Sociolinguistics and Human Language Technologies

Or why we all need large data sets, automatic tools and sharing!

Thesis

- LDC and others collect LARGE data sets to drive speech technology research (LID, ASR, DID, etc)
- LARGE =
  - *Hundred/Thousands of hours of data per language/dialect*
  - *Hundreds/Thousands of speakers*
  - *E.g. mixer, fisher, HUB4-5, etc*
- Many of the technologies that have been developed could support dialect/variation research!
  - *Analysis of large data (word usage, pronunciation, etc.)*
  - *Measurement of speaker/dialect variability (intra and inter)*
  - *Measurement of channel affects*

Case 1

British English vs. American English

- WSJ (US English): 200+ hours of read speech
- WSJ-CAM0 (British): 90+ hours of read speech
- 200+ speakers
- Use ASR techniques to learn pronunciation models

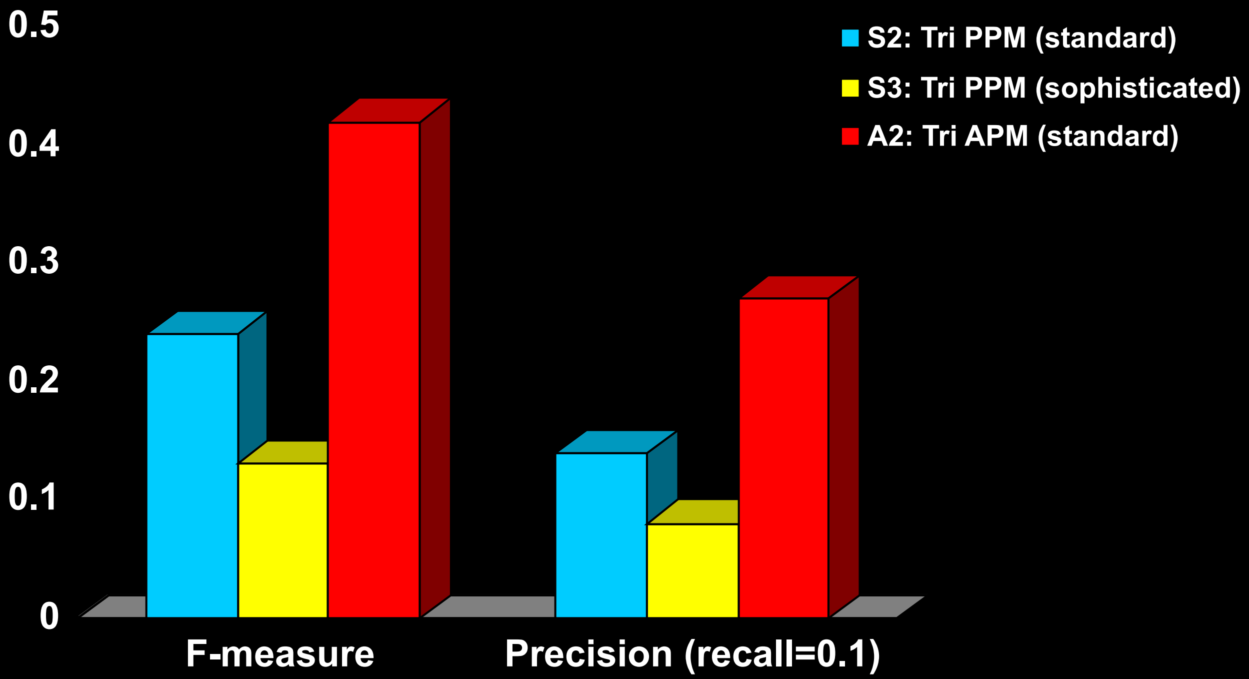
Literature	Proposed System	
Rule	**Learned Rule**	**Prob**
[ae] -> [aa] /_ [+fric, -voiced] (trap-bath split)	[ae] -> [aa] /_ [+fric, -voiced, +front]	0.84
	[ae] -> [aa] / [-voiced]_ [+fric, -voiced, -front]	0.52
[r] -> ø / _ [+cons] (R Dropping)	[er]$_{ins}$ -> [ah] / [+vowel] _ [+affric]	1.0
	[er] -> [ah] / l _ [+affric]	1.0

We rediscover known rules *AND automatically measured prevalen*

Case 2

AAVE/non-AAVE variability

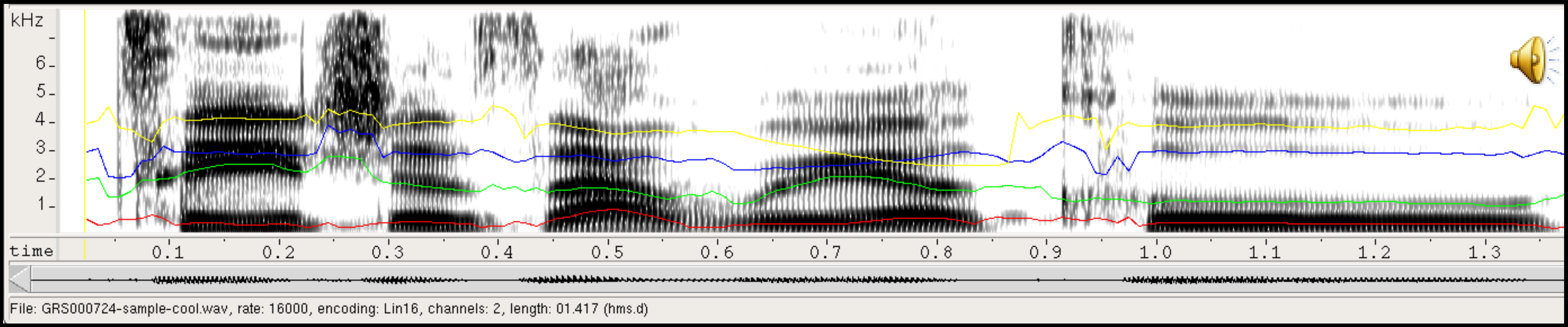
- StoryCorps: oral history collect of AAVE/non-AAVE talkers
- Simultaneous collection in 15 US cities for NPR
- 300+ speakers, 400+ hours / dialect
- Automatically identify and retrieve instances of AAVE specific transformations (21 from Wolfram 2005)



Mining data for analysis

Using the model to explore your corpus

Learned rules: uw-[l]: uw-l



File: GRS000724-sample-cool.wav, rate: 16000, encoding: Lin16, channels: 2, length: 01.417 (hms.d)

Sur.	t iy ch ih z	aa r	r iy l	k uw
Ref.	t iy ch er z	aa r	r iy l	k uw l
Words:	Teachers	are	real	cool

This is just the beginning

With more data we will be able to:

1. Characterize in-dialect speaker variability
2. Measure acoustic variability that is too subtle for categorical labeling (see [Shen 09] and [Chen/Shen 11])
3. Learn rare transformations that are difficult to observe in small data sets. [Chen 10] proposed 700+ AAVE-specific pronunciation transforms
4. Speed data analysis: find regions of dialectal difference using automatic methods