The SPADE project: large-scale analysis of a spoken language across space and time

Morgan Sonderegger, The SPADE Consortium

LSCP Language Group, Paris
26 Mar, 2019
SPADE
SPeech Across Dialects of English

http://spade.glasgow.ac.uk/
SPADE
SPeech Across Dialects of English

Project manager + data

Software development

http://spade.glasgow.ac.uk/

Graduates + undergrads: many

Postdocs
• **Software** for large-scale automated analysis of speech datasets

• **Data** from ~40 dialects
  • public & private
  • Focus: sociolinguistic data

• **Case studies**: investigate how "English" varies in time and space
Motivation

• Huge amount of annotated speech data exists
  – Corpora, academic labs…

• Across
  – Languages/dialects
  – Speech styles
  – Time

At least orthography + audio
Motivation

• Huge amount of annotated speech data exists
  – Corpora, academic labs, fieldwork…

• Across
  – Languages/dialects
  – Speech styles
  – Time

• + ever-better (semi)-automatic speech measurement tools
Motivation

• Great potential for speech analysis for different purposes
  – Bigger haystacks, same-sized needle…
  – ... need a bigger magnet

• Requires software for unified corpus analysis
  – Integrating speech datasets
  – Querying across them

• SPADE focus: sociolinguistic, phonetic datasets
Barriers

• Speech datasets:
  – Large
  – Complex
  – Diverse formats

• Access to many speech datasets
  – Costly or ethically restricted

• Result: requires lots of specialized code, €€, effort, computational power
Software goals

• Scalable & fast
• Require minimal technical skill from user
• Abstraction away from dataset format
• Querying dataset without access to raw data

⇒ Easier large-scale studies using speech corpora
• To motivate structure of software:
  – think about steps researcher goes through to do a (speech) corpus study
  – Running example: vowel formants

• Setting: sociolinguistic study, or laboratory phonology, phonetics, etc.
Speaker M01
Gender: M
Age: 35

Source: http://drammock.github.io/phonR/
Gender
Age
adjacent segments
formants
Data file (CSV)

How?

(R, Goldvarb...)

Vowel duration
Age
adjacent segments
formants
Gender

1. Process raw data
2. Make measures
3. Find relevant tokens
4. End up with usable spreadsheet
### Raw data

<table>
<thead>
<tr>
<th>SpkrID</th>
<th>Age</th>
<th>Gender</th>
<th>F1</th>
<th>F2</th>
<th>V_dur</th>
<th>C_left</th>
<th>C_right</th>
<th>Phone</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>35</td>
<td>M</td>
<td>340</td>
<td>1410</td>
<td>0.11 s</td>
<td>[g]</td>
<td>[s]</td>
<td>UW</td>
<td>goose</td>
</tr>
<tr>
<td>F01</td>
<td>22</td>
<td>F</td>
<td>480</td>
<td>1050</td>
<td>0.15 s</td>
<td>[r]</td>
<td>[r]</td>
<td>UW</td>
<td>truer</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Data file (CSV)

- **SpkrlD**: Speaker ID
- **Age**: Age of the speaker
- **Gender**: Gender of the speaker
- **F1**: Fundamental frequency of the first formant
- **F2**: Fundamental frequency of the second formant
- **V_dur**: Duration of the vowel
- **C_left**: Consonant on the left side
- **C_right**: Consonant on the right side
- **Phone**: Phone symbol for the word
- **Word**: Word represented by the phone symbols
<table>
<thead>
<tr>
<th>SpkrlD</th>
<th>Age</th>
<th>Gender</th>
<th>F1</th>
<th>F2</th>
<th>V_dur</th>
<th>C_left</th>
<th>C_right</th>
<th>Phone</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>35</td>
<td>M</td>
<td>340</td>
<td>1410</td>
<td>0.11 s</td>
<td>[g]</td>
<td>[s]</td>
<td>UW</td>
<td>goose</td>
</tr>
<tr>
<td>F01</td>
<td>22</td>
<td>F</td>
<td>480</td>
<td>1050</td>
<td>0.15 s</td>
<td>[r]</td>
<td>[r]</td>
<td>UW</td>
<td>truer</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Social factors:**
- Age
- Gender

**Acoustic measures:**
- F1
- F2
- Vowel duration
- Consistent left (C_left)
- Consistent right (C_right)

**Phone info:**
- Phone
- Word
<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>F1</th>
<th>F2</th>
<th>V_dur</th>
<th>C_left</th>
<th>C_right</th>
<th>Phone</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>35</td>
<td>M</td>
<td>340</td>
<td>1410</td>
<td>0.11 s</td>
<td>[g]</td>
<td>[s]</td>
<td>UW</td>
<td>goose</td>
</tr>
<tr>
<td>F01</td>
<td>22</td>
<td>F</td>
<td>480</td>
<td>1050</td>
<td>0.15 s</td>
<td>[r]</td>
<td>[r]</td>
<td>UW</td>
<td>truer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The diagram illustrates the data flow from speaker questionnaire to manual entry, through audio and software (e.g., FAVE), to formants, Vowel duration, adjacent segments, and scripts (Praat, python...).
<table>
<thead>
<tr>
<th>Phone</th>
<th>Word</th>
<th>Frequency</th>
<th>Stress</th>
<th>Syll #</th>
<th>Speech rate (sylls/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UW</td>
<td>goose</td>
<td>15</td>
<td>Y</td>
<td>1</td>
<td>0.11 s</td>
</tr>
<tr>
<td>UW</td>
<td>truer</td>
<td>25</td>
<td>Y</td>
<td>1</td>
<td>0.15 s</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Lexical info**

**Prosodic/suprasegmental**
Duplicated effort across researchers

**Lexical info**
- Phones
- Words
- Syllables
- Custom scripts
- Syllable frequency
- Syllable stress
- Syllable # in word
- Utterance duration
- Speech rate (sylls/sec)

**Prosodic/suprasegmental**
- Utterances
- Custom scripts

**Table**

<table>
<thead>
<tr>
<th>Phone</th>
<th>Word</th>
<th>Frequency</th>
<th>Stress</th>
<th>Syll_#</th>
<th>Speech rate (sylls/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UW</td>
<td>goose</td>
<td>15</td>
<td>Y</td>
<td>1</td>
<td>0.11 s</td>
</tr>
<tr>
<td>UW</td>
<td>truer</td>
<td>25</td>
<td>Y</td>
<td>1</td>
<td>0.15 s</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Different definitions of same variables
1. Process raw data

2. Make measures

3. Find relevant tokens

4. End up with usable spreadsheet
e.g. UW (GOOSE) tokens, stressed syllables > 50ms

Same primitives as other steps
Raw data

Analysis

• scripts
• manual entry
• software

Data file (CSV)

Gender
formants
Vowel duration
Age
adjacent segments
Why ‘Integrated’ Speech Corpus ANalysis?

– Practical reasons
  • Technical skill
  • Time/duplication of effort
  • Availability

– Methodological/theoretical reasons
  • Standardized, customizable linguistic measures

– much more difficult with 1+ corpora…
ISCAN: A SYSTEM FOR INTEGRATED PHONETIC ANALYSES ACROSS SPEECH CORPORA

Michael McAuliffe\textsuperscript{a}, Arlie Coles\textsuperscript{a}, Michael Goodale\textsuperscript{a}, Sarah Mihuc\textsuperscript{a}, Michael Wagner\textsuperscript{a}, Jane Stuart-Smith\textsuperscript{b}, Morgan Sonderegger\textsuperscript{a}

\textit{Proc. ICPhS 2019}

- **Implementation**
  - Python API
  - Graphical User Interface

\texttt{https://github.com/MontrealCorpusTools/iscan-spade-server}

\texttt{https://iscan.readthedocs.io/}

**Datasets (speech corpora, lexicons)**

**Database**

**Set of linguistic objects**

**Data file (CSV)**

\textit{import} \quad \textit{querying} \quad \textit{export}

\textit{enrichment}
• (show GUI here)

• Note:
  
  – **Server-client architecture** enables analysis without access to raw data
  
  – **Permissions system** controls who can see/hear tokens
  
  – Can be installed on web server (default) or personal computer
New World: US and Canada
Old World: UK and Ireland

• 43+ datasets, 4 countries, 115 years
• heterogeneous corpus formats
• public and private
SPADE: datasets

To date:

- Acquired: 20
- Measurements generated: 10
- ~10 dialect regions
  - ~500 hours (?)
SPADE: ethics and credit

• For private datasets (data guardians): ethics complex: GDPR + US laws
• Data transfer agreement
  – data use in keeping with original permissions, as far as is possible
• We welcome new datasets!

• “SPADE consortium” author on everything
• Datasets of measures -> data guardians at end of project
Case studies

- **Sibilants**
  - Stuart-Smith et al. *Proc. ICPHS 2019*

- **Vowels**
  - Mielke et al. *Proc. ICPHS 2019*

**Done**

**In progress**

- Vowels 2
- Stops
- Vowels 3

**Planned**

- Vowels (dynamic)
- r, l
/s/-retraction in English

• /s/ → [ʃ]-like sound in /str/
  – string, street

• Sound change, varies by dialect:
  – Ex: London, Philadelphia, NZ English
  – but not others, e.g. RP,Australian English
    (e.g. Baker et al, 2011; Stevens and Harrington, 2016)

• varies by individual speaker within dialect
Research questions

• **Q1**: what is the evidence for /s/-retraction across English?

• **Q2**: Do English dialects show a dichotomous pattern of /s/-retraction
  – or a continuum?

• Received wisdom: there are “retracting”/non-retracting dialects and speakers
Sample for this study: New World

- **Canada**
  - ICECAN Corpus
  - 28: 18m, 10f

- **Northern Cities, e.g. New York, Philadelphia**
  - Santa Barbara Corpus
  - 20: 9m, 11f

- **Columbus, Ohio**
  - Buckeye Corpus
  - 40: 20m, 20f

- **West coast/California**
  - Santa Barbara Corpus
  - 46: 20m, 26f

- **Raleigh, North Carolina**
  - Raleigh Corpus
  - 101: 50m, 51f

235 speakers

www.google.com/maps/
Sample for this study: New World

Canada
ICECAN Corpus
28: 18m, 10f

Northern Cities, e.g. New York, Philadelphia
Santa Barbara Corpus
19: 8m, 11f

West coast/California
Santa Barbara Corpus
43: 20m, 23f

Columbus, Ohio
Buckeye Corpus
40: 20m, 20f

Raleigh, North Carolina
Raleigh Corpus
101: 50m, 51f

235 speakers

reported to show /s/-retraction

www.google.com/maps/
Sample for this study: Old World

Highlands, Islands and North
SCOTS Corpus
54: 22m, 34f

East, e.g. Edinburgh
SCOTS Corpus
22: 11m, 11f

West, e.g. west coast
SCOTS Corpus
38: 19m, 19f

Glasgow
Sounds of the City
70: 35m, 36f

185 speakers
Data

• All instances of stressed, word-initial /s/

• Acoustic measures: peak, spectral Centre of Gravity (CoG)
  – 1-16 kHz
  – Middle 50%

• Data cleaning

• N = 76,440

• Prediction: /s/ > /str/ > /ʃ/
ISCAN usage

import, enrichment

Speech datasets
- Buckeye
- ICE-Canada
- SCOTS

Databases
- Buckeye database
  - she sells seashells
  - sells
  - seashells
  - σ
  - σ
  - σ
  - s
  - i
  - ʃ
  - ɛ
  - l
  - z

words
- utterances
- syllables
- phones

External tools/resources
- Lexicon
- Praat

query, export

- all syll-initial stressed /s/, /ʃ/
- phone: ID, COG, duration, foll_phone
- speech rate, speaker ID, dataset, ...

CSV file
Results

N = 76,440

- /str/ shows substantial variation across dialects
In US dialects, large differences in lowering of /str/ with respect to /s/.
In **Scottish** and **Canadian** dialects, smaller differences between /str/ and /s/ /s/ is lower in frequency overall.
/s/-retraction in individuals within dialects

A retraction pattern shared by many Raleigh and Glasgow speakers.

N = 76,440
/s/-retraction in individuals within dialects

N = 76,440

non-retracting pattern in individual speakers in West US and Northern Cities dialects
Both retracting and non-retracting patterns seen in some individuals in all dialects.
Discussion

• **Q1:** what is evidence for /s/ retraction across English?
  – Some dialects show retraction of /str/
  – Large differences by dialect and by country
  – Impression of “/s/-retraction” depends on which dialects are considered

• **Q2:** continuum vs. dichotomy in /s/-retraction
  – ?

• Scaling up analysis across dialects, with consistent measures, allows identification of new patterns
Study 2: vowel formants

• Influential hypothesis from sociolinguistics: (Labov, 1994)
  – **Intraspeaker variation** in vowel production ~ same axis as **diachronic change** in community

• Intuitively plausible, but unchecked

• (go to [poster..])

Mielke et al. *Proc. ICPhs 2019*
Thanks!

• SPADE Team, especially
  – Jane Stuart-Smith, Michael McAuliffe, James Tanner, Vanna Willerton, Jeff Mielke

• MCQLL lab RAs
  – Michael Goodale, Arlie Coles, Elias Stengel-Eskin

• Funding
  – Digging Into Data, SSHRC, NSERC
Questions