Current Issues and Methods in Speaker Adaptation

The Ohio State University
April 6-7, 2013

CIMSA
2013

Welcome to CIMSA

This workshop will be held in the Ohio Union on The Ohio State University campus. Talks and discussion sessions will take place in the Round Room, and poster sessions will take place in the Barbie Tootle Room.

This workshop has three primary aims.

1. **To integrate theoretically and methodologically diverse research** on how listeners with varying experiences and abilities (e.g., mono- and bilinguals across the lifespan) accommodate linguistic variation from a range of sources (e.g., variation due to anatomical, idiolectal or dialectal, or social factors).
2. **To examine the power and limitations of existing methods** for studying speaker adaptation by discussing the theoretical underpinnings and architectural assumptions of these methods.
3. **To inspire inter-field collaborations** between/among attendees that will transform the conversations that take place at the workshop into sustained lines of new research.

The CIMSA Organizing Committee

Mary Beckman (co-chair)
Bridget Smith (co-chair)
Abby Walker (co-chair)
Kodi Weatherholtz (co-chair)

Kathryn Campbell-Kibler
Cynthia Clopper
Micha Elsner
Kiwa Ito
Dahee Kim
Andrew Plummer
William Schuler

Acknowledgements

We are grateful to the following bodies for financial and organizational support.

- The Ohio State University Department of Linguistics
- The Buckeye Language Network
- The OSU Student Linguistics Association

Saturday April 6th

8:20am Registration, Breakfast, and Poster setup

8:50am Welcome

Session: Representations and Cognitive Architecture

Chair: Micha Elsner **Round Room**

9:00am Perceptual and cognitive constraints on vowel-space adaptation
Delphine Dahan

9:35am Factors that affect phonetic adaptation: Exemplar filters and sound change
Keith Johnson

10:10am Acquiring and adapting phonetic categories in a computational model of speech perception
Joseph Toscano

10:45am **Break**

11:00am Poster Session 1 **Barbie Tootle Room**

Table primes HAPPY: How acoustic variation activates meaning independent of rich lexical representations
Seung Kyung Kim & Meghan Sumner

Modeling adaptation to multiple speakers as (Bayesian) belief updating
Dave Kleinschmidt & Florian Jaeger

Visually-guided perceptual recalibration is phoneme-, cue-, and context-specific
Eva Reinisch, David R. Wozny, Holger Mitterer & Lori Holt

Effects of perceived talker characteristics on perceptual adaptation
Bridget Smith

12:00pm LUNCH BREAK

1:15pm Open Room Discussion on representations and cognitive architecture

2:15pm **Break**

Session: Individual Differences in Adaptation

Chair: Cynthia Clopper **Round Room**

2:30pm Neural and cognitive predictors of individual differences in perceptual learning
Frank Eisner

3:05pm Effects of indexical variation on the recall and recognition of spoken words
Meghan Sumner

3:40pm The voice of experience: The impact of individual and group attributes on talker-specific adaptation in speech
Lynne Nygaard

4:15pm **Break**

4:30pm Poster Session 2 **Barbie Tootle Room**

Spontaneous phonetic imitation as a predictor of perceptual recalibration
Molly Babel, Sophia Walters & Graham Haber

Listeners' pronunciations and how they perceive pin-pen merger
Kiwako Ito & Kathryn Campbell-Kibler

Improving non-standard dialect intelligibility in noise through associative priming
Abby Walker

5:30pm Open Room Discussion on individual differences in adaptation

6:30pm Day End

7:00pm Workshop Dinner at George Wells Knight House

Sunday April 7th

8:30am	Breakfast	
Session: Adaptation and Acquisition		
	Chair: Mary Beckman	**Round Room**
9:00am	Evidence for lexically driven adaptation in early development *Katherine White*	
9:35am	Bi-directional talker-listener adaptation across a language barrier *Ann Bradlow*	
10:10am	**Break**	
10:25am	Poster Session 3	**Barbie Tootle Room**
	Phonetic convergence and talker linguistic distance: Fine-grained acoustic and holistic measurements *Midam Kim & Ann Bradlow*	
	Evidence for talker-specificity and generalization in perceptual learning *Cheyenne Munson & Bob McMurray*	
	Aspects of modeling the learning of vowel normalization *Andrew Plummer*	
	Phonological inference and adaptation to cross-category vowel mismatches *Kodi Weatherholtz*	
11:25am	Open Room Discussion	
12:25pm	Conference End	

Perceptual and Cognitive Constraints on Vowel-Space Adaptation

Delphine Dahan[1]

1. Department of Psychology, University of Pennsylvania, U.S., dahan@psych.upenn.edu

Speech perception is an adaptive faculty: Experience that listeners have had with a talker's speech affects how they subsequently interpret speech from that talker. Encountering a talker's speech sounds in lexical contexts that constrain their categorization causes listeners to adjust the phonetic space in accordance to this experience, as revealed by subsequent sound categorization. My research investigates the perceptual and cognitive mechanisms that support restructuration and enable transfer. In this presentation, I will describe research examining the temporal constraints that operate on transfer in the context of vowel adaptation in a naturally occurring dialect of American English in which the vowel /a/ (as in rock) is shifted to resemble an instance of the vowel /æ/ (as in rack). When people are exposed to exemplars of the shifted vowel in words that can accommodate only one interpretation (thereafter anchor words, e.g., 'drop', where 'drap' in not a word of English), they are more likely to interpret ambiguous cases, such as 'rock', in a dialect-consistent manner than people who haven't experienced the shifted vowel in lexically constraining environments. However, the effect is modulated, among other things, by the temporal relationship between the anchors' vowels and test vowels: People are more likely to adopt a dialect-consistent interpretation of the test word when the anchor word appears within the same utterance than when the anchor word appears in the utterance immediately preceding the test word. This finding suggests that listeners' apparent adaptation to a talker's vowel space results from exposure to instances with a given vowel category accumulated over time, but also from a transfer between two vowel exemplars occurring within the same utterance, where their physical similarity may be most salient and transfer, most readily deemed appropriate. Features of the integration window over which such proximal transfer operates were revealed by a follow-up study where anchor and test words occurred within the same phrase. Results showed that transfer is as likely to take place whether the anchor precedes or follows the test word. I will discuss the implications of these findings with respect to the process and representation involved in phonetic-space adaptation.

Factors that Affect Phonetic Adaptation: Exemplar Filters and Sound Change

Keith Johnson[1]

1. Department of Linguistics, University of California Berkeley, U.S., johnson@berkeley.edu

This talk will summarize some recent research on phonetic accommodation and on compensation for altered auditory feedback. The data argue against a baseline view in which all exemplars of a category go into the formation of the category, and suggest instead that factors as varied as linguistic/phonetic crowding, and personal characteristics like prejudice and sense of personal power may limit the impact of experience on future performance. For example, Babel (2012) found that racial attitude influences whether listeners/speakers will accommodate a speaker, Katseff (2010, Katseff, Houde, Johnson, 2011) found that linguistic crowding influenced speaker's response to altered auditory feedback, and Dimov, Katseff & Johnson (2012) found that the speaker's sense of personal power modulated his/her response to altered auditory feedback. These findings and others like them help us understand some limits on the role of experience in the formation of linguistic/phonetic categories, and also helps us make a connection between language sound change as a group-level phenomenon and the individual personal patterns of behavior that must form a foundation for sound change.

Acquiring and Adapting Phonetic Categories in a Computational Model of Speech Perception

Joseph Toscano[1]

1. Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, jtoscano@illinois.edu

Recent work on perceptual adaptation has demonstrated that listeners can learn novel distributions of acoustic cues in unsupervised learning tasks with only a small amount of experience (Clayards, Tanenhaus, Aslin, & Jacobs, 2008, *Cognition*; Munson, 2011, dissertation). The learning problem faced by listeners in these tasks is similar to the one faced by infants acquiring the phonetic categories of their native language. In both cases, sounds are unlabeled and representations must be updated continuously as new input is received.

Can the same unsupervised learning algorithms that listeners use to acquire categories over development be used to adapt those categories in adulthood? Here, I present a computational model of speech perception (a Gaussian mixture model; McMurray, Aslin, & Toscano, 2009, *Developmental Science*) and simulations designed to address this question. The model represents phonetic categories as Gaussian distributions along acoustic cue dimensions, and it learns to map cues onto categories using a competitive statistical learning mechanism.

Previous work has shown that the model can successfully acquire phonetic categories. Given this, we can ask whether it can also adapt those categories in a perceptual learning task. These two processes are typically viewed as distinct: Language acquisition is seen as a slow process that occurs early in development and produces stable long-term representations, whereas adaptation is seen as a rapid process that can occur over the course of an hour and may produce only transient changes. As a result, it is not clear whether the learning rates that lead to successful development will also lead to successful adaptation. Moreover, it is unclear whether listeners' behavior in perceptual learning experiments can be explained via adaptation of long-term representations of phonetic categories.

We examined these issues in the context of the perceptual learning task presented in Munson (2011). In this study, listeners heard words varying in voice onset time (VOT) between minimal pairs differing in word-initial voicing. VOT-values were drawn from distributions in which the category boundary between voiced and voiceless tokens was either short (15 ms VOT) or long (35 ms). After hearing 300 tokens drawn from one of these distributions, listeners had adapted their category boundaries in the direction consistent with the distribution they heard.

The model was tested in the same task. First, we assessed its ability to correctly learn English voicing categories at a variety of learning rates. As expected, slower rates were more likely to yield successful acquisition and produced more stable categories. Next, we trained the model on VOT-values drawn from the distributions in Munson (2011) to ask whether a subset of the learning rates that worked for development also allowed for rapid adaptation. We found that this was the case: A common set of parameters can produce both successful acquisition and successful adaptation.

These simulations show that relatively simple unsupervised learning algorithms are sufficient for explaining speech sound learning on vastly different time-scales without changes in plasticity. Further, they suggest that some aspects of perceptual adaptation can be explained simply by the adjustment of listeners' long-term phonetic category representations.

Table Primes HAPPY: How Acoustic Variation Activates Meaning Independent of Rich Lexical Representations

Seung Kyung Kim[1] and Meghan Sumner[1]

1. Department of Linguistics, Stanford University, skim@stanford.edu

As language users, we have established associations between speech patterns and linguistic units, such as sounds and words. We have also established associations between speech patterns and speakers. Using these learned associations, we make inferences about a speaker's age, gender, accent, and emotional state. An understanding of variation as a carrier of information instead of redundant noise led directly to theories of representation and lexical access that incorporate variation in language production and perception via acoustically-rich lexical representations. Considering the massive amount of variation (and information) listeners have at their disposal when confronted with speech, we might ask what the role of variation is to the process of understanding spoken words beyond what detailed lexical representations offer. In this paper, we consider this perspective by investigating the recognition of words produced with different emotions.

Research examining emotional prosody is gaining ground, led by Nygaard and colleagues who have shown that listeners are faster to shadow words that are congruent with the emotional prosody (e.g., *happy* uttered with a happy prosodic contour) than words incongruent with the emotional prosody (e.g., ***happy*** uttered with a sad prosodic contour; Nygaard & Queen, 2008). Nygaard and colleagues also showed that specific word learning is predicted by the emotional prosody of novel words (Nygaard, Herold, and Nami, 2009). Studies of emotional prosody provide a path to tease apart word-specific effects from acoustic-variation effects more broadly. So far, though, these studies are consistent with word-specific explanations of spoken word recognition.

We test the hypothesis that meaning indexed from emotional prosody activates words related to that emotion independent of the lexical carrier of emotion. Specifically, we examine whether a prime, like ***table***, facilitates recognition to a visual target word that represents an emotion (e.g., *HAPPY*) when the prime is uttered in a happy voice than when it is uttered in a neutral voice. In other words, does a prosodic pattern prime an emotion-related lexical item? Critically, our primes are chosen at random and are semantically-unrelated to any particular emotion. To expand the set of words examined, we conducted a mass-testing pre-test to find semantic associates for different emotion words (e.g., happy, sad, bored, angry). We then paired random words with the target emotion and their associates (e.g., table – HAPPY; paper – SMILE). The main manipulation was whether the prime was produced in a neutral voice or an emotion voice (e.g., happy). We found robust facilitation for HAPPY-targets and associated targets (e.g., SMILE) when preceded by happy-voice primes compared to neutral-voice primes. We take this as evidence that listeners simultaneously process meaningful variation independent of the lexicon. To bolster this argument, we present results from a second study examining priming dependent on associate strength and argue that emotional prosody activates a concept in parallel with words, and that neutral prosody is the default, more strongly activating the lexicon. We couch these results in an approach that considers lexical specificity as given, but expands on this notion to include simultaneous activation of socially-meaningful cues and concepts independent of the lexicon.

Modeling Adaptation to Multiple Speakers as (Bayesian) Belief Updating

Dave Kleinschmidt[1] and Florian Jaeger[1]

1. Department of Brain and Cognitive Sciences, University of Rochester, dkleinschmidt@bcs.rochester.edu

The lack of invariance in language comprehension is due in large part to differences between speakers. Listeners deal with such systematic variation by rapidly adapting to changes in the statistics of their linguistic environment [1–6].

We propose that this adaptation can be modeled as incremental belief updating of the listeners the listener's uncertain, probabilistic representations of how a speaker will realize their intentions at different linguistic levels. Bayesian inference provides a powerful computational framework for describing such belief updating. Within this framework, we have developed an explicit model of phonetic recalibration/perceptual learning. Like other Bayesian models of speech perception, this model treats phonetic categories as probability distributions over acoustic/phonetic cues. In such a framework, phonetic categorization is accomplished by inferring the speaker's intended category label given a particular cue value using Bayes Rule (see e.g. [6,7]). Our model builds on existing work by adding an additional layer of inference (again via Bayes Rule), whereby the listener's beliefs about the speaker's categories are updated to take into account the speaker's recent productions. Thus, each noisy observation plays two roles: it provides information first about the current intended category and second about how this speaker realizes their categories generally.

One particular advantage of this Bayesian framework is that it quantifies how belief updating depends on the amount of recent experience and the strength of prior beliefs. Bayesian belief updating accounts for the effects of cumulative exposure on phonetic recalibration observed by [4], and, intriguingly, similar effects observed in selective adaptation experiments. Selective adaptation is widely attributed to separate mechanisms, but this model provides a conceptual bridge between the two phenomena, by emphasizing how they both arise from exposure to unusual statistics.

Thus far, this approach has dealt with adapting to a single speaker. We propose a hierarchical generalization of this model, where phonetic categories for different groups of speakers are represented in a loosely-linked fashion with similar speakers grouped together into a cluster and new clusters added as necessary. Such a representation can be learned as in distributional learning models of phonetic category learning [8–10]. Specifically, Bayesian non-parametric models can infer speaker clusters based on similarity, without assuming a certain number of clusters. This has the potential to resolve the paradox between pervasive adaptation and the overall stability of phonetic categories (since new representations are learned as needed, without overwriting old ones), and specifically, the fact that phonetic adaptation is both rapid (since a new cluster inherits only weak prior beliefs from other clusters) and robust [11] (since category beliefs are specific to clusters). It also potentially explains the fact that generalization of phonetic adaptation depends on the overall similarity structure of training and test speakers [1,3], because speakers are clustered according to similarity.

Bayesian belief updating provides a promising framework for understanding speaker adaptation, in a way that generalizes to different linguistic levels and which complements and extends the emerging consensus that language comprehension is (often near-optimal) inference under uncertainty.

References

[1] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech.," Cognition, vol. 106, no. 2, pp. 707–29, Feb. 2008.
[2] D. Norris, J. M. McQueen, and A. Cutler, "Perceptual learning in speech," Cognitive Psychology, vol. 47, no. 2, pp. 204–238, Sep. 2003.
[3] T. Kraljic and A. G. Samuel, "Perceptual adjustments to multiple speakers," Journal of Memory and Language, vol. 56, no. 1, pp. 1–15, Jan. 2007.
[4] J. Vroomen, S. van Linden, B. de Gelder, and P. Bertelson, "Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build- up courses.," Neuropsychologia, vol. 45, no. 3, pp. 572–7, Feb. 2007.
[5] M. H. Davis, I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan, "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences.," Journal of experimental psychology. General, vol. 134, no. 2, pp. 222–41, May 2005.
[6] M. A. Clayards, M. K. Tanenhaus, R. N. Aslin, and R. a Jacobs, "Perception of speech reflects optimal use of probabilistic speech cues.," Cognition, vol. 108, no. 3, pp. 804–9, Sep. 2008.

[7] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, “The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference.,” Psychological review, vol. 116, no. 4, pp. 752–82, Oct. 2009.
[8] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, “Learning phonetic categories by learning a lexicon,” Proceedings of the 31st Annual Conference of the Cognitive Science Society, pp. 2208–2213, 2009.
[9] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, “Unsupervised learning of vowel categories from infant-directed speech.,” Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 33, pp. 13273–8, Aug. 2007.
[10] B. McMurray, R. N. Aslin, and J. C. Toscano, “Statistical learning of phonetic categories: insights from a computational approach.,” Developmental Science, vol. 12, no. 3, pp. 369–78, Apr. 2009.
[11] F. Eisner and J. M. McQueen, “Perceptual learning in speech: Stability over time,” The Journal of the Acoustical Society of America, vol. 119, no. 4, pp. 1950–3, 2006.

Visually-guided perceptual recalibration is phoneme-, cue-, and context-specific

Eva Reinisch[1], David R. Wozny[2], Holger Mitterer[3] and Lori L. Holt[2]

1. Department of Phonetics and Speech Processing, Ludwig Maximilian University Munich, evarei@phonetik.uni-muenchen.de
2. Department of Psychology, Carnegie Mellon University
3. Department of Cognitive Science, University of Malta

Listeners use lexical and visual (lipread) context information to interpret ambiguous sounds. For example, the last sound in the word “giraffe” is likely to be interpreted as /f/ even if its acoustics are ambiguous between /f/ and /s/. In the lexical context of “gira_” only /f/ leads to the interpretation of a real word. Importantly, when such ambiguous sounds are later encountered in a neutral context (e.g., "lea_" where both "leaf" and "lease" are words), listeners still tend to interpret them in line with the previous context. Similar effects have been shown with lipread context information. After hearing an ambiguous auditory stimulus between "aba" and "ada" coupled with a clear visual stimulus (e.g., lip closure in "aba"), an ambiguous auditory-only stimulus is perceived in line with the previously seen visual stimulus. Listeners are thought to have "recalibrated" their perception. However, for both types of context it remains unclear what exactly listeners are recalibrating: the perception of phonemes, or specific acoustic cues.

This question was addressed in a series of experiments using visual context to guide recalibration. An auditory "aba"-to-"ada" continuum was created for exposure such that only the formant transitions in the vowel cued the consonants' place of articulation. If listeners recalibrate phoneme categories they should interpret an auditory-only /b/-to-/d/ continuum in line with previous exposure not only in the context of "a_a" but also in the context of "i_i" where place of articulation is mainly cued by burst and frication but less so by formant transitions (different-cues-same-phoneme condition). If, in contrast, listeners recalibrate specific cues independently of phonemes, then generalization to "ama"-"ana" should be found where cues are made identical to exposure (formant transitions) but the perceived phoneme differs (same-cues-different-phoneme condition). Whereas recalibration was robust for all exposure stimuli, no generalization was found for either of the two conditions. This was the case when exposed to "aba"-"ada" as well as when "aba"-"ada" was the generalization continuum. To further explore this apparent specificity of visually-guided recalibration, generalization between "aba"-"ada" and "ubu"-"udu" was tested where in both cases formant transitions were made the only informative cues to consonant identity (same-cues-same-phoneme-different-context condition). Critically, again, robust effects were found for the exposure contrasts but generalization across acoustic contexts did not occur. This suggests that visually-guided recalibration is robust but restricted to the phoneme category experienced during exposure and to the specific manipulated acoustic cues in the specific acoustic context.

These findings contrast with studies of lexically-guided retuning where various kinds of generalization have been shown: across words, position in the word, and even across speakers. Therefore, we speculate that lexically-guided and visually-guided category retuning might not build on the same underlying processing mechanisms. Differences in the standard experimental paradigms (not discussed here) as well as the nature of the disambiguating information, and variability during exposure are likely candidate explanations for differences in effects.

Effects of perceived talker characteristics on perceptual adaptation

Bridget J. Smith[1]

1. Department of Linguistics, The Ohio State University, bsmith@ling.osu.edu

Perceptual learning is the process in which listeners use top-down processing to infer the phonemic status of an ambiguous sound in a word, and make generalizations from that to alter their perception of that sound (e.g., Norris, McQueen & Cutler 2003). Shadowing is a technique that elicits imitative speech patterns from listeners to model talkers, including sub-phonemic variation (e.g., Goldinger 1998). In an experiment that combined perceptual learning with shadowing, participants were exposed to pronunciation variants of the /tw/ cluster in familiar and unfamiliar words, with the intent of creating a small-scale sound change in perception and production.

One set of participants read definitions and sentences that were written in an exaggerated Appalachian dialect as in (1). Another group read definitions and sentences that were written in standard academic English, as in (2), with the definitions taken from the Oxford English Dictionary, and sentences written in a standard style. Both of these groups heard a retracted affricated variant for /tw/. A control group heard plain alveolar [tw], but read the standard English sentences and definitions. A second control group had no training, but only participated in an identification task, whose results illustrated the perceptual space for the related /tu/ before training.

(1) *Tweeter:* A loudspeaker that plays mostly high pitched sounds, so that if you turn it up real high, your dog will yowl.

Twive: Clayton waited to reach me the rope til the twive brung the raft closer to the pier.

(2) *Tweeter:* A small loudspeaker designed to reproduce accurately high-frequency sounds whilst being relatively unresponsive to those of lower frequency

Twive: As he boarded the boat, the unfamiliar twive of the vessel made him a bit sea-sick.

A lexical decision task and an identification task measured the degree of perceptual learning, and generalization to new talkers, new words, and new phonological environments. It was expected that the Appalachian condition would have less robust effects of perceptual learning, due to the stigmatized nature of the dialect, but the reverse was true. Participants in the Appalachian condition exhibited perceptual broadening of their /tw/ and /tu/ categories to include more variant pronunciations than any other group. It is hypothesized that listeners' expectation that the talkers spoke an unfamiliar dialect increased their receptivity to pronunciation variants, thereby allowing greater effects of perceptual adaptation over the course of the experiment.

References

Goldinger, Steven. 1998. Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105, 251–279.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. Cognitive Psychology, 47, 204-238.

Neural and Cognitive Predictors of Individual Differences in Perceptual Learning

Frank Eisner[1]

1. Max Planck Institute for Psycholinguistics, f.eisner@mpi.nl

Listeners can adjust to a wide range of variability in speech through processes of perceptual learning. This kind of learning is thought to alter the way in which acoustical cues in the speech signal are mapped to abstract linguistic units. Learning-induced changes are relatively long-lasting, and have a facilitatory effect on word recognition. Perceptual learning in speech has been studied at different levels, and I will discuss two recent lines of investigation in this talk: adjustments in response to relatively small deviations in the production of individual speech sounds, and adjustments to a global, spectro-temporal degradation of the speech signal.

Perceptual learning can occur after repeated exposure to idiosyncratic productions of a speech sound by a particular talker. Words which contain such unusually articulated sounds may initially be difficult to process by the listener, but after sufficient exposure the perceptual system is able to exploit various kinds of information that are present in the signal in order to infer the intended identity of the sound. Such potential sources of information include lexical, phonotactic, or visual-articulatory features. Learning then results in a change in the representation of the affected phoneme category, which affects how subsequent occurrences of that phoneme are processed. The change is coded in a manner that can be quite specific to the context in which learning occurred, for example to a specific talker or group or talkers, and does not necessarily generalise to other listening situations.

When speech is degraded globally through artificial processing of the signal, for example after resynthesis or vocoding, linguistic content as well as the identity of individual talkers or accents is often too difficult to perceive initially. However, many listeners can learn to overcome a global spectro-temporal degradation with sufficient exposure. For this kind of perceptual learning, individual differences in both the steepness of the learning curve, as well as the level of asymptotic performance, have been documented. Because these findings have implications for practical applications of perceptual learning, for example for rehabilitation from hearing impairment, there has been recent interest in what cognitive and neural mechanisms underlie such inter-individual variability.

The ability to adjust to variability in speech is an essential function of the perceptual system that facilitates comprehension in a variety of different listening situations, but has not yet been properly incorporated into cognitive models of speech comprehension. The picture emerging for perceptual learning is complex, suggesting that learning can be both generalised and specific, interact with higher-level cognitive systems as well as low-level perceptual processes, and is subject to considerable inter-individual differences.

The Voice of Experience: The Impact of Individual and Group Attributes on Talker-Specific Adaptation in Speech

Lynne Nygaard[1]

1. Department of Psychology, Emory University, lnygaar@emory.edu

The acoustic speech signal conveys an enormous amount of information not only about linguistic content, but also about specific characteristics of individual speakers. Vocal properties provide socially relevant information to the listener about factors such as talker identity, age, sex, social status, health, and psychological state and listeners appear to readily detect, use, and accommodate to this *indexical* information during spoken language communication. Less well known is what accounts for variation in listeners' ability to identify and accommodate to particular talkers or groups of talkers. I will present data from a series of perceptual learning and adaptation studies examining task-, listener-, and talker-related factors that mediate accommodation to informative talker-specific variation in spoken language. In particular, I will focus on the contribution of 1) relatively short-term task-related changes in attention and expectation and 2) relatively long-term differences in individual listener's perceptual sensitivity. The outcome of this research suggests that although listeners dynamically adapt to systematic changes in linguistic structure as a function of experience, this adaptation depends on both the structure of the learning environment and individual differences in sensitivity to socially relevant variation. The considerable behavioral and representational plasticity that is characteristic of speech perception and spoken language processing may depend in part on the social, linguistic, and contextual relevance of talker-specific variation.

Effects of Indexical Variation on the Recall and Recognition of Spoken Words

Meghan Sumner[1]

1. Department of Linguistics, Stanford University, sumner@stanford.edu

Variation in speech carries a multitude of information to listeners. Indexical acoustic cues provide information about a speaker's age, gender, and accent. While we have come to view variation as engrained in lexical representations, we have yet to grasp exactly how listeners integrate the social and linguistic information extracted from the speech signal beyond effects mediated by a specific lexicon. In this talk, I examine the recall of spoken words across speakers using the false memory paradigm. In Exp. 1, I investigate the effects of indexical variation in the false recall of spoken words, presenting listeners with semantically-associated study lists produced by three speakers with different accents. The major finding is that the rates of falsely recalled items differ greatly depending on the speaker of the list. In Exp. 2, I examine this speaker effect in greater detail, attempting to assess the cause of the difference, using the same paradigm with full- or divided-attention, as attention has been shown to modulate false memories. The results from the second experiment suggest that speaker difference are process-based, not representation-based. Through two more experiments, I attempt to identify what the listener experience during multi-speaker tasks is, and suggest that as a whole, the data are best accounted for by appealing to differences in the allocation of attentional resources during encoding. I claim that indexical information that cues real or perceived speaker characteristics that influence the way listeners encode spoken language. Overall, the data support a picture in which spoken word recognition is a complex interactive process between linguistic and social cues and categories beyond a detailed lexicon.

Spontaneous Phonetic Imitation as a Predictor of Perceptual Recalibration

Molly Babel, Sophia Walters and Graham Haber[1]

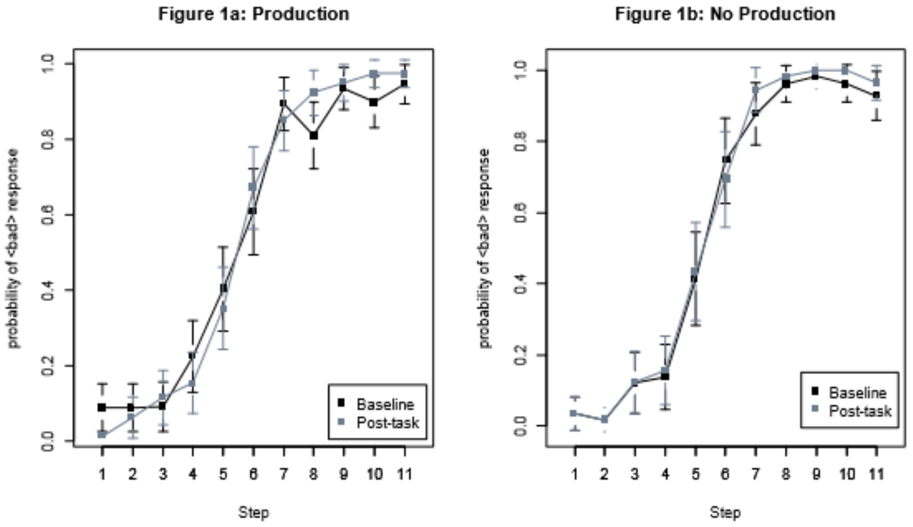
1. Department of Linguistics, University of British Columbia, molly.babel@ubc.ca

There is evidence for a connection between perception and production from the earliest stages of language acquisition. Infants' perceptual abilities are shaped by contrasts in the input (e.g., Cristià 2011, Werker & Tees 1984), and the speech infants eventually produce is driven by those contrasts. The perception/production link in adults has shown that production predicts individuals' abilities in perception (e.g., Ghosh et al. 2010, Hay et al. 2006). Research has also shown that listeners' perceptual categories are highly adaptable (e.g., Norris et al. 2003, Clayards et al. 2008), and phonetic imitation shows this malleability for speech production too (Goldinger 1998). Of interest here is how flexibility in production may relate to flexibility in perception both mechanistically and at an individual level. Some have touched on this: In sum, Shiller et al. (2009) find that changes in production lead to perceptual changes; Kraljic et al. (2008) show no production changes as a result of perceptual shifts; and Baese-Berk (2010) finds changes in production as a result of perceptual changes, but not the reverse. Such findings are contradictory, but result from different methods. Shiller et al. and Kraljic et al. manipulated real words, probing the flexibility of pre-existing categories, and Baese-Berk trained listeners on new categories.

In this paper we integrate recalibration and imitation paradigms to examine adaptation using natural variation and real words. New Zealand English was used as stimuli. The vowels of NZE are such that TRAP is more like [ɛ], compared to [æ] in American English. Participants (N=12) completed the following tasks: (1) categorization of an 11-step *bed/bad* continuum; (2) production of baseline tokens using a picture naming task with 100 words, 20 target words with /æ/ and 80 fillers; (3) model talker exposure such that all participants heard the NZE model's productions – those in a Production group shadowed the model talker while those in a Listen group listened quietly; (4) finally, participants completed the *bed/bad* categorization task again.

Perception data were scored as *bed* or *bad* . Logistic regressions for the Production group found main effects of Step ($\beta = 0.89$, $p < 0.001$) and Block ($\beta = 1.28$, $p < 0.01$), and a Step x Block interaction ($\beta = -0.24$, $p < 0.001$). Analysis for the No Production group found a main effect of Step ($\beta = 1.09$, $p < 0.001$), but no effect of Block, suggesting that only the Production group adjusted their perceptual categories. Figure 1 shows categorization functions for both groups. To assess how imitation relates to recalibration, the amount of F1 shift in /æ/ and the overall perceptual shift toward *bad* were computed for each participant in the Production group. There was a correlation between perceptual learning and phonetic imitation [$t(4) = -3.61$, $p < 0.05$, $r = -0.87$], shown in Figure 2. Positive perceptual accommodation indicates perceptual learning, and negative speech imitation shows a decrease in F1 – i.e., imitation of the model. Individuals who accommodated more in production, accommodated more in perception.

While we are collecting more data, these interim results suggest that encoding for speech production affects how perceptual input is used in recalibration and that an individual's propensity to imitate may be related to perceptual sensitivity. These results suggest the connection between perception and production may be more streamlined for established phonetic categories, as opposed to the connection forged when learning new phonetic categories.



Figures 1a and 1b. The probability of responding *bad* in baseline and post-task for both groups.

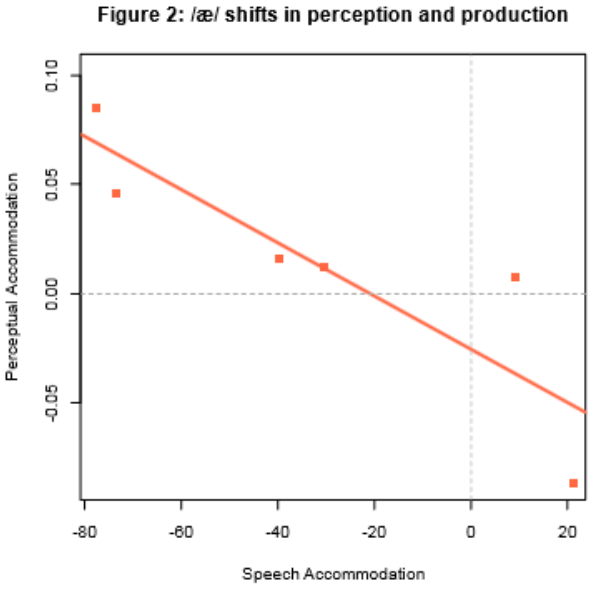


Figure 2. The relationship between perceptual recalibration and imitation for the Production group.

Listener's pronunciations and how they perceive the PIN-PEN merger

Kiwako Ito[1] and Kathryn Campbell-Kibler[1]

1. Department of Linguistics, The Ohio State University, ito@ling.ohio-state.edu

Listeners quickly adapt to speaker-specific pronunciations and use such knowledge in subsequent lexical processing [1, 2] . Previous work suggests that visual sociolinguistic information about speakers may bias the perception of dialect s [3]. While individuals' experiences with the target variety must largely influence its processing, little work has investigated how listener variability affects their dialect perception. This study reports how listeners' own pronunciations relate to the speaker - adaptation and the use of sociolinguistic cues for processing pin-pen merger by comparing those who merge the vowels themselves and those who do not.

Participants performed a visual search task following instructions (e.g., "Click on the pencil.") given in four voices: two that lowered their /I / to /ɛ/ and two that separated the vowels. A photo of a White or Black face in professional or non-professional dress accompanies each voice. Trials were divided into three blocks: Block 1 presented /ɛn/-words (fence) produced clearly with /ɛ/ by all four speakers. Block 2 presented /In/-words that were pronounced with /I/ by the non-merged speakers but with /ɛ/ by the merged speakers (/mɛnt/ for mint). Block 3 presented /ɛn/-words again to test whether learning speaker-specific pronunciations modulates listeners' responses to words with /ɛ/, which then should activate only /ɛn/-words for non- merged voices but both /ɛn/- and /In/-words for merged voices.

Before the eye-tracking task, participants read aloud the labels for objects presented one by one. Participants' productions of /ɛn/ and /In/ were analyzed using principal component analyses. Comparing the overlaps between the two vowels, 15 most-merged (MM) and 15-least merged (LM) listeners were selected among the 80 total participants.

Group differences surfaced in Block 1, before participants heard the critical dialectal differences. The LM group fixated the /ɛn/-targets faster with the merged than with the non- merged voices (the merged voices' /ɛ/ had slightly higher F1 than the non-merged), while the MM group showed no voice effect. Thus, the LM group showed sensitivity to subtle acoustic differences in the /ɛ/ vowels. Having experienced the merger in Block 2, both groups showed numerically faster fixations to the /ɛn/-targets with the non-merged voices and slower fixations with the merged voices in Block 3 than in Block 1. However, the LM group showed only the facilitation effect for the non-merged voices whereas the MM group showed only the slowing effect for the merged voices, exhibiting the biased sensitivities to the pronunciations closer to their own. In addition, the LM group showed an interaction between voice and race in Block3, where the non-merged voices led to faster target fixations with White than with Black faces and the merged voices showed faster target fixations with Black than with White faces. In the MM group, the non-professional dress led to faster target fixations than professional dress, although with a decrease from Block 1 to 3. These results demonstrate that listeners more sensitively respond to the pronunciation patterns similar to their own, and that speakers with different sociolinguistic backgrounds may attend to different social cues during processing.

References

[1] Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? Cognition 108 , 710–718.

[2] Trude, A. M. & Brown-Schmidt, S. (2012).Talker-specific perceptual adaptation during online speech perception. Language and Cognitive Processes, 27, 979-1001.

[3] Koops, C., Gentry, E. & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. University of Pennsylvania Working Papers in Linguistics, 14 (2), 93–101.

Improving non-standard dialect intelligibility in noise through associative priming

Abby Walker[1]

1. Department of Linguistics, The Ohio State University, ajwalker@ling.osu.edu

This study combines two fields of inquiry by asking whether priming a dialect improves the intelligibility of that dialect in noise. Work by Hay and colleagues (2006, 2008, 2010) has suggested that merely priming a dialect region may be sufficient to cause perceptual adaptation to that dialect, but this has only been shown using the Niedzielski paradigm (1999). In work on sentence intelligibility in noise, Clopper and Bradlow (2008) found that, for American English listeners, a General American (Gen Am) dialect (collapsed Midland, West, New England) was more intelligible than Northern, Southern and Mid-Atlantic dialects, and that this affect appeared to be independent of where the listener came from. In this study, we test whether we can improve the intelligibility of Mid-Atlantic speakers in noise by associatively priming the Mid-Atlantic accent. Additionally, we test what happens to the intelligibility of other non-standard accents (Southern, Northern) when the Mid-Atlantic is primed.

This study uses the same materials as Clopper and Bradlow: 98 unique HP SPIN sentences from 6 speakers (3 M, 3 F) from each of the dialect regions Gen Am, Northern, Southern, and Mid-Atlantic, mixed with speech shaped noise at -2dB SNR. The sentences were divided evenly into two blocks, and all participants heard both blocks. Before each block, participants saw the names of three Mid-Atlantic or Midland cities. In the implicit condition, these place names flashed on the screen before the block. In the explicit condition, participants were told the speakers they were about to hear came from these cities. In both the implicit and explicit conditions, participants do both a Mid-Atlantic and a Midland primed block.

Preliminary analysis of 63 native listeners shows that listeners perform significantly better on Mid-Atlantic accents when primed by Mid-Atlantic place names compared to Midland place names (though Mid-Atlantic accents are still the most difficult accent). At this stage, there is no difference in whether the prime is implicit or explicit. The primes do not affect the intelligibility of Northern and Midland accents. However, the primes do significantly affect the intelligibility of Southern accents: priming Mid-Atlantic causes a decrease in the intelligibility of Southern speakers compared to priming Midland, and it is worse when the prime is explicit.

Data collection is ongoing, but these results appear to confirm the findings of Hay et al. that associatively priming a dialect results in perceptual shifts to that dialect. Additionally, the results suggest that one of the reasons behind the difficulty in understanding non-standard accents in noise is that they are not expected. With more participants, we hope to be able to further investigate the effect of explicit vs. implicit primes, the longevity of primes through the two blocks, and the familiarity with and attitudes towards the dialects that our participants report at the end of the experiment.

Bi-directional talker-listener adaptation across a language barrier

Ann Bradlow[1]

1. Department of Linguistics, Northwestern University, bradlow@northwestern.edu

A language barrier is both a persistent source of potential miscommunication and an impetus for speech and language flexibility. In the present research we explore talker and listener mutual adaptation in response to the presence of a language barrier as a possible short-term individual-level mechanism of long-term population-level contact-induced language change.

I will begin this presentation by briefly alluding to work that examined talker adaptation to the listener, that is to work on native and non-native perception and production of clear speech (i.e. listener-oriented talker modifications with the purpose of enhancing speech intelligibility under difficult listening conditions). Then, I will focus on the other side of the talker-listener adaptation cycle, that is, on listener adaptation to the talker. This work demonstrates increasingly generalized listener adaptation to foreign-accented speech in response to training regimens that incorporate increasingly expansive dimensions of systematic variation. Specifically, we observe talker-dependent adaptation following exposure to a single foreign-accented talker, talker-independent (but accent-dependent) adaptation following exposure to multiple foreign-accented talkers from the same native language background, and finally accent-independent adaptation following exposure to multiple foreign-accented talkers from different native language backgrounds. In an attempt to understand the task requirements for listener adaptation to foreign-accented speech, we investigate a training regimen that includes trials requiring focused attention to the task of foreign-accented speech recognition ("active" task performance) as well as trials that involve exposure to foreign-accented speech while performing an unrelated visual attention task ("passive" exposure only). A comparison of this combined active+passive training regimen and the more typical all-active laboratory training regimen showed that active+passive training can provoke as much perceptual adaptation as all-active training, suggesting that immersion conditions can promote highly efficient perceptual adaptation to foreign accented speech.

Finally, I will discuss talker-listener adaptation under more naturalistic conditions of spontaneous dialogues in which talker-to-listener and listener-to-talker adaptation can occur interactively. In this work we examined communicative efficiency and phonetic convergence in conversations between pairs of talkers that varied along a "language distance" scale. The closest end of the language distance scale was represented by talker pairs with two native talkers of the target language from the same dialect region and the farthest end of the scale was represented by talker pairs with one native talker and one relatively low proficiency non-native talker of the target language. Results from this dialogue-based work showed a negative correlation between language distance and communicative efficiency, as well as a negative correlation between language distance and phonetic convergence. However, we also found evidence for a mitigating effect of phonetic convergence on the negative correlation between language distance and communicative efficiency, suggestsing that phonetic convergence between interlocutors with a relatively great language distance between them may be an effective mechanism for overcoming the detrimental effects of the language barrier.

Taken together, these studies build a picture of speech communication across a language barrier as an opportunity for bidirectional talker-listener adaptation, raising the possibilitity that these relatively short-term adaptations may lay the foundation for longer-term speech and language change.

Evidence for Lexically Driven Adaptation in Early Development

Katherine White[1]

1. Department of Psychology, University of Waterloo, white@uwaterloo.ca

There is abundant evidence that adults can adapt to novel accents after a short period of exposure to either artificially controlled or naturalistic accents. Two important characteristics of this adaptation process have emerged: First, lexical knowledge plays a powerful role in driving adaptation. When faced with atypical phonetic properties (e.g., a speaker with unusually short VOTs), activation of intended lexical items can drive reinterpretation of those phonetic properties. Further, in some cases – when there is a complete mismatch between the phonemic properties of the accented form and the stored version (e.g., when a speaker pronounces "pen" as "pin") – knowledge about the intended lexical item is critical for adaptation. Second, adaptation in adults appears to involve the re-tuning of phoneme categories or system-wide remappings between the phonological level and the lexicon. For example, exposure to an atypical speech sound in the context of a lexical decision task affects later categorization of similar speech sounds (Kraljic & Samuel, 2005; Norris,McQueen & Cutler, 2003), in some cases, extending even to new phonemes not presented during the exposure (Kraljic & Samuel, 2006). Moreover, hearing atypical pronunciations can affect the recognition of phonologically related words not heard during the exposure period (Maye et al., 2008; McQueen, Cutler & Norris, 2006). Thus adaptation does not simply involve memorization of specific items heard during the exposure phase, but is more generalized.

An important question is whether these characteristics of adaptation in adults are also present earlier in development. There are reasons to think that they may not be. First, young word learners are, in many cases, highly sensitive to the phonetic properties of words (Swingley & Aslin, 2000; White & Morgan, 2008) and they are strongly biased to assume that novel wordforms apply to novel referents (Markman, 1990). This sensitivity may make them overly conservative in the face of phonetic deviations, causing them to fail to map accented forms onto known lexical items and thus preventing adaptation. Second, it is likely that top-down knowledge exerts less influence in children than it does in adults, as has been demonstrated in other areas of language processing (Trueswell et al., 1999).

I will present studies from my lab and others demonstrating that, in fact, toddlers and young children *do* make use of lexical information to interpret novel accents (both artificially controlled and naturalistic). Further, as in adults, adaptation generalizes beyond the specifics of the exposure (White & Aslin, 2011; McQueen et al., 2012; Van Heugten & Johnson, in press). As a whole, these findings suggest that the features of the adaptation process are quite similar in child and adult listeners.

Phonetic convergence and talker linguistic distance: fine-grained acoustic and holistic measurements

Midam Kim[1] and Ann Bradlow[1]

1. Department of Linguistics, Northwestern University, midamkim@gmail.com

This study explores the linguistic conditions of phonetic accommodation, its measurement, and the extent of generalization. Specifically, we investigated native English talkers' speech production modifications after passive auditory exposure to model talkers with varying linguistic distances from the participants. In a prior study, Kim et al. (2011) found a negative effect of talker linguistic distance on phonetic convergence in a conversation. In this study, we ask the same question in a non-interactive condition. That is, where is the linguistic limitation of convergence without social interactions between talkers? Second, while accommodation was measured with only holistic human judgments in Kim et al. (2011), we used both fine-grained acoustic measurements and holistic measurements to see what really changes in phonetic accommodation. Lastly, following Nielsen (2011), we examined whether accommodation to exposed materials transferred to unexposed materials.

For the experiment, we recorded four female model talkers reading 126 words and 64 sentences in English. Two of the model talkers were native English talkers with US northern dialects, while the other two were Korean nonnative English talkers with high English proficiency. In the experiment, 67 female native American-English talkers read all words and sentences before and after exposure to half of the materials either visually or auditorily. Among the 67 participants, 20 were assigned to the control group, where they were exposed to half of the materials visually rather than auditorily during the exposure phase. Each of the 47 participants in the experimental groups heard half of the materials read by one of the four model talkers during the exposure phase. Approximately half of the participants in the experimental groups had US Northern dialects, while the other half did not. Various acoustic measurements were performed on the words recordings. With the sentences recordings, we performed human and computational holistic measurements: an XAB perception test with a separate group of 55 participants and dynamic time warping analyses.

We found evidence of native English talkers' convergence to all model talkers, regardless of their regional dialects and native status. With the acoustic measurements, we found that the pre-exposure acoustic distances between model talkers and participants positively affected phonetic convergence; the farther the pre-exposure acoustic distance was, the larger their degree of phonetic convergence was. Moreover, while human and computational holistic judgments revealed different accommodation patterns towards the four model talkers, the human judgments exhibited convergence towards all model talkers, and importantly, this perceived accommodation was positively predicted by the computational judgments. Finally, the accommodation patterns on exposed items and unexposed items were not significantly different. Altogether, these results show acoustic and perceptual evidence of phonetic accommodation to both native and nonnative talkers and its possibility to change talkers' linguistic systems in the long term.

References

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. Laboratory Phonology, 2(1), 125-156.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. Journal of Phonetics, 39, 132-142.

Evidence for talker-specificity and generalization in perceptual learning

Cheyenne Munson[1] and Bob McMurray[2]

1. Department of Psychology, University of Illinois at Urbana-Champaign, cheyenne@illinois.edu
2. Department of Psychology and Department of Communication Sciences and Disorders, University of Iowa, bob-mcmurray@uiowa.edu

While the acquisition of sound categories is generally considered to be a process that is largely completed during infancy, there is considerable evidence that these categories remain malleable even in adulthood and that this may be helpful for processes like talker and dialect adaptation. A large body of work has shown evidence for a form of perceptual learning in adult listeners that allows them to rapidly adjust their category boundaries (e.g., adjusting to a new voicing boundary along a VOT continuum) to better match manipulated speech input that they are exposed to in laboratory settings (e.g., Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2005; Clarke-Davidson, Luce, & Sawusch, 2008).

One aspect of this process that has attracted significant attention is the degree to which perceptual learning generalizes both to novel contrasts and to novel talkers. This is a particularly important aspect of perceptual learning because it can help reveal the fundamental units over which learning occurs. For example, if VOT boundaries learned for one talker generalize to another talker, it implies that the speech perception system includes some level of representation that is talker independent. In contrast, if all learning is talker-specific it implies a more or less direct mapping between the signal (including indexical characteristics) and phonetic categories. Models of speech perception that posit abstract units (e.g., TRACE and Merge) can accommodate generalization across talkers but not talker-specific learning. In contrast, models that posit veridical representations of speech input (e.g., exemplar models) can accommodate talker-specific learning, but it is unclear what degree of generalization should be expected.

Perceptual learning research on generalization to novel talkers has shown evidence for talker-specific boundary learning (Allen & Miller, 2004; Theodore & Miller, 2010) but also for generalization across talkers (Eisner & McQueen, 2005; Kraljic & Samuel 2006; 2007). While talker-specific learning is not necessarily inconsistent with generalization across talkers, the apparent discrepancy in these results suggests that there may be some circumstances that lead to talker-specific learning and others that lead to generalization.

We performed two experiments using a statistical learning paradigm to assess (1) whether listeners learn talker-specific VOT boundaries in a task that does not require talker identification and has no instructions related to the talkers, and (2) whether simultaneous exposure to multiple talkers leads to talker-specific learning and sequential exposure leads to generalization across talkers. We found that listeners can learn talker-specific boundaries without prompting even when talkers are irrelevant to the task. In a second experiment we found that listeners also generalize perceptually-learned boundaries to a novel talker, and yet, they can still retain talker-specific boundaries after acquiring new boundaries for a second talker.

All of these results were obtained in a purely unsupervised learning paradigm, suggesting a powerful and flexible form of perceptual learning. These results indicate that models of speech perception must incorporate the use of talker-specific information at the level of phonological categories. In addition, these models must simultaneously maintain some route for generalization across talkers.

References

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 115(6), 3171-3183.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, 70(4), 604-618.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67(2), 224-238.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51(2), 141-178.
Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, 13(2), 262-268.
Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers*.* *Journal of Memory and Language*, 56(1), 1-15.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204-238.
Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *Journal of the Acoustical Society of America,* 128(4), 2090- 2099.

Aspects of modeling the learning of vowel normalization

Andrew Plummer[1]

1. Department of Linguistics, The Ohio State University, plummer@ling.osu.edu

We put forward a framework for the investigation of the learning of vowel normalization based on the idea that infants perform abstractions over their psychophysical representations of the vowels of individual speakers by mapping them to mediating spaces of representations, guided by vocal imitative interaction with their caretakers, as a first step in the phonological acquisition process. The framework is accompanied by a computational methodology for modeling the abstraction, which involves the "alignment" of cognitive structures, called "manifolds," that an infant builds from the psychophysical representations of the vowels of individual speakers. We conclude with a simple demonstration of the main algorithm involved in implementation of the methodology.

We take *vowel normalization* to be a cognitive process "in which interspeaker vowel variability is reduced in order that perceptual vowel identification may then be performed by reference to rela- tive vowel quality rather than absolute [psychophysical] parameters of vowels" (Johnson, 1990, p. 230). In this connection, we take the following to be a minimal collection of aspects essential to the modeling of the learning of normalization. The first is a "reference frame," which "can be thought of as a coordinate frame that best captures the form of information represented in a particular part of the nervous system" (Guenther, 2003, p. 209), though construed more broadly to include physical and cognitive domains (a` la Saltzman, 1995). The second essential aspect is that of a "manifold," a structure embedded within a reference frame and used to organize representations. Specifically, we make use of "vowel manifolds" (Jansen & Niyogi, 2006, 2007), physical structures over vowel sig- nals hypothesized to motivate an infant's formation of "perceptual manifolds" (Seung & Lee, 2000; Niyogi, 2004), and "cognitive manifolds" (Plummer, 2012), cognitive structures used by an infant in the normalization process. The third essential aspect is a computation over manifolds, called "manifold alignment" (Wang, 2010), which maps representations on two (or more) manifolds to a mediating "latent space" (Ham et al., 2005; Ma & Fu, 2012), where vowel identification may take place, or further cognitive computations. Alignment computations are guided by social interaction between an infant and adult caretakers characterized by specific types of vocal exchanges (Howard & Messum, 2011; Masataka, 2003; Fitch, 2010; Gros-Louis et al., 2006; Goldstein & Schwade, 2008). These exchanges broadly involve: (i) structured turn-taking between an infant and caretak- ers (Masataka, 2003), and (ii) caretaker responses differentiated according to the nature of infant vocalizations (Gros-Louis et al., 2006; Goldstein & Schwade, 2008).

We briefly exemplify the approach with the following simplified dramaturgical dyadic exchange. Let VI be a set of representations of vowels derived from an infant, and VA a set derived from an adult caretaker, both within a single acoustic reference frame (Figure 1). The adult may impart their systematic knowledge of the vowel categories [i], [u], and [a] to the infant by responding in a positive manner to infant productions in VI judged by the adult to be good examples of [i], [u], and [a], respectively, with their own productions taken to be good examples of [i], [u], and [a]. The yellow, orange, and green points, respectively, in Figure 1 approximate a series of vocal exchanges involving good examples of infant (left) and adult (right) [i], [u], and [a], as judged by the adult. The infant may then represent the positive interaction, pairing representations of their good productions with the corresponding representations of the positive adult responses. These pairs of positive representations guide the alignment of the manifolds the infant constructs over VI and VA, yielding the aligned structures in Figure 2, which may then be used for vowel identification, as well as for further cognitive computation.

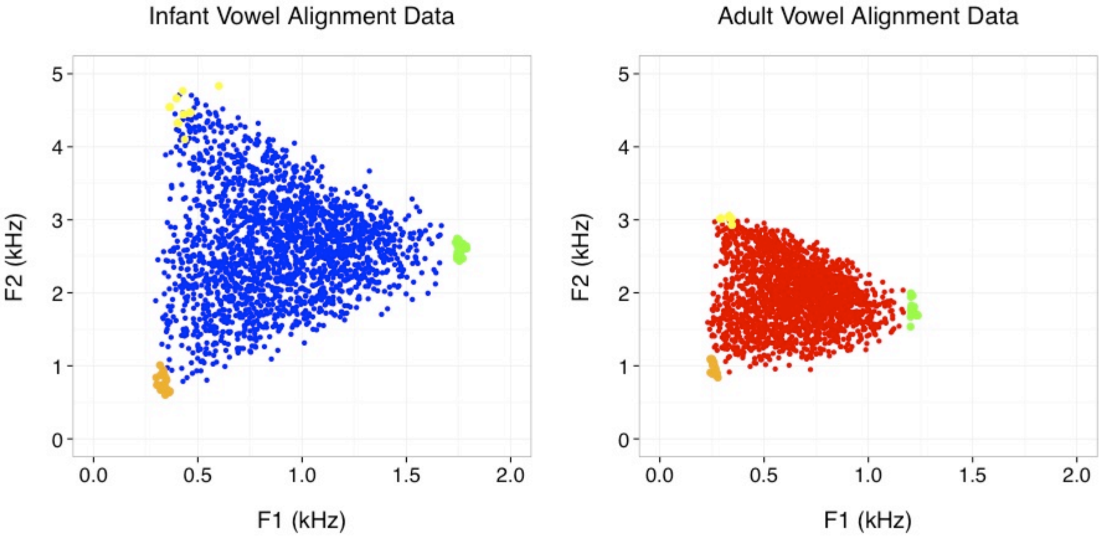


Figure 1: Infant vowel representations V_I (left) and adult vowel representations V_A (right), in a formant-based acoustic reference frame, together with “good” examples of infant and adult [i] (yellow), [u] (orange), and [a] (green) as judged by the adult.

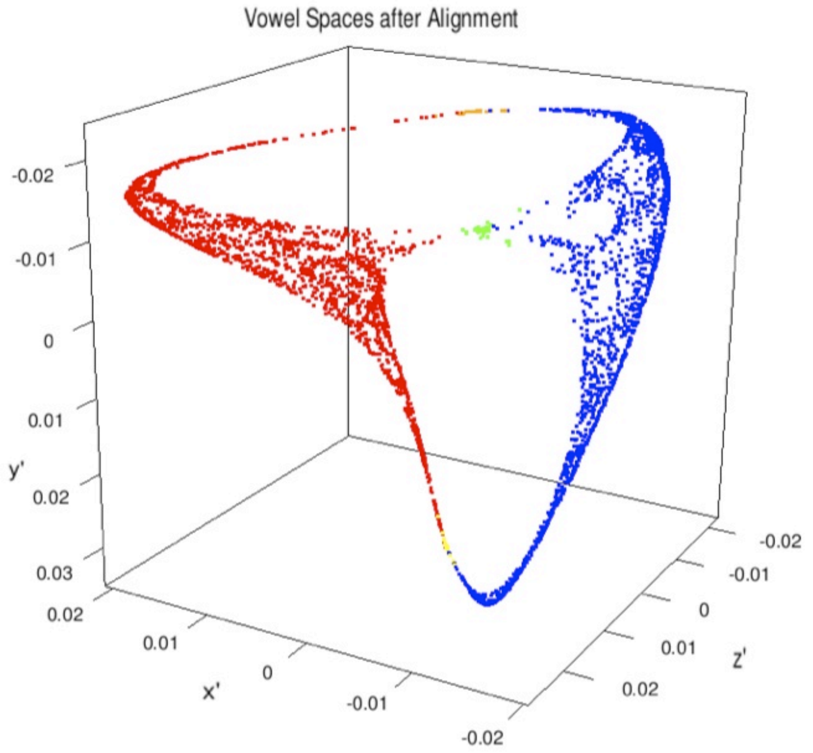


Figure 2: Representations in a reference frame where the alignment of manifolds over adult and infant formant representations has been achieved using the “good” infant and adult productions.

References

Fitch, W. T. (2010). The Evolution of Language. Cambridge University Press. Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. Psychological Science, 19(5), 515–523.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. International Journal of Behavioral Development, 30(5), 112–119.

Guenther, F. H. (2003). Neural control of speech movements. In N. O. Schiller, & A. Meyer (Eds.) Phonetics and Phonology in Language Comprehension and Production: Differences and Similarites, (pp. 209–239). Walter de Gruyter.

Ham, J., Lee, D. D., & Saul, L. K. (2005). Semisupervised alignment of manifolds. In Z. Ghahra- mani, & R. Cowell (Eds.) Proc. of the Ann. Conf. on Uncertainty in AI, vol. 10, (pp. 120–127).

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. Motor Control, 15, 85–117.

Jansen, A., & Niyogi, P. (2006). Intrinsic fourier analysis on the manifold of speech sounds. In in IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing, (pp. 241–244).

Jansen, A., & Niyogi, P. (2007). Semi-supervised learning of speech sounds. In Proceedings of INTERSPEECH 2007.

Johnson, K. (1990). Contrast and normalization in vowel perception. Journal of Phonetics, 18, 229–254.

Ma, Y., & Fu, Y. (2012). Manifold Learning Theory and Applications. CRC Press.

Masataka, N. (2003). The Onset of Language. Cambridge, UK: Cambridge University Press.

Niyogi, P. (2004). Towards a computational model of human speech perception. In Proceedings of the Conference on Sound to Sense, MIT (In Honor of Ken Stevens' 80th birthday).

Plummer, A. R. (2012). Aligning manifolds to model the earliest phonological abstraction in infant caretaker vocal imitation. In 13th Annual Conference of the International Speech Communica- tion Association (INTERSPEECH 2012). Portland, OR.

Saltzman, E. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. Mind as motion: Explorations in the dynamics of cognition, (pp. 149–173).

Seung, H. S., & Lee, D. D. (2000). The manifold ways of perception. Science, 290(5500), 2268– 2269.

Wang, C. (2010). A Geometric Framework For Transfer Learning Using Manifold Alignment. Ph.D. thesis, University of Mass. Amherst.

Phonological Inference and Adaptation to Cross-Category Vowel Mismatches

Kodi Weatherholtz[1]

1. Department of Linguistics, The Ohio State University, kweatherholtz@ling.ohio-state.edu

This study investigated perceptual adaptation to and generalization of learning about cross-category vowel variation. In two experiments, listeners were familiarized to a novel accent characterized by a cross-category remapping of the American English vowel space and then tested for whether this familiarization improved the recognition of words pronounced with either accent-consistent or accent-inconsistent vowel shifts. Results from Experiment 1 demonstrated that exposure to cross-category front vowel lowering improved recognition of words produced with both lowered (accent-consistent) and raised (accent-inconsistent) front vowels, despite the lack of evidence for a system of vowel raising. Experiment 2 replicated this finding with a different set of test materials: exposure to cross-category back vowel lowering improved recognition of words produced with lowered and raised back vowels. Taken together, these findings provide evidence that perceptual adaptation to a specific system of vowel variation generalizes beyond the input, enabling listeners to cope with a broader range of variability. This generalization finding is discussed in terms of processing benefits for listeners encountering accents characterized by partially opposing vowel shifts (such as the Northern Cities Shift and the Southern Shift).