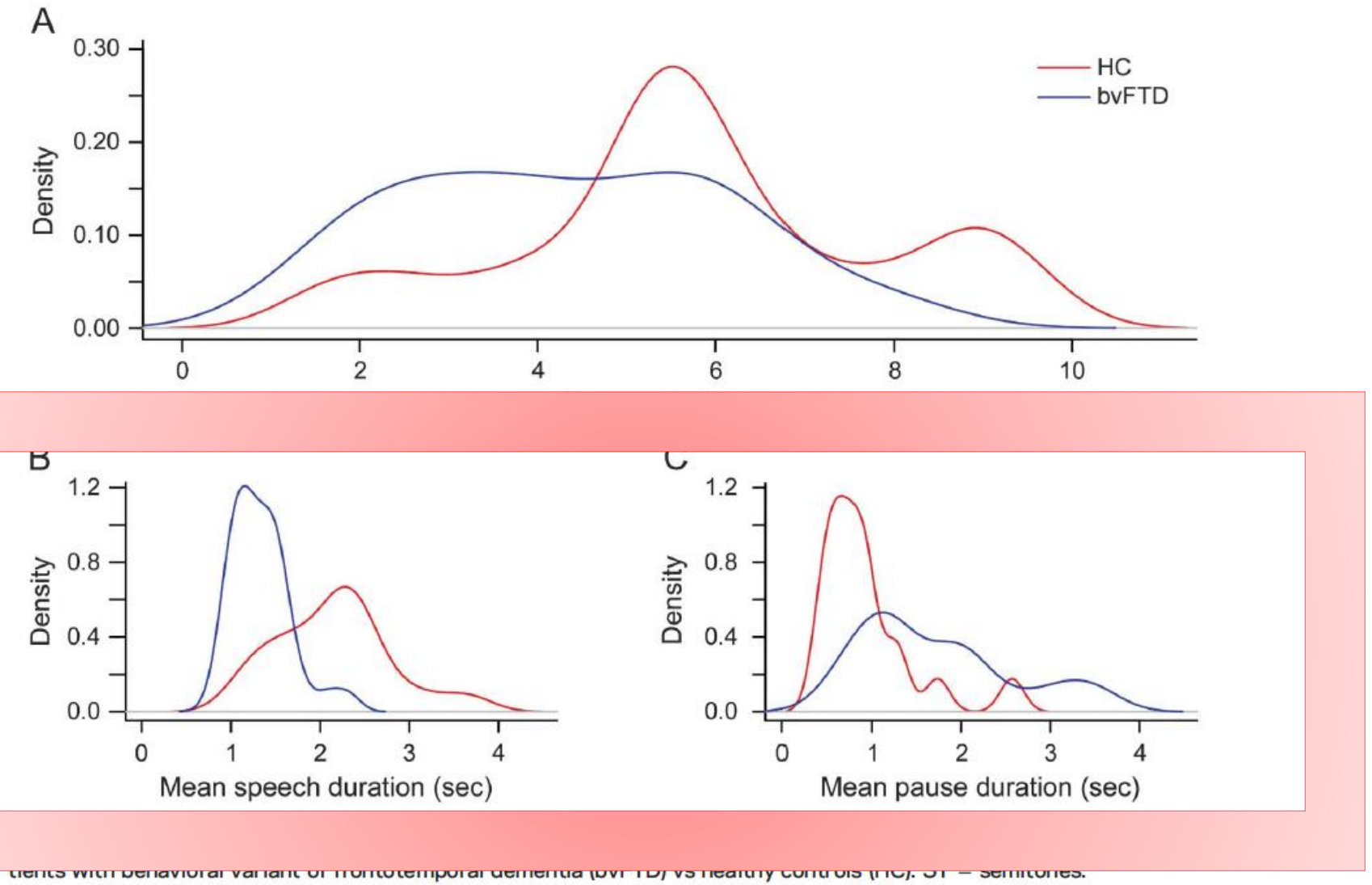
for example, the many diverse varieties
of neurodegenerative disorders,
such as Frontotemporal Degeneration,
Parkinsonism, and Alzheimers.

Again, in every case that we've looked at,
simple properties of speech and language data
correlate with clinical categories.

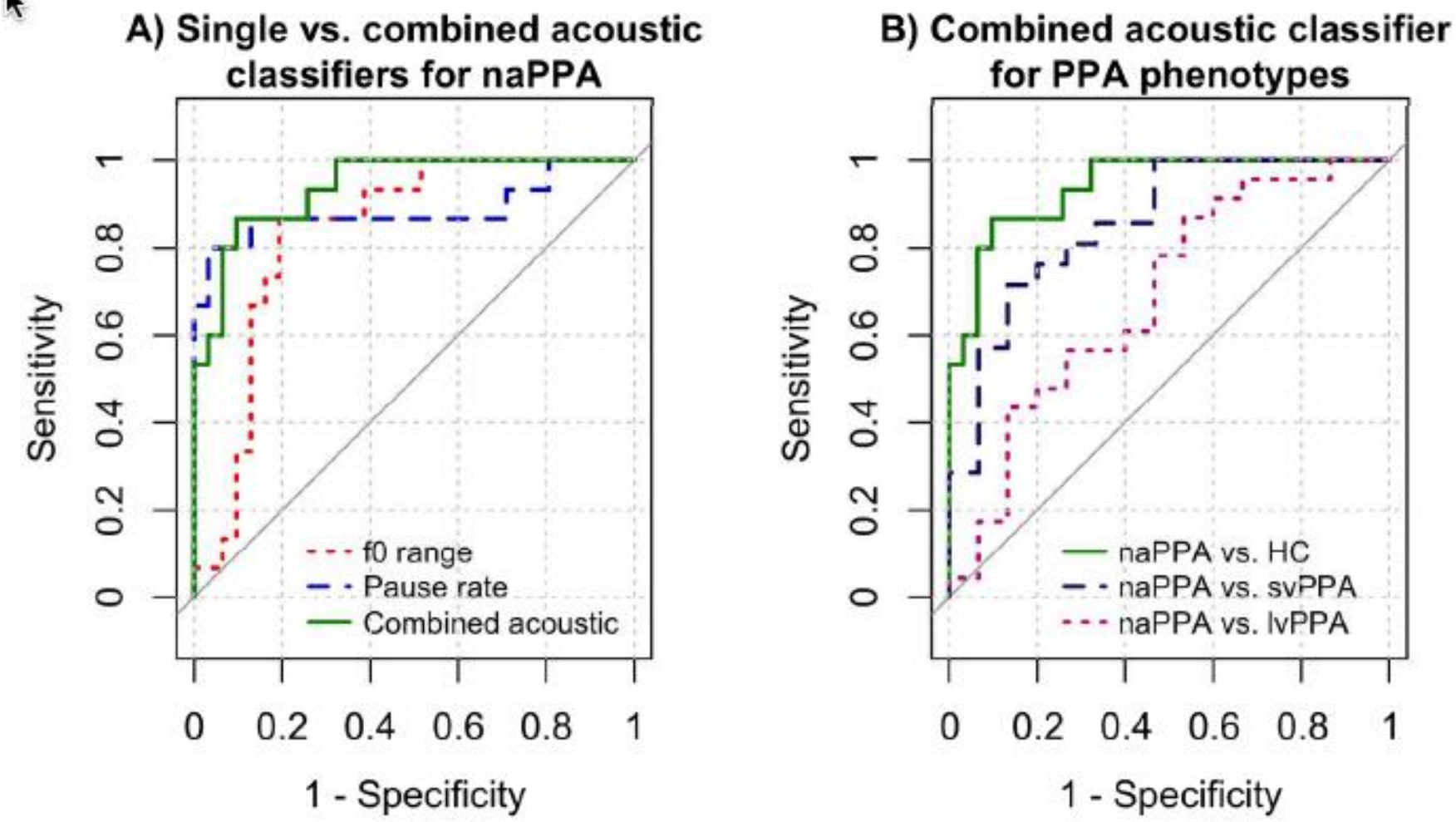
We're working with Penn's Frontotemporal Dementia Center
on a dataset of picture-description recordings
from ~1200 patients and elderly controls.

Simple (language-independent) acoustic-phonetic measures
have significant potential value in diagnosis and monitoring.

Figure 3 Speech measures distributions



tients with behavioral variant of frontotemporal dementia (bvFTD) vs healthy controls (HC). ST = semitones.


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

The good news:

- Picture descriptions give a lot of diagnostic information (…that's why neurologists use them…)
- The task is quick, and easy to automate (e.g. using a web app)
- So this task could be part of a longitudinal test battery measuring linguistic and cognitive skills across time

The bad news:

- Repeated description of the same picture is problematic
- There are only a couple of commonly-used pictures
- Even for those, there’s no basis for psychometric norming

Therefore we're planning to

- Create ~50 suitable pictures or short animations
- Get thousands of descriptions of each picture
  to permit psychometrically stable automated measures
- Combine with other standard tasks that can be automated
  - Digit span
  - "Fluency"
  - …etc…

With subject consent for data publication!