1pSC: Exploring the interface between linguistic processing and talker recognition

Rachel M. Theodore
University of Connecticut

Tyler K. Perrachione
Boston University
1pSC: Exploring the interface between linguistic processing and talker recognition

@UConnNBL
#1pSC
#ASA177
Listener sensitivity to structured phonetic variation

Rachel M. Theodore, Ph.D.
Department of Speech, Language, and Hearing Sciences
Connecticut Institute for the Brain and Cognitive Sciences
University of Connecticut
Like many on this panel, work in our lab examines how listeners map the acoustic speech signal to meaning, focusing on the early stages of this process – where listeners recognize the speech sounds that form building blocks for higher levels of comprehension, such as words.
To perceive speech sounds, listeners have to solve a massive computational problem; much to the dismay of our students in speech science courses, there’s no one-to-one relationship between speech acoustics and individual consonants or vowels.
As an illustration, consider just one of the acoustic-phonetic properties present in the speech signal, voice-onset-time, which is temporal property of stop consonants. On any given day, a listener will hear a wide range of VOTs, even for the same speech sound.
But the input is structured such that some VOTs for a given stop are more frequent than others —
and distinct contextual distributions are formed. For example, VOTs for voiceless stops are longer than those for voiced stops, and VOTs for a slow speaking rate are longer than those at a faster speaking rate. So even though there's a wide range of acoustic variation in the input, there are systematic patterns – or structure – to this variability.
Speech sounds have a graded internal structure

Do you hear /b/ or /p/?

How good is this as /p/?

Identification as /p/ vs. VOT (ms)

Goodness as /p/ vs. VOT (ms)

Volaitis & Miller (1992); Theodore et al. (2015); Myers & Theodore (2017)

Research from our team and others has shown that this structure is reflected in the representations for speech sounds; while forced-choice categorization tasks show that many different signals are identified as the same phoneme, consistent with categorical perception of speech — other tasks including explicit goodness judgments show that listeners code the degree to which the signal is a good vs. poor exemplar; not all category members are equally good members.

Phonetic category structure is also observed in fMRI paradigms, which show that activation in temporal regions is graded with respect to prototypicality in the input.
- Lack of invariance between speech acoustics and linguistic representations
- Speech sound categories have a graded internal structure that reflects input frequency

Because speech sound categories have a graded internal structure, perception can be viewed as a process in which listeners make inferences about the sounds they hear in the face of some degree of uncertainty.
• Lack of invariance between speech acoustics and linguistic representations
• Speech sound categories have a graded internal structure that reflects input frequency
• Perception is probabilistic

In other words, perception is probabilistic. Given a specific VOT, for example, /p/ might be activated to strong degree, along with relatively weaker activation for other stop consonants.
• Lack of invariance between speech acoustics and linguistic representations
• Speech sound categories have a graded internal structure that reflects input frequency
• Perception is probabilistic
• Perception is dynamic

We also know that speech perception is dynamic. We don’t wait until we’re 100% sure about the sounds we hear before we start making guesses about a speaker’s words; and the probabilistic nature of speech sound activation has cascading effects along the way.
• Lack of invariance between speech acoustics and linguistic representations

• Speech sound categories have a graded internal structure that reflects input frequency

• Perception is probabilistic

• Perception is dynamic

Another example of the dynamic nature of speech perception is that the specific mappings from acoustics to speech sounds change as a function of context…
• Lack of invariance between speech acoustics and linguistic representations

• Speech sound categories have a graded internal structure that reflects input frequency

• Perception is probabilistic

• Perception is dynamic

...such as speaking rate, dialect, or even who in particular is speaking. Listeners need to track variation, and adapt to it; which in and of itself is a skill.
I'm going to tell you about some of our recent work that informs (1) how listeners modify the mapping to speech sounds to accommodate structure in the input, and (2) how sensitivity to phonetic variation can also be used to facilitate nonlinguistic processing, including talker recognition.
Talkers provide structure to phonetic variation

- Talkers differ in their characteristic formant patterns specifying vowels

The specific type of structure that we've considered is talker-specific phonetic variation. This plot shows formant patterns produced for English vowels across a large number of speakers. Each IPA symbol represents a specific talker's productions, and while variation across speakers is vast, individual talkers are relatively more consistent in the cues they produce for vowels. So talker provides structure to the overall variation.

Hillenbrand et al., 1995
Talkers provide structure to phonetic variation

- Talkers differ in their characteristic spectral centers for fricatives

This holds also for spectral properties of speech, such as center frequency of fricatives. This plot shows histograms of the center frequency for S and SH for two subjects. Though each talker clearly differentiates S from SH, they do it in different acoustic space; and if listeners were to apply the same frequency boundary to both of these speakers, they’d be catastrophically wrong for one of the talkers; one could even say they would be SIT out of luck…
Talkers provide structure to phonetic variation

- Talkers differ in their characteristic VOTs for stop consonants

Allen et al. (2003); Theodore et al. (2009); Chodroff & Wilson (2017)

Talker differences have also been observed for temporal properties of speech, including VOT; some talkers have characteristically longer VOTs than other talkers.
An overarching hypothesis in our lab is that the structure present in speech input can be used to solve the mapping problem in speech perception, that is, listeners can derive talker-specific mappings to reduce the overall computational complexity in mapping speech to meaning. Initial tests of this hypothesis asked whether listeners could keep track of a talker’s prototype; that is, can they learn that one talker produces short VOTs and another produces long VOTs?
To answer this question, we exposed listeners to two talkers, and manipulated how each talker produced the voiceless stop /p/. One talker produced /p/ with shorter VOTs compared to the other talker.
Talker-specific influences on category structure

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Talker 1</strong></td>
<td><strong>Talker 1</strong></td>
</tr>
<tr>
<td>Short-VOT <em>pain</em></td>
<td>Short-VOT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Talker 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-VOT <em>pain</em></td>
<td>Short-VOT</td>
</tr>
</tbody>
</table>

Theodore & Miller, 2010

Following this exposure, we presented the listeners with two VOT variants in each talker's voice, and simply asked them to choose which one was most representative of the talker.
And what we found is that listeners picked the version that was consistent with their exposure during training. This effect held even when they were tested on novel words, and when they were tested on novel stop consonants; learning how a talker produced /p/ informed listeners as to how that talker produced /k/.

These results showed that listeners can track a talker's prototype. In subsequent studies, we examined whether this also promotes changes to internal category structure.
To do so, we presented two groups of listeners with the same talker’s voice, we fictitiously referred to her as Joanne. And one group of people heard Joanne produce /k/ with short VOTs, but the other group heard Joanne produce /k/ with long VOTs.
At test, we presented both groups with a wide range of VOTs and asked them to rate each one as goodness as a member of the /k/ category.
And here are the results, with goodness as /k/ on the ordinate as a function of VOT. What we observed is that which VOTs were given the highest goodness ratings differed between the two exposure groups, in line with their previous experience with Joanne’s voice; this difference was observed across a range of VOTs, not just the specific ones that had been presented during training. This pattern also held when listeners were tested on novel words.
Listeners adjust phonetic category structure in line with a talker’s characteristic productions.

Listeners track distributional information for a given talker, which can be modeled using ideal observer frameworks.

\[
p(k|\text{VOT}) = \frac{p(\text{VOT}|k)}{p(\text{VOT}|k) + p(\text{VOT}|g)}
\]

Clayards et al. (2008); Kleinschmidt & Jaeger (2015); Theodore & Monto (2019)

Listeners’ ability to modify perception in line with distributional information in the input follows predictions made by ideal observer computational frameworks.

This modified version of Bayes’ theorem predicts that the listener’s response, in terms of the probability of hearing /k/ given a specific VOT, for example, is determined by the probability that this specific VOT is produced for /k/ and the probability that this specific VOT is produced for a different category, such as /g/.
Meghan Clayards and colleagues demonstrated that listeners’ responses follow predictions of ideal observer models. Two groups of listeners were presented with different VOT input distributions — one group heard VOTs cueing voiced and voiceless stops that had narrow variances, while the other group heard VOT distributions that had wider variances. The ideal observer framework predicts that this specific difference in input distributions will lead to a difference in listeners’ identification slopes; those who hear narrow input will show a steeper — or more categorical — identification function compared to those who hear the wide input. And this is exactly the pattern that they observed among humans listeners.
This same equation can of course be used to derive predictions for input distributions that differ in their modal values, like those shown here, where the input specifying /g/ and /k/ is shifted in acoustic-phonetic space. The ideal observer framework predicts that for these distributions, listeners will show a difference in the location of the perceptual boundary, which will be located at a shorter VOT value for the short compared to the long input distributions. And these predictions too have been shown to match human behavior.
However, the time-course of adaptation is currently underspecified in distributional learning accounts of speech adaptation. And by this I mean — do listener’s responses reflect cumulative experience with a talker’s voice, or rather do listeners rely more heavily on recent experience? The studies I just told you about have used a between-subjects design, which doesn’t allow optimal examination of this question.
We used a within-subjects design to get at this more directly. Two groups of listeners completed two blocks of phonetic categorization task for single words that differed in word-initial VOTs; *goal, coal, gain cane*. In one block, VOTs formed narrow distributions; VOT. In the other block, VOTs formed wide distributions. Each block contained 236 tokens, and we manipulated the order in which listeners completed the two blocks: NW vs. WN. In terms of the type of within-talker variation this input might represent, you can think of narrow input as clear speech and the wide input reflecting a more casual speech register.
Predictions for the local vs. global statistics hypotheses were derived using the modified version of Bayes’ theorem from Clayards and colleagues. For the local statistics predictions, identification functions were determined based on the input distributions in each block, which predicts that the slope of the ID function will differ between the two order groups in each block.
Predictions for the global statistics hypothesis are disassociated in block two, which were determined by calculating the predicted ID function for the narrow and wide input blocks combined; leads to the prediction that the two groups will show equivalent ID slopes in block two, given that cumulative experience is equivalent between the two order groups.
Time-course of distributional learning for speech

- The two order groups differed in block 1 but not in block 2, consistent with the global statistics hypothesis.
- Only the NW order group showed a change in slope across blocks.

Here are the results; the slope of the ID function is shown on the ordinate for both order groups in each block. Higher values indicate steeper slopes.

Three things to note: (1) In block 1, the two groups differ, those who heard narrow input in block 1 have a steeper slope compared to those who heard wide input; (2) in block 2, no difference between the two order groups, consistent with the global statistics hypothesis, and (3), the convergence in block 2 reflects within-subjects change of only the NW order group. There’s an asymmetry in slope movement that was not predicted by the simplified version of Bayes’ theorem…
Trial-level analyses confirmed slope movement for only the NW order group. I’ve shown here the slope (beta estimate for VOT) for each order group at three trials; trial 200 is near the end of the first block (236 trials in block 1), trial 450 is near the end of the second block (472 trials total), and trial 325 is about 1/3 of the way through the second block.

So, what can explain this asymmetry? The version of Bayes’ theorem we used to derive predictions did not take into account prior knowledge; that is, what expectations listeners have when they come in to the study; it also can’t generate predictions over time, at least not easily…
To try and explain the asymmetry in learning, we turned to the Bayesian belief-updating model from Kleinschmidt & Jaeger. In their model, the user specifies prior distributions for two categories, along with a confidence parameter. The input is trial-level observations of a perceptual parameter, such as VOT, and a response category, such as /k/.

With this input, the learning algorithm updates the category-specific distributions on each trial, integrating the observed VOT and response with the prior distribution, weighted by confidence. The output is the posterior distribution on each trial, reflecting the likelihood of the prior distribution (formed by global experience with English) given the observed evidence (from the specific talker). The algorithm is iterative at each trial, and thus can simulate how beliefs in priors change across trials.
Bayesian belief-updating model of adaptation

I'm going to talk you through our simulation procedure, and want to give a shout-out to some great open science tools...

github.com/kleinschmidt
“beliefupdatr”

github.com/nick-monto
“slopeExtractR”

osf.io/8krg3/
Computational simulations of distributional learning

- Nine simulations were performed
  - Three prior specifications
  - Three confidence specifications

We performed 9 simulations, formed by crossing three prior specifications and three confidence levels. For all three prior specifications, the SD of /g/ and /k/ was set to match that of a typical talker, drawing on a great open data corpus provided by Eleanor Chodroff. What we changed across the prior specifications was the mean of /g/ and /k/, which was set to match our actual input distributions in addition to means that were shifted up and down 10 ms.

Happy to answer questions about our prior specifications — Confidence is specified by a pseudo-count of prior observations, so a low number indicates less confidence and a high number indicates more confidence.
Nine simulations were performed

Three prior specifications

Three confidence specifications

Input was 40 randomizations of trial-level VOT and response category for each order group, simulating 80 listeners

The input for each simulation was 80 randomizations of trial-level VOT and response, 40 simulating exposure for the NW group and 40 simulating exposure for the WN group.

[Response patterns for the 80 simulated listeners matched the intended category for all VOTs except the four most intermediate. For the two most intermediate VOTs (60 ms, 69 ms), a random 50% of the responses were set to match the opposite category. For the next two intermediate VOTs (51 ms, 83 ms), a random 25% of the responses were set to match the opposite category. This procedure added a degree of response noise to simulate the imperfect categorization of midpoint stimuli observed in the behavioral test.]
Nine simulations were performed

Three prior specifications

Three confidence specifications

Input was 40 randomizations of trial-level VOT and response category for each order group, simulating 80 listeners

For each randomization, the slope of the predicted identification function was extracted at trials 200, 325, and 450.
Computational simulations of distributional learning

- Greater slope movement is observed for the NW compared to the WN order, consistent with the behavioral results.

Here are the simulation results for prior specification #1 for the three confidence levels. Higher slope parameters mean steeper identification functions; and these values represent the mean and SE of the slope for the 40 simulated listeners in each order group.

A few things to note: As confidence in priors goes up, slopes become steeper overall; this makes sense as increased confidence in the priors leads to diminished change to the priors. Critically, with respect to the ORDER in which the input distributions were simulated, an asymmetry in learning is observed for all three confidence levels, just as we found in the human data.
Same pattern held for the other two prior specifications…
In this framework, the observed asymmetry in learning can be explained by considering how the posterior distributions change over time as a consequence of the relationship between the input and the priors. This plot shows the posterior distributions over time as calculated by the belief-updating model. For the NW group, there is a steady change over time, as the input moves from narrow to wide distributions. But for the WN group, all of the movement in posteriors occurs very early in order to adapt to the initial wide input. Later occurring “narrow” input does not generate prediction error because it is contained within the wide input, resulting in limited change to beliefs over time.
Predicted identification functions over time

- For the NW order, slope becomes shallower over time
- For the WN order, shallow slope occurs early and then remains constant over time

I’m happy to field questions on this if there’s interest; I imagine we’ll hear more about this model later in this session…
Outline

1. Listeners dynamically modify the mapping to speech sounds to accommodate structured phonetic input

2. Sensitivity to structured phonetic variation facilitates talker recognition
Phonetic variation and voice recognition

- Linguistic processing and talker processing are tightly linked
  
  Experience with a talker’s voice changes the mapping to speech sounds

  Voice recognition is heightened in a familiar compared to an unfamiliar language

- What aspects of the language architecture interface with talker processing?

Nygaard & Pisoni (1998); Theodore & Monto (2019); Goggin et al. (1991)

The links between linguistic processing and talker processing are apparently bidirectional — just as experience with a talker's voice can influence the mapping to speech sounds, experience with a language can facilitate talker processing.

An open question, however, is what aspects of the language architecture facilitate this nonlinguistic ability.
<table>
<thead>
<tr>
<th>Phonetic variation and voice recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Listeners who have regular exposure to a nonnative language show better talker identification in that nonnative language</td>
</tr>
<tr>
<td>• Listeners can identify voices from sine-wave speech analogs</td>
</tr>
<tr>
<td>• Listeners can learn to use VOT as a cue to talker identity for “voices” that are otherwise identical</td>
</tr>
<tr>
<td>• Brain regions associated with voice processing are sensitive to talker-specific phonetic variation</td>
</tr>
</tbody>
</table>

Orena et al. (2015); Remez et al. (1997); Francis & Driscoll (2006); Knöshe et al. (2002)

There is some evidence that talker differences in phonetic properties of speech can be used to identify talkers. For example, talker identification in a nonnative language is facilitated by exposure to that specific nonnative language. Talker identification is possible for sine-wave speech analogs, a signal manipulation that removes traditional indexical properties. Listeners can learn to use VOT as a cue to talker identity for voices that are otherwise identical, AND, brain regions associated with voice processing are sensitive to talker-specific phonetic variation.
Phonetic variation and voice recognition

- Voice networks show sensitivity to talker-specific phonetic variation when listeners perform a phonetic identification task
- Can listeners exploit structured phonetic variation for talker identification?

In our earlier study, we found that right temporoparietal regions implicated in voice processing showed sensitivity to the consistency between VOT variant and talker exposure during a phonetic identification task, raising the possibility that talker-specific phonetic variation is informative for voice processing.
To get at this directly, we ran a talker identification experiment where we manipulated phonetic structure across talkers. Two groups of listeners completed an associative learning talker identification task where listeners learned to pair a voice with a cartoon face with feedback on every trial. For one group, the structured group, there was a talker-specific structure to VOTs of the *coal* and *cane* words. The other group heard the same VOTs, but no talker-specific structure was present. Note that three talkers had perceptually distinct voices; listeners could completely ignore their VOT productions and still excel at the talker identification task.
Following the training, listeners were then tested on trained and novel items in the same identification task, but feedback was removed. The novel items consisted of words at a novel place of articulation. The test stimuli were the same for both the structured and unstructured exposure groups.
Results

- With only brief exposure to talkers’ voices, no evidence that talker identification is facilitated by structured phonetic variation.

And here are the results in terms of proportion correct talker identification during training, at left, and test, at right. Feast your eyes on this pile of null effects... In the short term, structured variation doesn't seem to offer much for recognizing voices. This is perhaps not surprising given that there were many other cues available for discriminating the voices.

So we thought to ourselves that maybe these patterns reflect the very limited training period; perhaps listeners need additional time in order to learn the talker-specific phonetic structure. So we ran a second experiment with a different group of listeners and this time tripled the training; instead of one block of training they did three blocks.
Results

- Given extended exposure, structured phonetic variation facilitates talker identification
- Facilitation generalizes to novel place of articulation

And here are the training results for the three training sessions. Those who heard a structured relationship between phonetic variation and talkers’ voice get better at voice recognition over time compared to those who have unstructured exposure.
Results

- Given extended exposure, structured phonetic variation facilitates talker identification
- Facilitation generalizes to novel place of articulation

We observed a similar benefit for structured compared to unstructured variation at test, for both trained and novel items. These data suggest that talker-specific pronunciation patterns can be exploited for voice recognition.
Future directions

- To what degree does sensitivity to talker-specific phonetic variation pervade the cognitive and linguistic architectures?
- How can we account for individual differences in perceptual learning for speech, and their contributions to language impairment?
  - Deficits in perception vs. deficits in learning
  - Causal path from perceptual learning to broad behavioral phenotype

I'll end by pointing towards some of our ongoing work; we are currently examining whether the role of talker as context for linguistic processing pervades to higher levels of processing including semantic integration in memory;

AND, we're trying to better understand factors that contribute to individual differences in perceptual learning with an eye towards implications for language impairment; you can get a preview of this work if you check out my students' posters later this week.
I extend gratitude to my collaborators, and a great team of research assistants in the SLaP Lab —
Acknowledgements

NIH R21 DC016141  Theodore
NSF BCS 1827591  Theodore
NSF BCS 1827598  Shattuck-Hufnagel
NSF DGE 1144399  Magnuson
Stetson Fellowship  Monto

@UConnNBL  slaplab.uconn.edu  @UConnNBL  osf.io/8krg3/

And to the fundings sources for this research. THANK YOU.