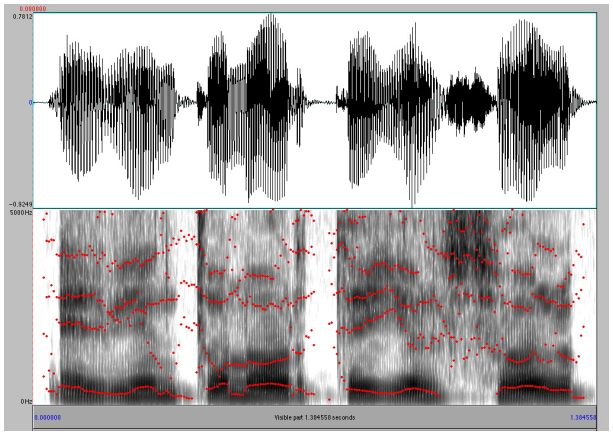
Is the Future Almost Here?

Large-Scale
Completely Automated Vowel Extraction
of Free Speech

Sravana Reddy and James N. Stanford

Dartmouth College

Motivation



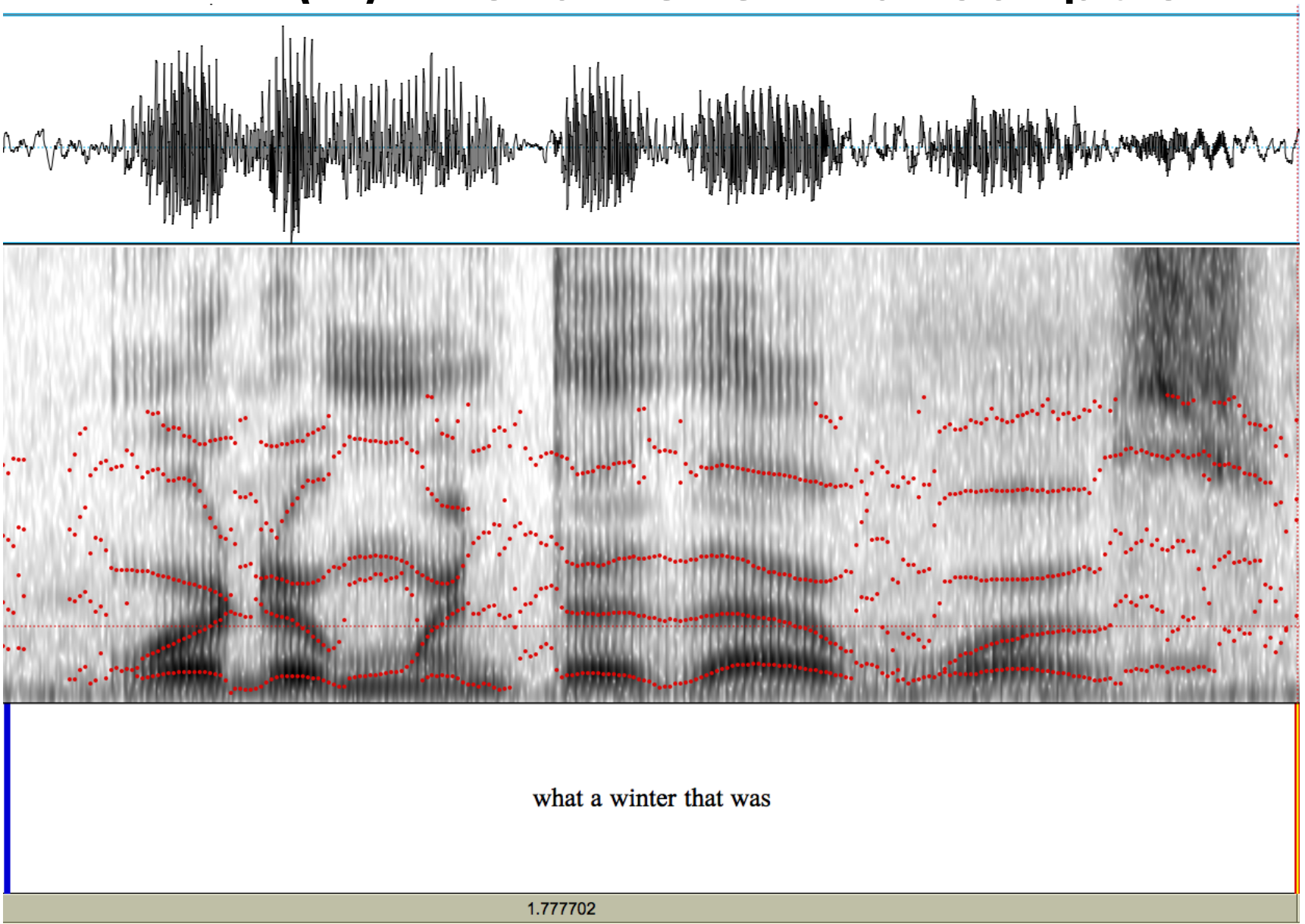
Transcription

Can this be completely automated?

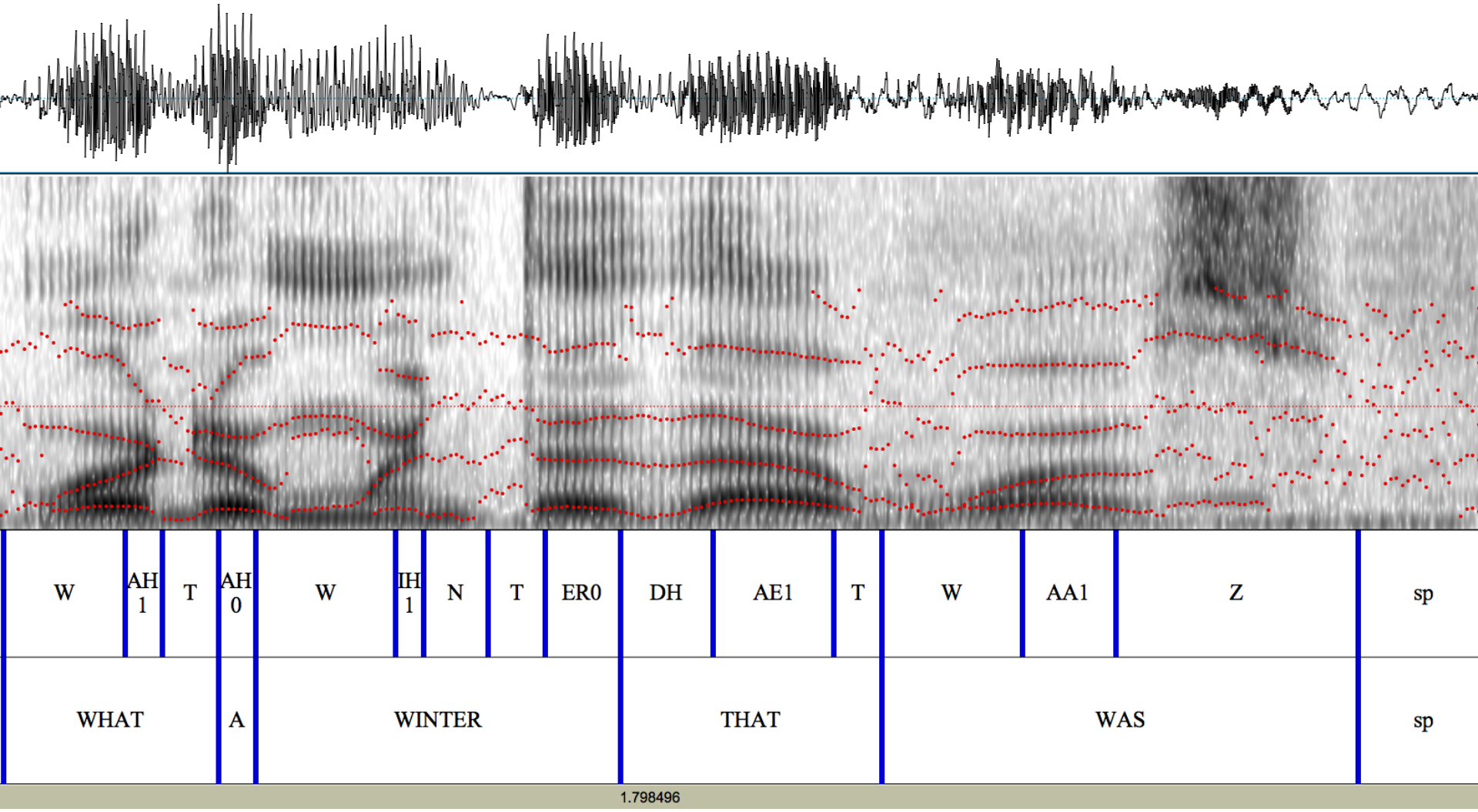
Current Level of Automation

- Penn Aligner (Yuan & Liberman 2008)
  - Evanini (2009)
  - Evanini, Isard & Liberman (2009)
- ProsodyLab (McGill) Aligner (Gorman et al. 2011)
- WebMAUS (Kisler et al. 2012)
- FAVE: Forced Alignment Vowel Extraction (Rosenfelder, Fruehwald, Evanini & Yuan 2011)
  - Used for Philadelphia data analysis in Labov, Rosenfelder & Fruehwald (2013)
  - Fruehwald & Kendall at this conference

FAVE: (1) Word-Level Transcription



FAVE: (2) Forced Alignment



FAVE: (3) Vowel Extraction

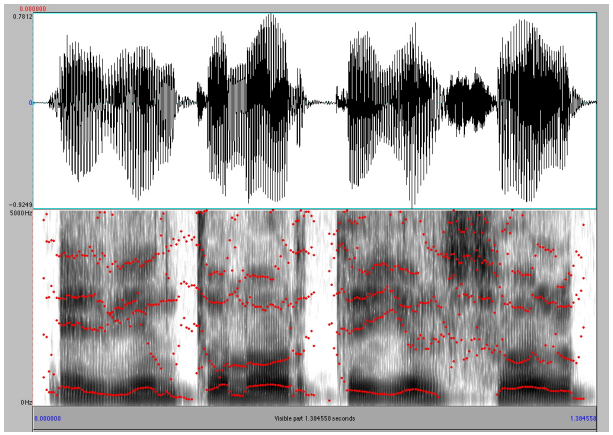
vowel stress word F1 F2 F3 B1 B2 B3 t beg end dur cd fm fp fv ps fs style glide F1@20% F2@20% F1@35% F2@35% F1@50% F2@50% F1@65% F2@65% F1@80% F2@80% nFormants

OW 1 NO 611.9 1644.7 2058.7 65.5 99.5 815.6 10.317 10.28 10.55 0.27 63 0 0 0 4 0 657.7 1599.0 610.0 1455.2 580.4 1160.2 546.1 1059.3 507.3 1037.8 5

AA 1 NOT 732.2 1493.6 2861.9 232.1 82.6 289.4 10.9 10.8 11.101 0.301 5 1 4 1 4 0 698.8 1484.7 739.8 1496.1 790.9 1503.2 796.4 1568.6 788.2 1646.4 4

AE 1 HAVE 592.4 1810.1 2135.6 49.8 125.7 699.1 11.467 11.43 11.54 0.11 3 3 2 2 0 0 610.0 1852.0 589.3 1800.2 577.6 1733.9 552.3 1656.6 479.2 1567.0 5

This Work



Word-Level Transcription

ASR (Automatic Speech Recognition)

FAVE

Vowel Formants

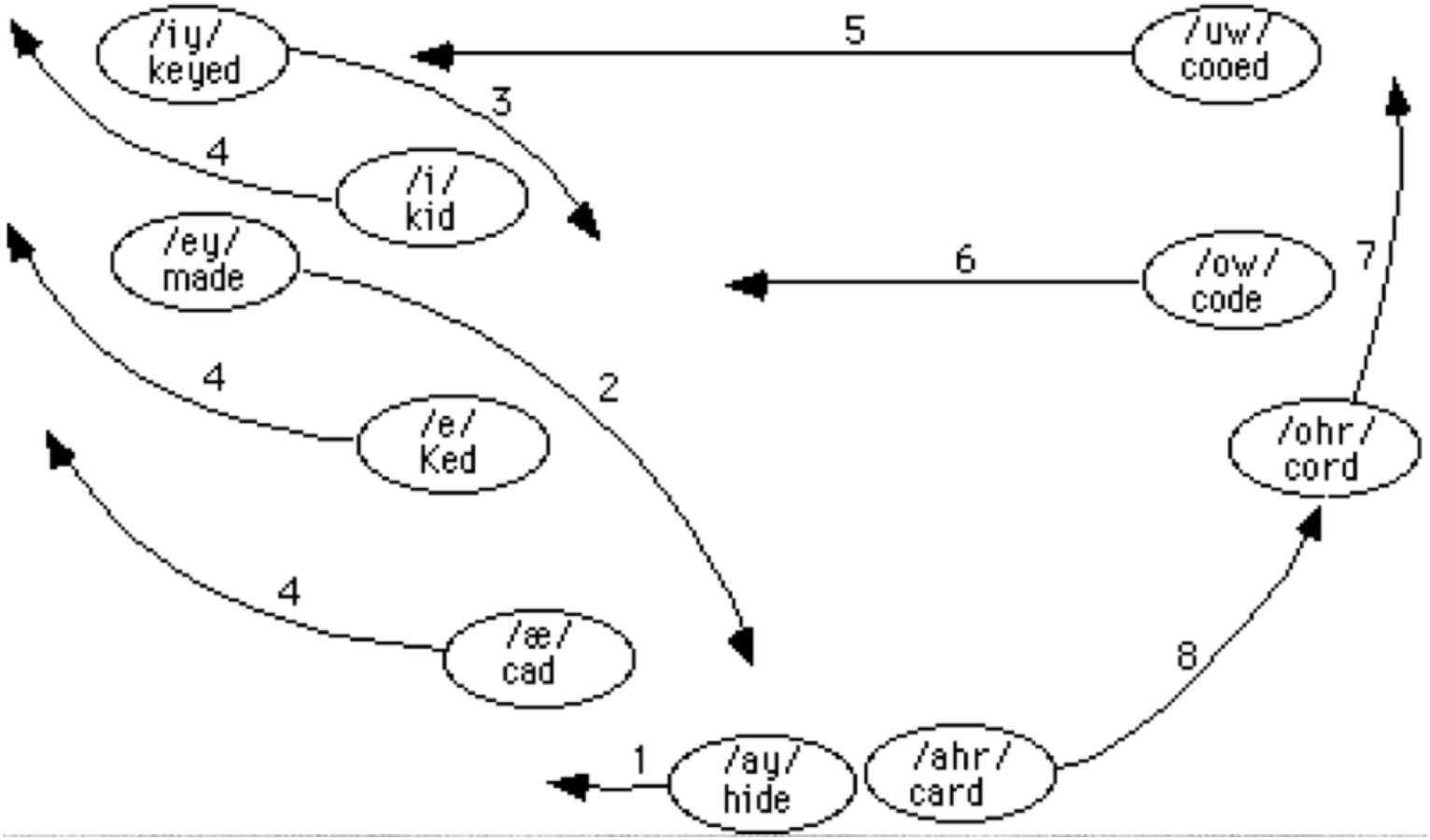
This Work

CAVE:
Completely Automated
Vowel Extraction

A future full of possibilities!

Analyze hours of speech from the radio and TV, terabytes of data from YouTube, live interviews, dialects of any language…

The Southern Shift



(Labov 1996)

Examples of ASR Errors

- REF: give me your first impressions
  HYP: give me yours first impression

- REF: it's one of those
  HYP: it's close

- REF: no it's it's wood turning
  HYP: no it it would turn it

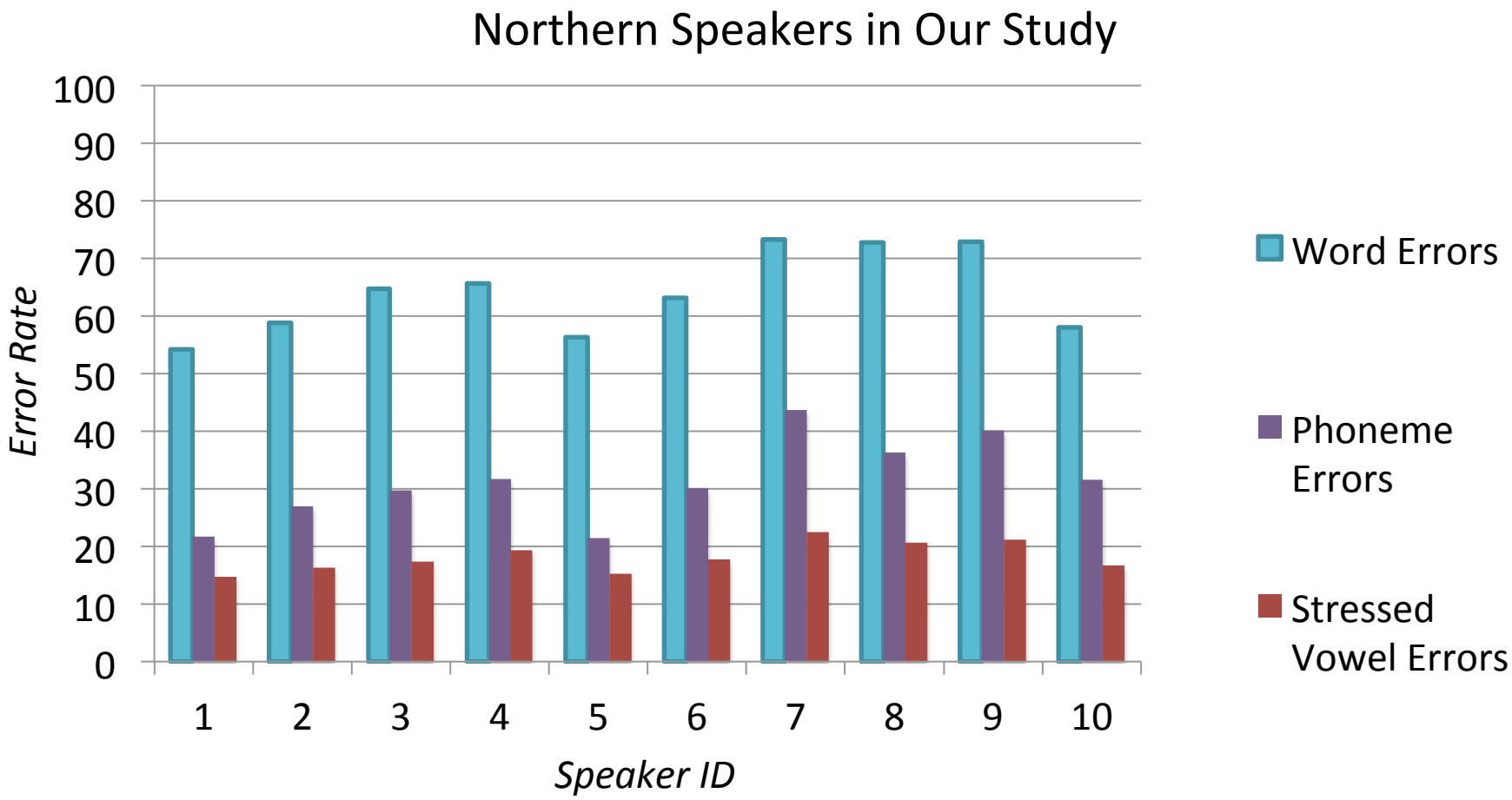
- REF: and we really don't spend on anything much
  HYP: and we don't depend on anything much

- REF: a real dog and cat and all the other animals
  HYP: a real docking tap and on the other animals

Poor understanding of meaning and syntax…

but the (stressed) vowels are ok!

ASR Word and Phoneme Errors



Our Idea

ASR vowel error rates are low.

With large amounts of data,
can get hundreds of tokens per vowel.

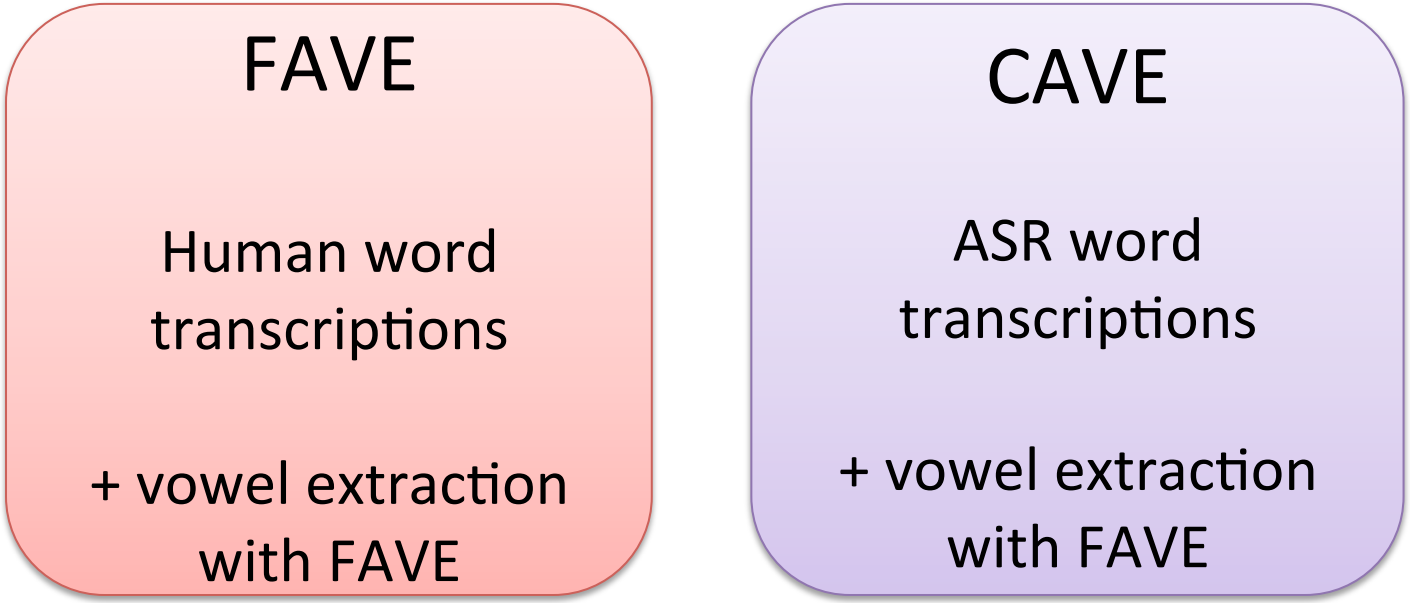
Therefore, ASR transcriptions should be nearly as good as human for analyzing vowels in sociolinguistics.

Technology behind FAVE

- Same models in automatic speech recognition
  - Forced alignment using MFCC features, acoustic models, dynamic programming…
- Natural question: take it further?

This Work

- Compare



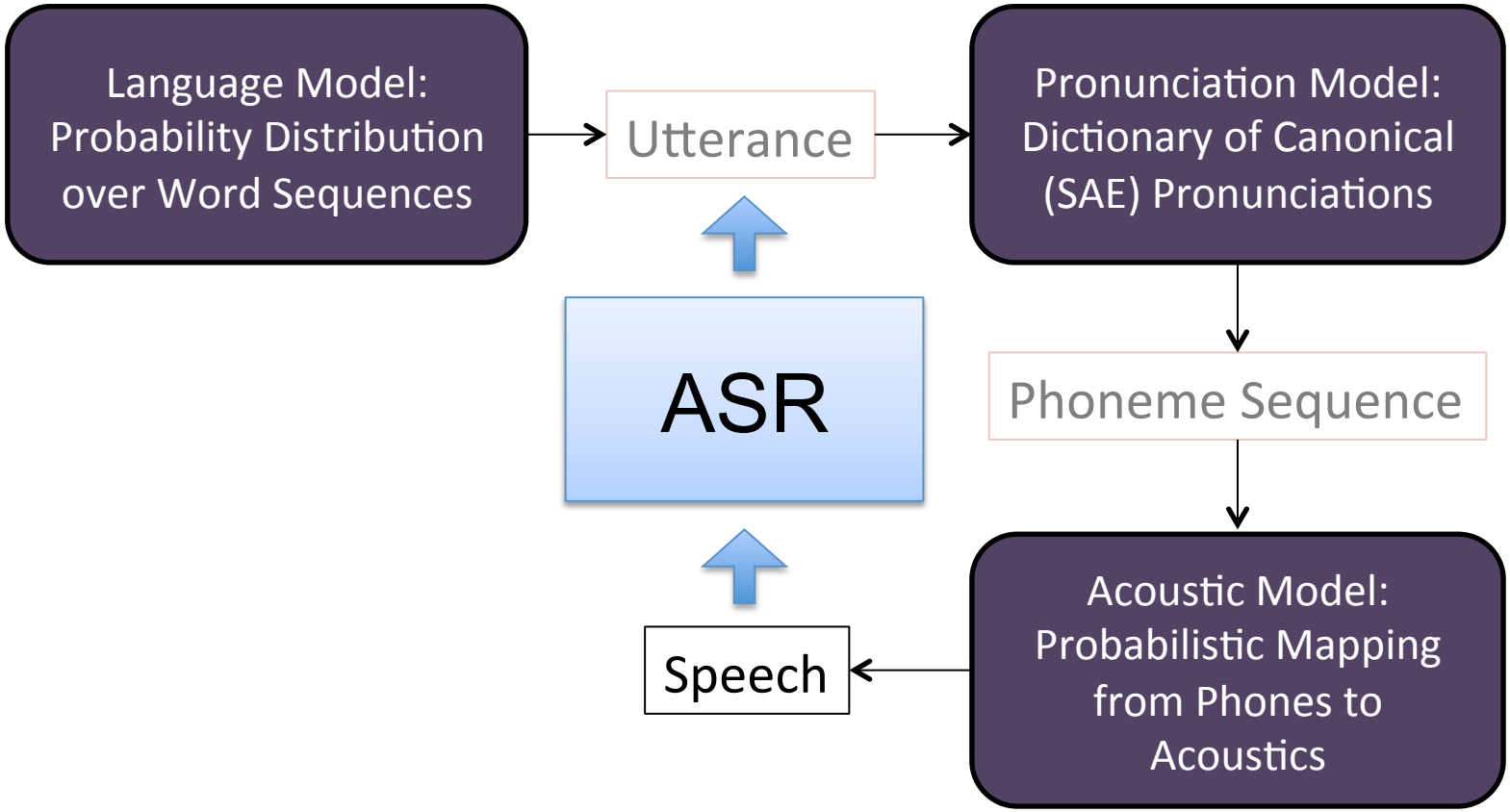
- Feasibility Test: Do the vowel spaces show a distinction between Northern and Southern dialect features?

Data

- Switchboard-1 Corpus (1997), available from the LDC
  `https://catalog.ldc.upenn.edu/LDC97S62`
- Two-sided telephone conversations between US speakers
- Includes human word-level transcriptions
- Randomly selected 20 speakers (15 hours of speech, 143266 stressed vowel tokens, approx. 300 tokens per vowel per speaker)

	Northern	Southern
Male	5	5
Female	5	5

Automatic Speech Recognition

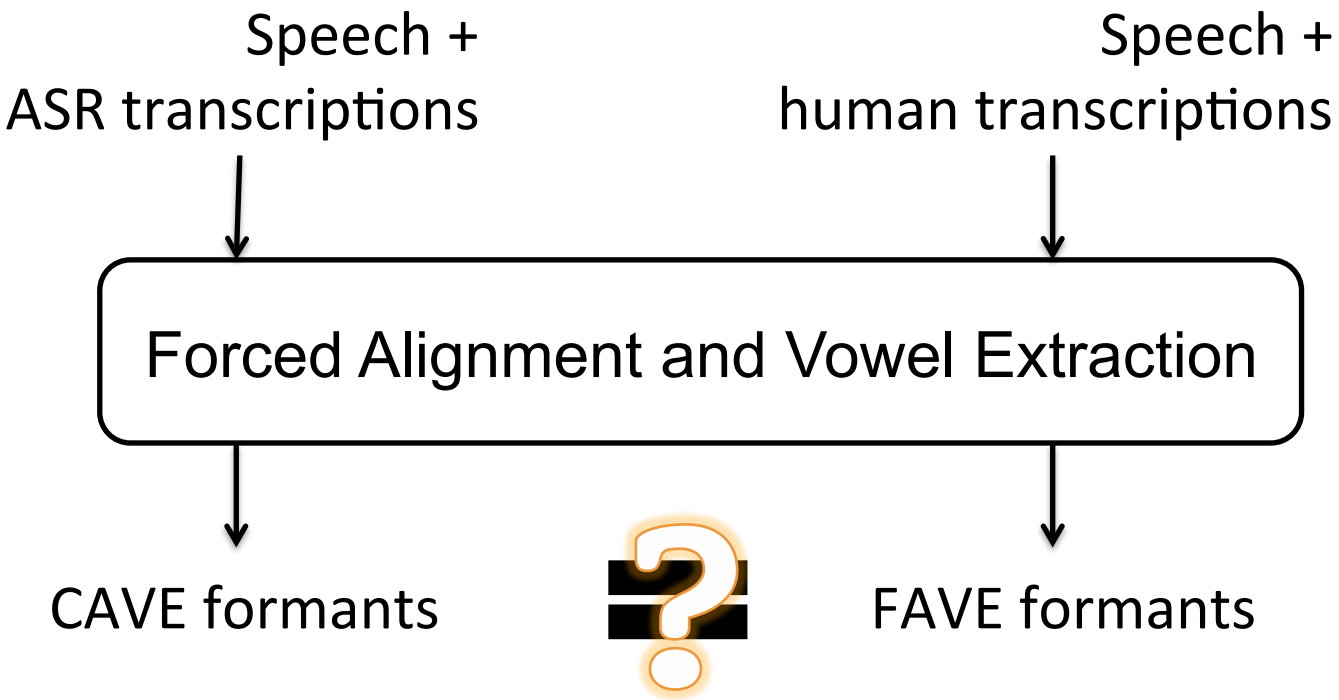


ASR System

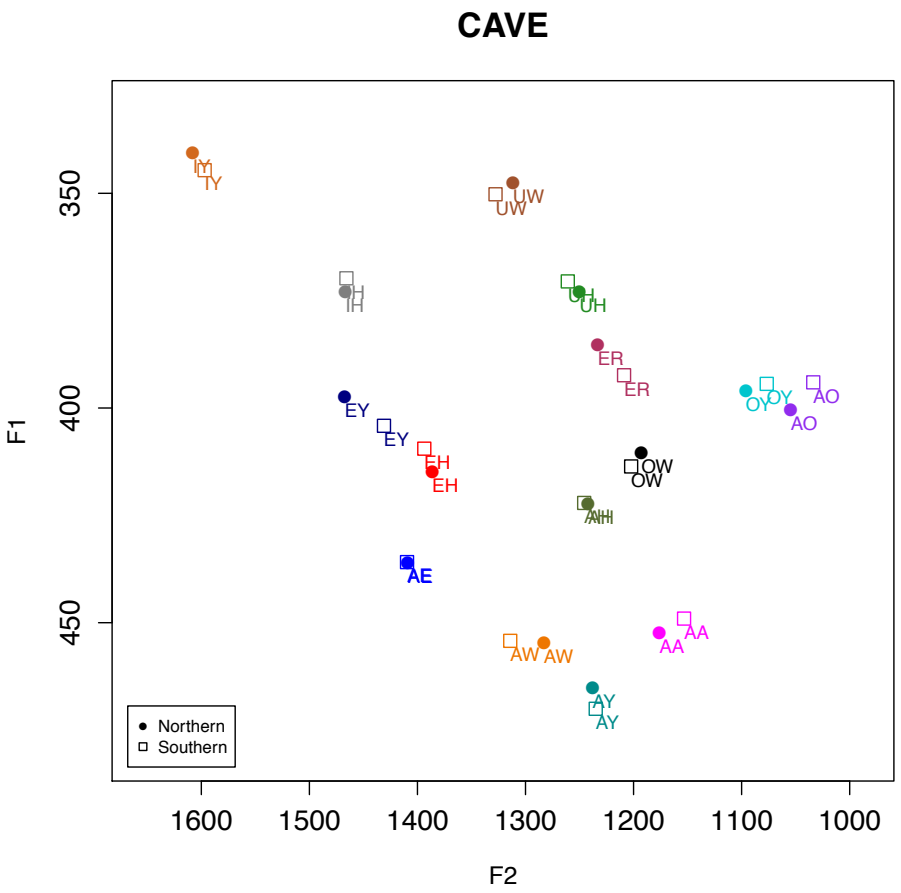
- We trained an acoustic model on US English speech (mostly newswire, some telephone)
- and a trigram language model on assorted US English corpora
- CMU pronouncing dictionary

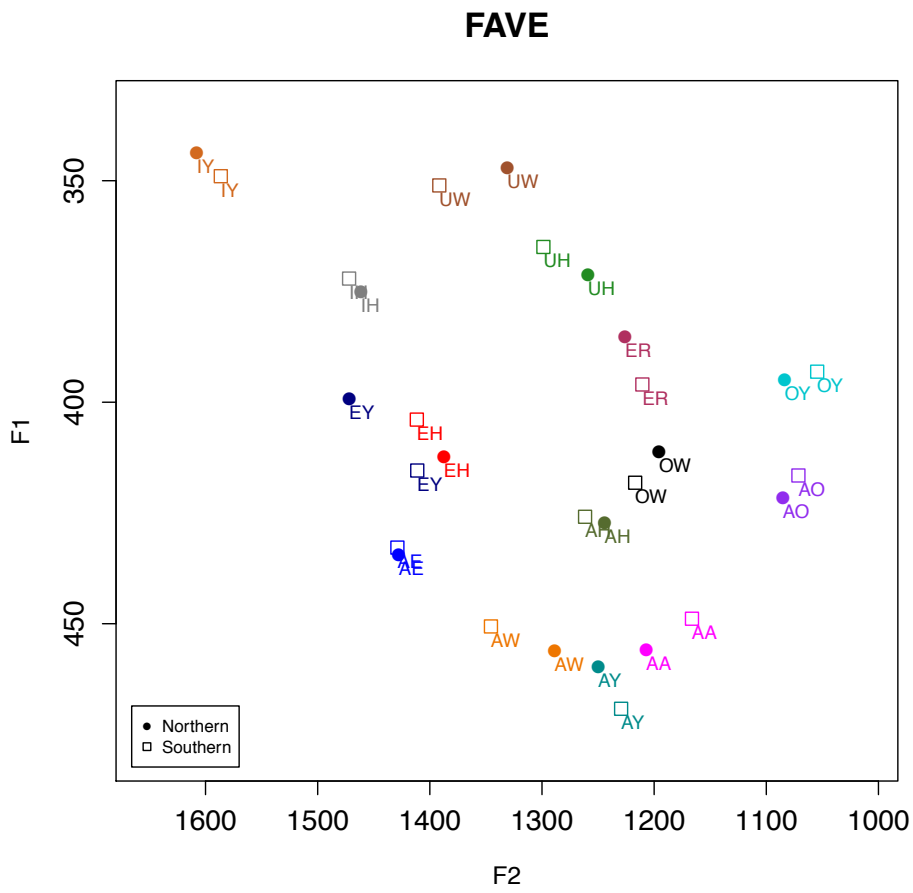
- Decoding with CMU Sphinx
  `http://cmusphinx.sourceforge.net`

Stressed Vowel Extraction



Results





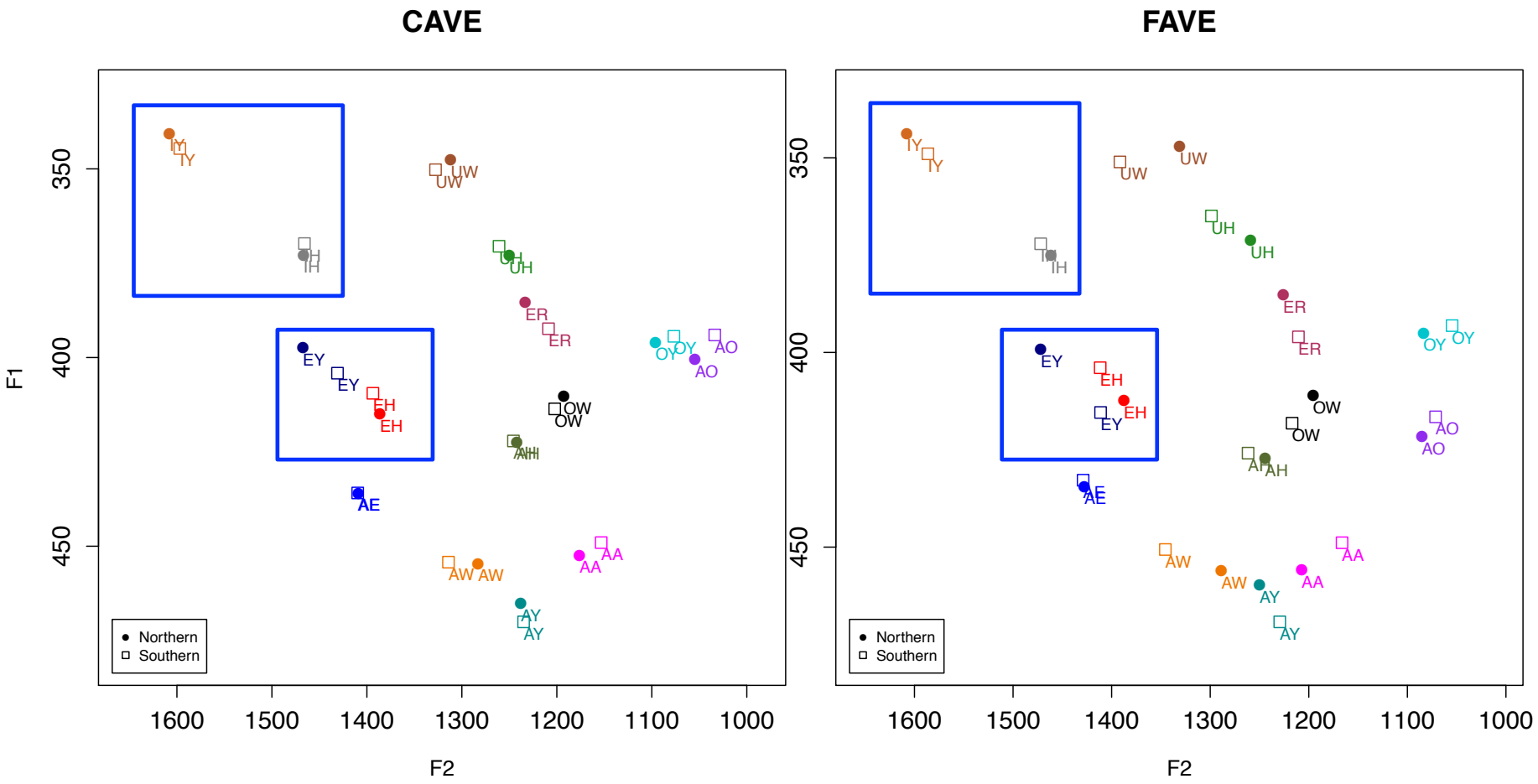
Normalized with Lobanov (Kendall & Thomas 2010)

Evidence of the Southern Vowel Shift

Both CAVE and FAVE

- ✔Clear north-south contrasts in EY/EH and IY/IH in the expected directions
  - ✔EY (bait) and IY (beet): lowered/backed for southerners
  - ✔EH (bet) and IH (bit): raised/fronted for southerners

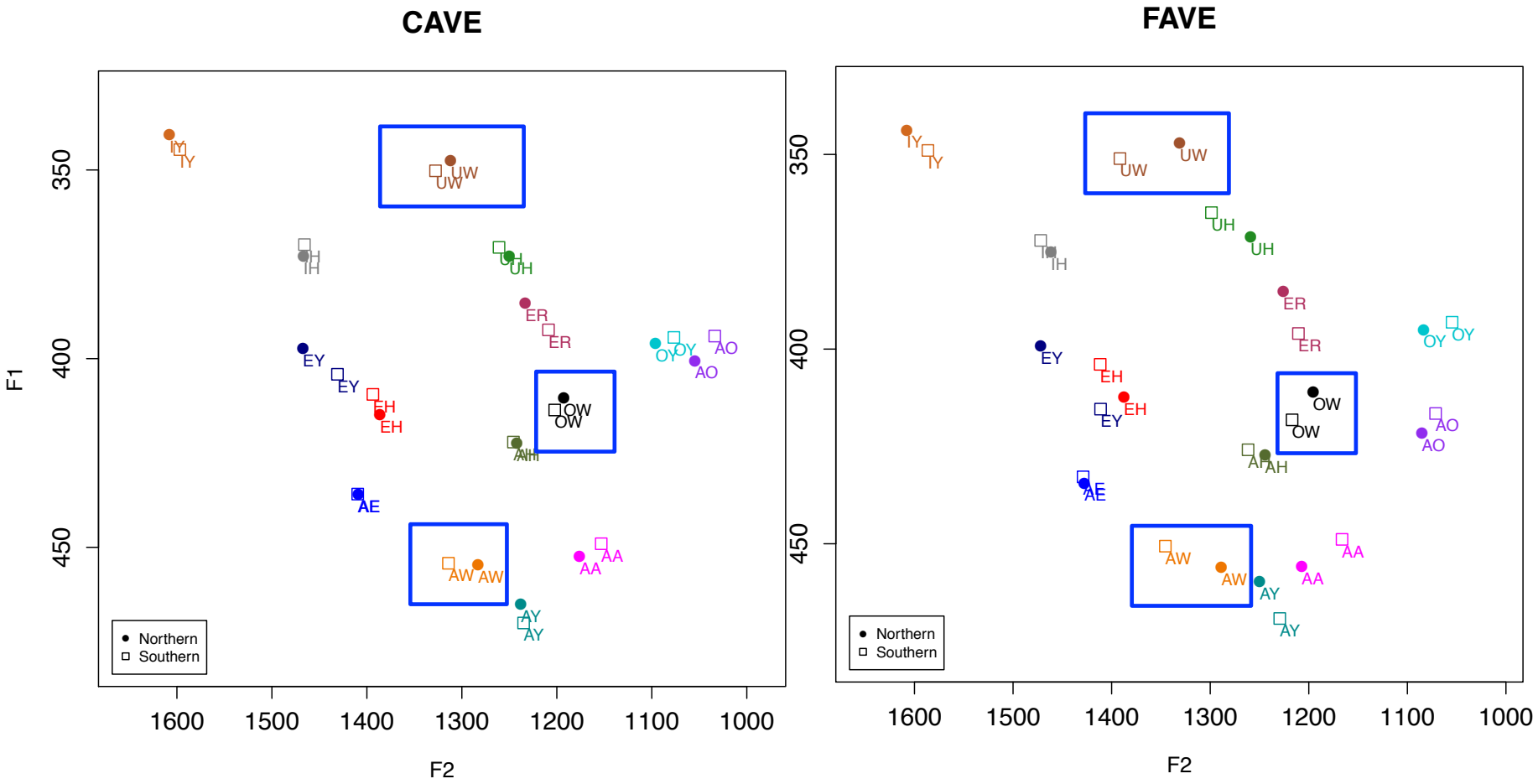
Tense/lax shifts



Evidence of the Southern Vowel Shift

Both CAVE and FAVE also show Southern fronting of AW, UW, and OW

Fronting: UW, AW, OW



FAVE vs. CAVE comparisons

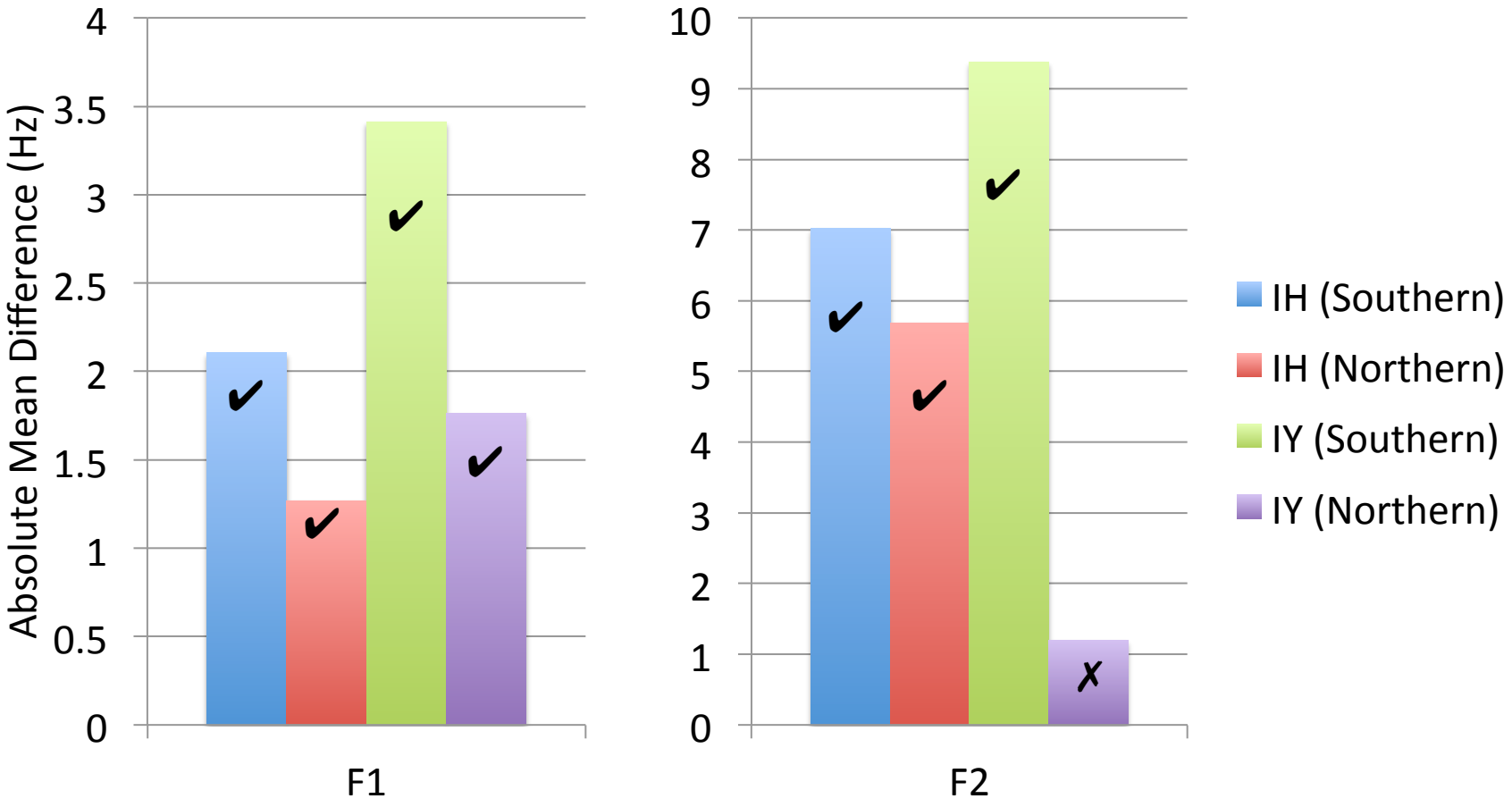
Kendall & Fridland (2012:296) use Euclidean distance between EH and EY as a measure of the tense/lax shift

Repeated Measures ANOVA results:

	FAVE	**CAVE**
EH-EY distance:	North mean=79 Hz South mean=31 Hz Sig. different (p=0.001)	North mean=83 Hz South mean=39 Hz Sig. different (p<0.0001)
IH-IY distance:	North mean=150 Hz South mean=117 Hz Sig. different (p=0.011)	North mean=145 Hz South mean=134 Hz n.s. (p=0.284)

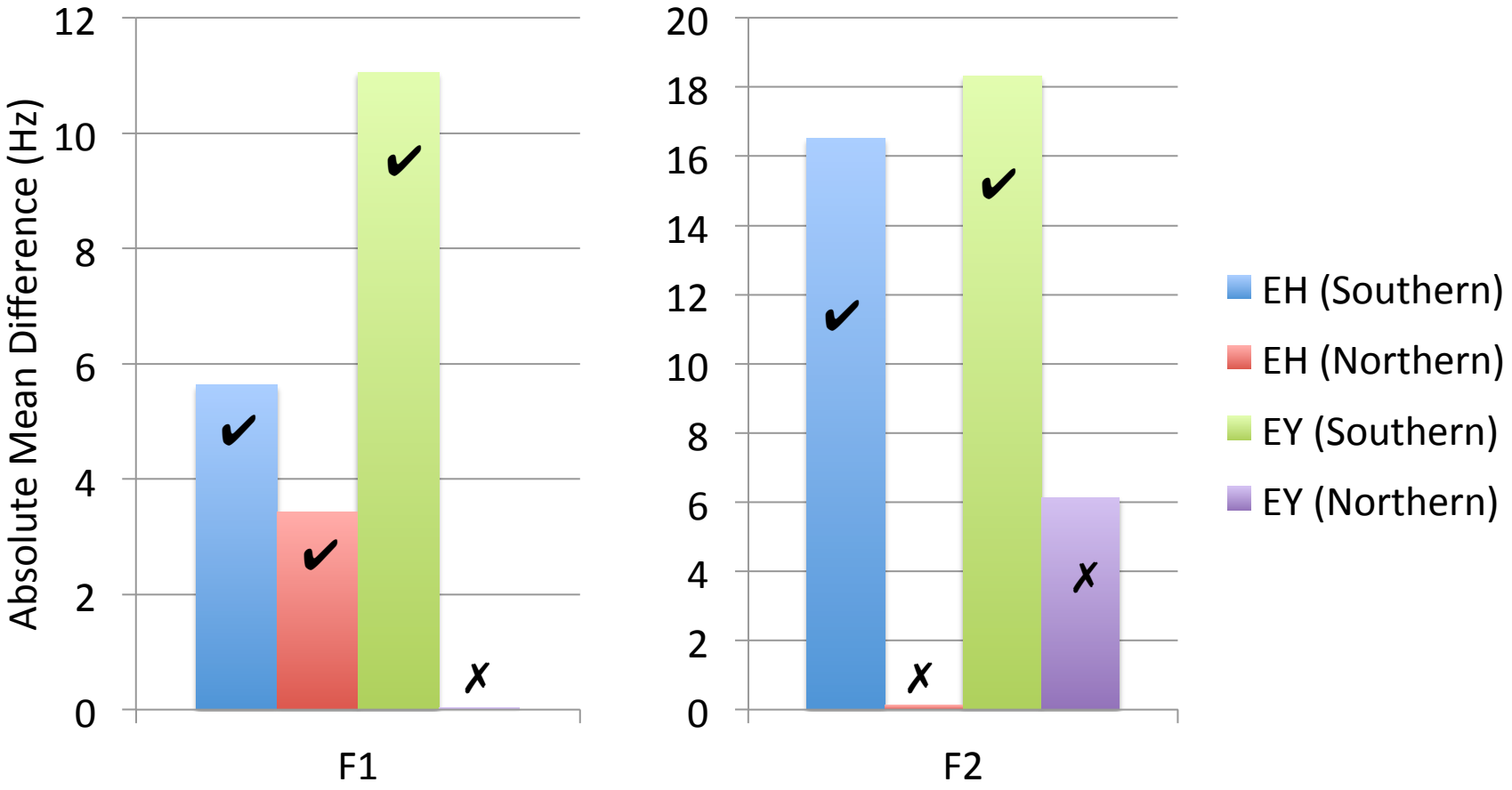
Kendall & Fridland also find EH-EY shift more advanced than IH-IY

Formant Mean Differences between FAVE and CAVE



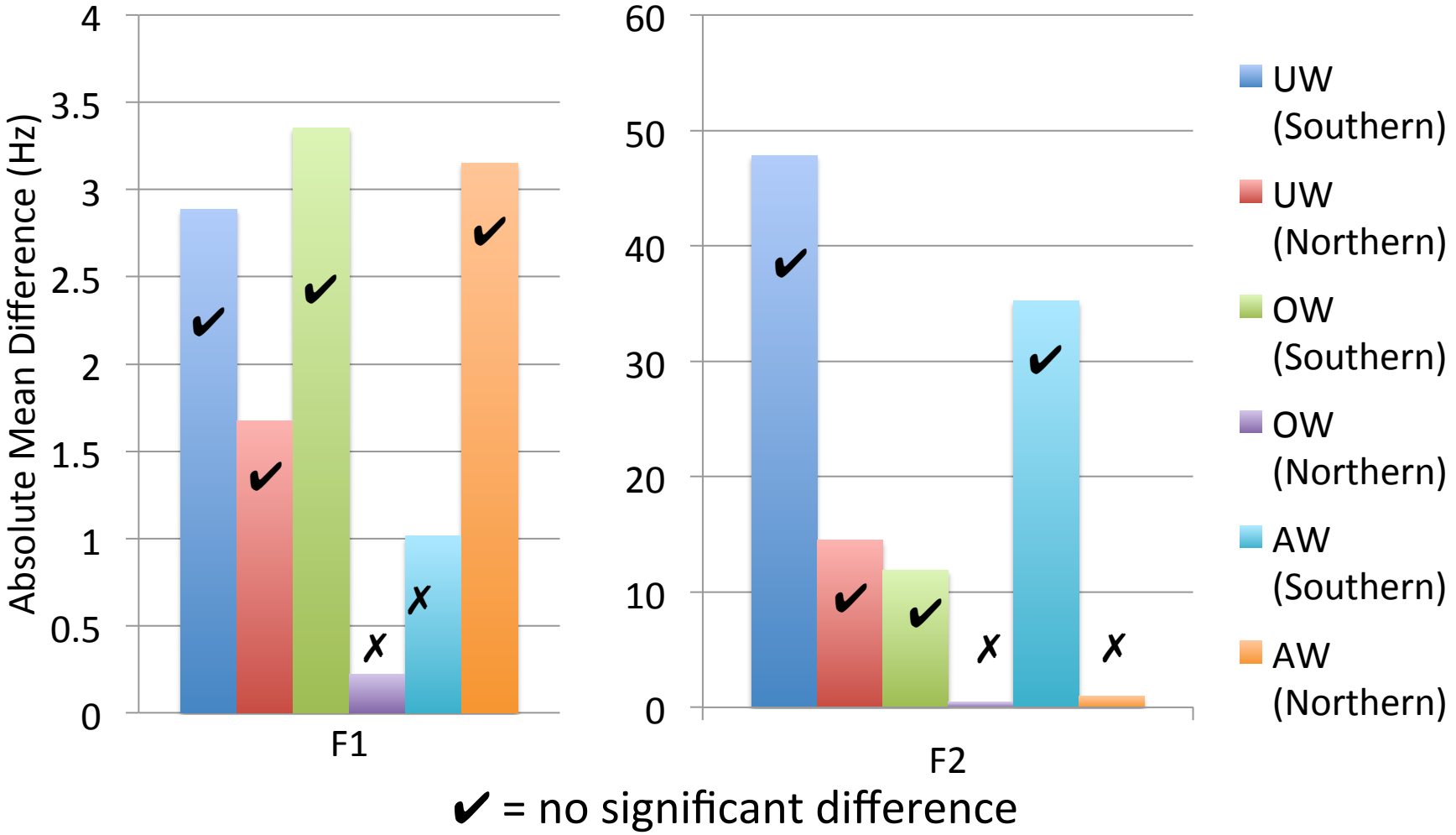
✔ = no significant difference

Formant Mean Differences between FAVE and CAVE



✔ = no significant difference

Formant Mean Differences between FAVE and CAVE



Formant Mean Differences

Differences comparable to findings on inter-analyst differences (Evanini 2009: 92-94)

- Labov et al. (1972:32)
  - F1: 31.5 to 40.5 Hz
  - F2: 38 to 84 Hz
- Deng et al. (2006)
  - F1: 55 Hz
  - F2: 69 Hz
- Hillenbrand et al. (1995:3101)
  - F1: 9.2 Hz
  - F2: 17.6 Hz

Future Work

- Test on other data and dialects

- Tailoring ASR for sociolinguistic applications
  - Get multiple candidate transcriptions and take a weighted average: more resistance to errors
  - Build unified ASR decoding and vowel extraction that directly optimizes for good formant outputs rather than transcription

Conclusions

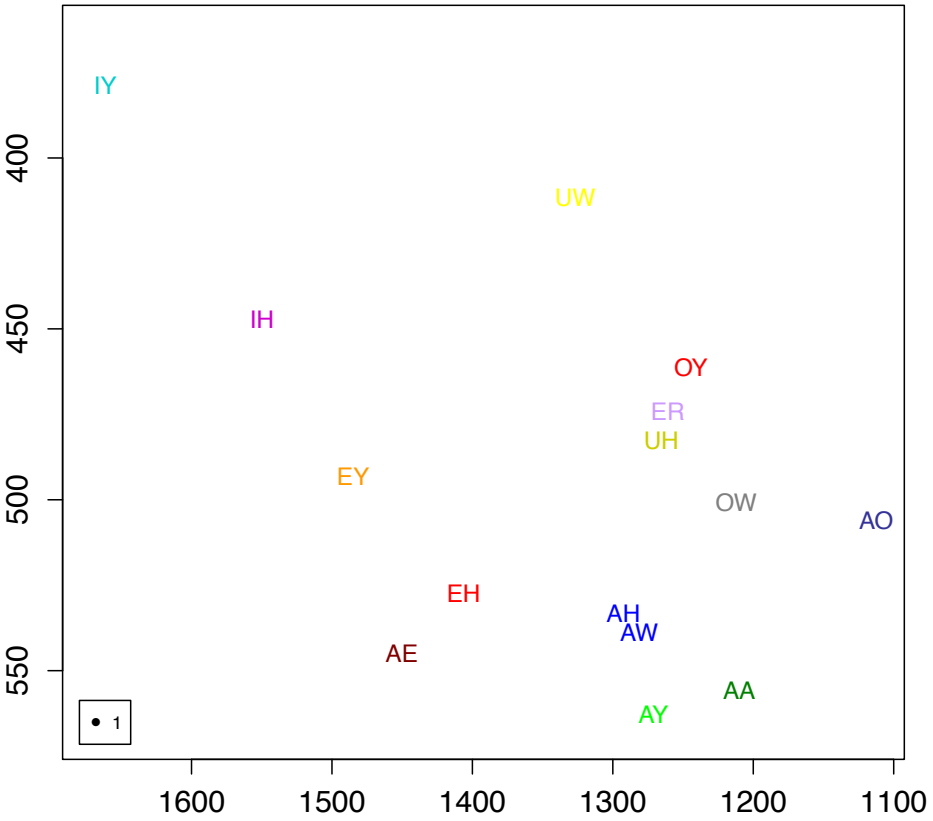
Feasibility Test: Southern Vowel Shift evident with automated transcription and analysis

Suggests that meaningful sociophonetic results can be drawn from a completely automated method

As ASR improves, automated methods will become more reliable for fast analyses of vast amounts of speech

Obama's Vowel Space from CAVE

2014 State of the Union speech (65 min)



Acknowledgments: This project was supported by the Neukom Institute and the Karen Wetterhahn Award at Dartmouth