Large Scale Learning of Speaker Variation

Eleanor Chodroff

Co-mentors:
Sanjeev Khudanpur (Electrical and Computer Engineering, WSE)
Colin Wilson (Cognitive Science, KSAS)
Individual talkers vary significantly in the realization of speech. Morgan's pitch: very, very low. Fran’s nasality: very, very high. Kim’s creakiness: very, very high. But all talkers vary in how they realize pitch, nasality, creakiness, and all other sorts of phonetic variables.
Individual talkers vary significantly in the realization of speech

Today’s case study: mean frequency of [s] (~pitch of [s])

Low
~4,000 Hz

German < English
Japanese < English

Male < Female
(not related to physiology)
Male and straight < Male and gay
Living in rural Redding, CA < Living in urban Redding, CA

Et cetera
Individual talkers vary significantly in the realization of speech

Is the way a talker produces [s] indicative of how they produce other related speech sounds, [z] ‘z’, [ʃ] ‘sh’, and [ʒ] ‘zh’ (sibilant fricatives)?

Research hypothesis: There are strong relations of mutual predictability among phonetic variables measured at the individual talker level (such as talker’s mean frequency of [s], [z], etc.)

Null hypothesis: The way a talker produces an [s] is independent of how he/she produces other speech sounds (even those closely related in articulation).
Individual talkers vary significantly in the realization of speech

**Insight from ASR:** Assuming there are relationships among speech sounds helps a lot in automatic methods of speaker adaptation.

**Cognitive scientist:** Why are some relationships stronger than others, and are some more reliable than others? Also, do humans use these relationships in learning about a new talker?

**Help from ASR and machine learning:** Tools and techniques available to process large amounts of speech data to answer such questions.
American English: Mixer 6 Corpus

180 native talkers of American English
~45 minutes of speech per talker
Controlled sentential contexts: same set of sentences read in the same order

Corpus: Brandschain et al. 2010, 2013
Corpus audit: Chodroff et al. 2016
Alignment: Yuan & Liberman 2008
Mixer 6 Sibilants

[s, z, ʃ]: word-initial, word-medial, a few word-final sibilants before vowels

Measured the mid-frequency peak ($F_{\text{M}}$), which is highly related to the mean frequency

Adapted from Shadle et al. 2011, Koenig et al. 2013, Shadle 2016

Excluded tokens ±2.5 standard deviations from talker-specific category mean

55,304 sibilants in $F_{\text{M}}$ analysis

<table>
<thead>
<tr>
<th>Fricative</th>
<th>Range per talker</th>
<th>Median # Tokens</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>110 - 314</td>
<td>223.5</td>
<td>39,431</td>
</tr>
<tr>
<td>[z]</td>
<td>21 - 44</td>
<td>33</td>
<td>6,006</td>
</tr>
<tr>
<td>[ʃ]</td>
<td>30 - 84</td>
<td>54</td>
<td>9,867</td>
</tr>
</tbody>
</table>
Talker variation in mean $\text{Freq}_M$: American English

- **[z]**
  - $\mu = 5735$ Hz
  - Range of talker means: 3573 – 6753 Hz

- **[s]**
  - $\mu = 5656$ Hz
  - Range of talker means: 3713 – 6856 Hz

- **[ʃ]**
  - $\mu = 3181$ Hz
  - Range of talker means: 2178 – 5341 Hz
**Covariation of mean Freq_M: American English**

[s] – [z]
95% CI: [0.95, 0.97]

Females: $r = 0.92^* [0.85, 0.96]$
Males: $r = 0.92^* [0.86, 0.95]$

[s] – [ʃ]
95% CI: [0.60, 0.74]

Females: $r = 0.49^* [0.35, 0.58]$
Males: $r = 0.38^* [0.16, 0.60]$

* = $p < 0.001
Do listeners have perceptual knowledge of covariation among speech sounds?

Two experiments:
• Expose to talker’s [z], test whether the perceptual boundary between [s] and [ʃ] shifts
• Expose to talker’s [v], test whether the perceptual boundary between [s] and [ʃ] shifts

Predictions from acoustics:
• The mean frequency (COG) of [s] and [z] are highly correlated, so if a talker has a high COG for [z], then they should also have a high COG for [s].
• The mean frequency (COG) of [v] and [s] are not correlated, so even if a talker has a high peak for [v], no strong inferences can be made regarding [s].
Perceptual generalization

**Exposure**
HIGH OR LOW COG*

**Test**
[s] or [/ʃ]

trial: 1 exposure stimulus (2 reps) + 1 test block: 20 trials
Listeners less likely to choose [s] after exposure to high COG [z] than low COG [z]

Listeners not less likely to choose [s] after exposure to high COG [v] than low COG [v]
Related research

Extended analysis of **structured variation** in the phonetic realization of speech sounds to:
  • Other phonetic variables associated with fricatives and stop consonants
  • American English, Czech, and other languages
  • Child speech patterns

Examined perceptual learning of novel talkers in:
  • Fricatives
  • Stop consonants
Research Dissemination

Why?
Increased public awareness (general knowledge, appreciation, funding)
Improve methodologies in the field

Two projects:
1. High school outreach
2. Corpus phonetics tutorial
**Dissemination Project 1: High School Outreach**

**Broad Goals**

Introduce high school students to high level concepts in phonetics, automatic speech recognition, and automatic speaker recognition

Increase awareness of and inspire interest in these topics

Make smarter consumers (of both technology and language) and potential researchers

**Specific Goals**

Explain how voice recognition works
Understand types of features that go into a voiceprint

**Audience**

High schoolers
Dissemination Project 1: High School Outreach

“Hey Siri!” voice recognition
Dissemination Project 1: High School Outreach
Dissemination Project 1: High School Outreach

Differences in pitch
Student feedback:
“Linguistics was an area of study I’ve never seriously considered but after the activities, I’ve changed my stance.”

“I was surprised on the information I learned on the voice recognition. There’s a lot of work put in to decode someone’s voice.”
Dissemination Project 2: Corpus Phonetics Tutorial

Broad Goals

Facilitate data processing in both scale and speed for better and more efficient research

Advance the state-of-the-art in speech science and technology

Specific Goals

Provide accessible (online) resource on how to use ASR-based tools for doing large scale corpus phonetics

Expand community of researchers using ASR-based tools in research

Audience

Speech scientists and engineers
Dissemination Project 2: Corpus Phonetics Tutorial

https://eleanorchodroff.com/tutorial/intro.html

Tutorials:
Penn Forced Aligner
AutoVOT
Kaldi

Corpus Phonetics Tutorial
Eleanor Chodroff

Kaldi

What is it? Kaldi is a state-of-the-art automatic speech recognition toolkit, containing almost any algorithm related to ASR. It also contains recipes for training your own acoustic models on commonly used speech corpora such as the Wall Street Journal Corpus, TIMIT, and more. These recipes can also serve as a template for training acoustic models on your own speech data.

What are acoustic models? Acoustic models are the statistical representations of a phoneme's acoustic information. A phoneme here represents a member of the set of speech sounds in a language. N.B., this use of the term 'phoneme' only loosely corresponds to the linguistic use of the term 'phone'.

The acoustic models are created by training the models on acoustic features from labeled data, such as the Wall Street Journal Corpus, TIMIT, or any other transcribed speech corpus. There are many different ways these can be trained, and the tutorial will try to cover some of the more standard methods. Acoustic models are necessary not only for automatic speech recognition, but also for forced alignment. If you open up the Penn Forced Aligner, you can find those acoustic models in the 'models' directory. They were trained on the SCOTUS corpus using the Hidden Markov Model Toolkit (HTK).

Kaldi provides tremendous flexibility and power in training your own acoustic models and forced alignment system. The following tutorial covers a general recipe for training on your own data. This part of the tutorial assumes more familiarity with the terminal; you will also be much better off if you can program basic text manipulations.

Please also refer to the Kaldi website for thorough documentation: http://www.kaldi-asr.org/.
Dissemination Project 2: Corpus Phonetics Tutorial

Tutorial on the Kaldi Automatic Speech Recognition Toolkit

In the past 9 months:
> 5 seconds: 1,214 users
> 1 minute: 916 users
> 5 minutes: 508 users
> 10 minutes: 294 users
Special thanks to:

Science of Learning Institute -- Johns Hopkins
The Distinguished Science of Learning Fellowship Program

Science of Learning Institute
Barbara Landau
Kelly Fisher
Kristin Gagnier

Education Outreach
Margaret Hart

Fellow Fellows
Kara Blacker
Emily Coderre Plante
Aaron White
Bob Wiley

Mentors
Sanjeev Khudanpur
Colin Wilson

Collaborators
Jack Godfrey
Yenda Trmal

Research Assistants
Alessandra Golden
Elsheba Abraham
Gigi Edwards
Chloe Haviland
Spandana Mandaloju
Monica Sohn
Ben Wang

Thank you!