précis of Perception in a Variable but Structured World: The Case of Speech Perception

Dave F. Kleinschmidt

Human speech recognition is a remarkable feat of perception in the face of variability and uncertainty. One of the biggest contributors to the difficulty of speech perception is that individual talkers vary substantially in the way they produce speech. Understanding how listeners cope with this talker variability—known as the "lack of invariance"—is one of the longest-standing problems for models of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). My dissertation develops a theory of speech perception that addresses this problem. This framework—the *ideal adapter* framework—is a computational-level theory (in the sense of Marr, 1982) which treats speech perception as a problem of *inference under uncertainty*. The main contribution of this theory is that it identifies three hierarchical levels of uncertainty that jointly inform each other. According to the ideal adapter, the listener must infer

1. *what* the talker is saying,
2. *how* the talker says things, and
3. *who* the talker is, in relation to the listener's prior experience.

This framework provides a unified perspective on a large empirical literature on how listeners deal with talker variability. More importantly, it also opens new doors to future work on speech perception and perception in general, by providing a theoretical framework and formal, computational tools for quantitative modeling of human perception behavior.

Chapter 2: The ideal adapter

(The work that constitutes this chapter was published as Kleinschmidt & Jaeger, 2015, Psychological Review)

A large body of empirical work shows that listeners employ a variety of superficially distinct strategies to deal with talker variability. In some experiments, listeners rapidly adapt to an unfamiliar talker, suggesting that speech perception is highly *flexible*. However, in others, listeners have highly stable, long-term memories for specific individual talkers that they heard hours, days, or months ago, memories which facilitate speech recognition (S. D. Goldinger, 1996; Nygaard & Pisoni, 1998). Moreover, in some cases these representations

are highly *talker-specific*, while in other cases listeners *generalize* what they have learned about one talker to another.

The ideal adapter provides a unified account of these phenomena. Like previous models of speech perception in the ideal observer/rational analysis tradition (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Norris & McQueen, 2008), the ideal adapter treats speech perception as a process of inference under uncertainty. Because speech production is variable, even within a single talker and linguistic context (Allen, Miller, & DeSteno, 2003; Miller, 2001), each linguistic unit (word, phonetic category, etc.) is realized as a *distribution* of acoustic cues. Thus, given a particular observed cue value, the best a listener can do is use their knowledge of these distributions to *infer* the probability that the talker intended to produce each possible linguistic unit, by comparing how well each distribution predicts the observed cue.

The ideal adapter proceeds from this with two additional central insights. The first is that, because of talker variability, these cue distributions *vary* from situation to situation. As a result, the listener never knows the true underlying distributions, but rather only has *uncertain beliefs* about those distributions. Because effective speech perception depends on knowing these distributions, the listener needs to infer not only *what* that talker intended to say but *how* they say things in general. Each observed cue value provides information about both the underlying distributions, and which category it was most likely generated from. Formally, this can be treated as a process of jointly inferring the intended category for each observed cue value *and* the underlying distributions (e.g., the mean and variance of each category's cue distribution, if these distributions are normal distributions).

The second insight of the ideal adapter is that there are two sources of information about the cue distributions in the current situation: the cues that the listener has currently observed, and their prior experience in other situations. In the language of Bayesian inference, the current observations determine the *likelihood* of each possible set of underlying distributions, while prior experience determine the *prior* probability assigned to each of these possibilities. This is critical because cue distributions do not vary arbitrarily from one situation to the next: individual talkers are relatively consistent over time (Heald & Nusbaum, 2015), and talkers of the same gender, age, regional origin, native language background, etc. produce similar cue distributions (e.g., Clopper, Pisoni, & Jong, 2005; Labov, Ash, & Boberg, 2005). This means that listeners can, in principle, benefit a great deal from their prior experience in many situations.

By considering These two factors together, the ideal adapter provides a unified perspective on how listeners cope with talker variability. When listeners are in a novel situation (like a laboratory experiment or meeting a completely unfamiliar talker), their prior experience is not very relevant, and they must rely primarily on the likelihood from the cues they observe in that environment. In such situations, the ideal adapter predicts rapid, incremental belief updating, as each cue is highly informative about the underlying distributions in the absence of strong prior beliefs. This is exactly what is observed (Kraljic & Samuel, 2007; Norris, McQueen, & Cutler, 2003; Vroomen, Linden, Gelder, & Bertelson, 2007), and a simple Bayesian model that implements this belief updating provides a very good fit ($R^2 > 0.8$) to data on how human listeners incrementally recalibrate their phonetic category

boundaries based on experience with a novel talker (from Vroomen et al., 2007, and my own extensions). Furthermore, this same model—using the same parameters—provides an equally good fit to the effect of repeated exposure to the same stimulus, which is typically considered to be a qualitatively different phenomenon ("selective adaptation", Eimas & Corbit, 1973; Vroomen et al., 2007, see the following section).

In many situations listeners *do* have relevant prior experience, such as when they encounter a particular familiar talker. The ideal adapter predicts that if a listeners recognizes a familiar talker in such situations, then their past experience is sufficiently informative that little to no additional information is needed to make an accurate inference about the underlying cue distributions. That is, if a listener's beliefs about the current cue distributions are *conditional* on the identity of the current talker (or more generally, the type of talker), then they can continue to update those beliefs incrementally *across situations*. As a result, each time they encounter that talker again, less and less information is required to hone in on the correct cue distributions. This is how the ideal adapter explains the apparent contradiction between the extreme *flexibility* of the speech perception system (as demonstrated by rapid recalibration; e.g. Norris et al., 2003) and the stability of talker-specific effects over long periods of time (Creel & Bregman, 2011; e.g., Creel, Aslin, & Tanenhaus, 2008; S. D. Goldinger, 1996; Nygaard & Pisoni, 1998; Palmeri, Goldinger, & Pisoni, 1993; Remez, Fellowes, & Rubin, 1997).

The power of treating speech perception as inference under uncertainty is that such recognition can be formalized as another level of inference, where the listener infers *who* is talking in the current situation. This inference combines information from top-down cues (a name, face, etc.) with bottom-up information from the speech signal itself (both from phonetic cues and others, like voice quality or pitch). The ideal adapter thus treats speech perception as multiply hierarchical inference under uncertainty, where listeners' prior expectations about the underlying cue distributions are themselves *conditional* on the type of talker that they think is currently speaking. One of the underexplored implications of this idea is for understanding how listeners decide whether to generalize what they have learned about the way that one talker produces a certain phonetic category to a different talker. The literature on this is apparently contradictory, with listeners sometimes showing talker-specificity (no generalization, Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007) and in other experiments generalizing across talkers (Kraljic & Samuel, 2006, 2007). If listeners are actively inferring how each talker relates to their previous experience with other talkers, then whether they generalize from experience with one talker to another depends on whether they infer the two talkers to be of the same type (in the relevant way). This inference, as with other inferences, takes into account both the bottom-up similarity in the cue distributions the talkers produce, as well as top-down expectations that a listener has about whether or not two different talkers will have substantially similar cue distribution for a particular category. At the most coarse level, the ideal adapter thus predicts that listeners should generalize experience for phonetic categories that are relatively consistent across talkers (like voice onset time, a cue to word-initial stop voicing), but should adapt to talkers separately for contrasts where there is substantial variability across talkers (like fricatives and vowels). This is, in fact, the general pattern that is observed. If these biases are really due to an active inference process, they should be able

to be overcome with sufficient experience that contradicts them. This prediction, too, is borne out in the empirical data (e.g., Munson, 2011; Reinisch & Holt, 2014).

In summary, the ideal adapter is a computational level framework for understanding speech perception that provides an answer for how listeners deal with talker variability, one of the oldest puzzles in speech perception. The central claim of this framework is that speech perception can be thought of as a problem of inference under uncertainty at multiple levels. Because of talker variability, listeners must engage in life-long distributional learning. The same argument applies to language comprehension at other levels as well: as long as there is variability in how people realize their linguistic intentions, listeners need to continuously update their beliefs about the statistical properties of language (for examples in syntactic expectations, see Kleinschmidt, Fine, & Jaeger, 2012; and in the interpretation of quantifiers like "some" and "many", Yildirim, Degen, Tanenhaus, & Jaeger, 2016).

To make this distributional learning tractable, listeners cannot just track changes in the underlying cue distributions of speech sounds. Rather, they need to build up internal models of how these distributions vary across situations, based on the type of talker and other variables. This second point has implications for theories of perception and motor control in general: in a variable but structured world, agents can benefit from not simply tracking changing sensory statistics but by learning the structure of the world, a point that has been largely overlook in work on sensory adaptation and perceptual learning.

Chapter 3: Adaptation in speech and general perception

(The work that constitutes this chapter was published as Kleinschmidt & Jaeger, 2016, Psychonomic Bulletin and Review)

Adaptation is ubiquitous in sensory systems: repeated presentation of a stimulus leads to reduced sensitivity to that stimulus, in terms of both behavioral and neural measures (Kohn, 2007; for a review, see e.g. Webster, Werner, & Field, 2005). Adaptation—in speech perception and in general—is typically thought of as a sort of "feature detector fatigue". In this view, repeated presentation of a particular stimulus leads to decreased response to that stimulus because neural populations quickly habituate to presentation of their preferred stimuli. Despite recent findings in non-linguistic adaptation that cannot be accounted for as simple detector fatigue, this view remains common, especially in work on speech perception.

The ideal adapter suggests an alternative way of looking at adaptation: as another manifestation of distributional learning, just like perceptual recalibration. Repeated presentation of a prototypical exemplar of a phonetic category leads to a sharpening of the representation of that category, because such repeated presentation corresponds to an unusually peaked distribution. This provides a good quantitative fit to the incremental build-up of adaptation with further exposure, as well as puzzling adaptation-like effects with long-term exposure to recalibrating stimuli (Samuel, 2001; Vroomen et al., 2007).

At a high level, the ideal adapter suggests that this type of selective adaptation is best thought of as a *computational* property of sensory systems, rather than a *mechanistic* property. Specifically, sensory systems at all levels must cope with changes in the statistical properties of the sensory world. Adaptation often leads to functional benefits, by (for instance) increasing discriminability of stimuli near the adapting stimulus or reducing metabolic costs by reducing firing rates for common stimuli (Kohn, 2007). Thus, the ideal adapter provides a framework for understanding and modeling adaptation *in general*, and opens the door to new computational-level models of perception that put the variability of the sensory world front and center. In doing so, it reveals unexpected connections between speech perception and low-level computational neuroscience.

Chapters 4-5: New approaches to studying speech perception

The final two chapters demonstrate two complementary future directions for work on speech perception that the ideal adapter enables. The first focuses on measuring listeners' knowledge about how talkers vary in terms of their accents. The second addresses how to quantify the variability in talkers' accents that actually occurs in the world using speech corpora in order to make precise predictions about listeners' adaptation behavior.

Prior experience guides adaptation

Even when adapting to a totally unfamiliar talker, listeners can benefit from experience with other talkers to narrow down the range of possible accents that are reasonable. This makes adaptation more efficient for talkers that fall within the expected range. But it will slow or even prevent adaptation to a talker whose cue distributions falls outside the range the listener expects. Chapter 4 presents a series of two studies in which listeners were exposed to cue distributions that are more or less similar to what English speakers typically produce. I found, as predicted, that listeners adapt rapidly to cue distributions that are in the normal range of English talkers, but adapt only incompletely to distributions that are outside the normal range. Thus, rapid adaptation to unfamiliar talkers is *constrained*, and these constraints are qualitatively consistent with the predictions of the ideal adapter. Again, these constraints are *adaptive*, in that they generally make adaptation more efficient, but can cause problems when we encounter talkers that fall outside the range we expect (e.g., in the extreme case of second language learning, Pajak, Fine, Kleinschmidt, & Jaeger, 2016).

Moreover, these constraints, as revealed by the patterns of adaptation (or lack thereof) across different cue distributions, are also *quantitatively* captured by a simple Bayesian belief updating model. This model assumes that all listeners start from the same prior expectations about the mean and variance of each category's cue distribution, and incrementally update them based on the distribution of cues that they experience in the

experiment. By treating these prior expectations as free parameters and fitting them to the patterns of adaptation to different cue distributions, it is possible to *infer* what listeners think an arbitrary, unfamiliar talker will sound like (in terms of their cue distributions), and how confident they are in these assumptions. Listeners' prior beliefs (as inferred by my model) align reasonably well with the distributions of the same cue measured from actual speech data, and furthermore effectively *predict* how listeners in a second study adapt to talkers who produce very different distributions of the same cues.

This is important for two reasons. First, it is an alternative to collecting and annotating lots of production data. Second, and more importantly, listeners' prior expectations are *subjective*, and may diverge from the actual variability across talkers in the world that can be measured from production data. By providing a formal framework for modeling adaptation, the ideal adapter allows these otherwise inaccessible, subjective beliefs to be gleaned from listeners' behavior, which is a critical step in understanding how listeners bring prior experience to bear in adapting to unfamiliar talkers.

How much do talkers actually vary, and how is this variability structured?

Chapter 2 shows that the ideal adapter has broad explanatory power for how listeners cope with talker variability, and many of its qualitative predictions about when listeners adapt, recognize, and generalize are borne out in the data. But to make these prediction more precise and quantitative, we need to know how cue distributions vary, across individual talkers and within and between groups of talkers. Chapter 5 presents a method for quantifying this variability based on existing speech corpora. It focuses on how talkers vary in terms of their voice onset times (VOT) for word initial stops (like /b/ and /p/ in "beach" vs. "peach") and in vowel formant frequencies.

The first basic finding is that, as commonly assumed (but not yet, to my knowledge, quantified), there is more variability across talkers in vowels than VOT. Based on this, that listeners should be more likely, a priori, to expect that two talkers will have different cue distributions for vowels than stop voicing, and thus less be more likely to generalize from one talker to another when considering VOT. This is, in fact, how listeners behave (Kraljic & Samuel, 2007).

Next, I found that socio-indexical variables (like gender, age, and regional dialect) are more *informative* about vowel distributions than about VOT. Moreover, these variables are highly *useful* for speech perception: using distributions of cues *conditioned* on, for instance, the gender or regional dialect of the current talker leads to substantial increases in correct vowel recognition. The ideal adapter predicts, based on this, that listeners should use information about, for instance, a talker's gender or dialect to guide vowel recognition. This prediction, too, is borne out in empirical data (Hay & Drager, 2010; gender, Johnson, Strand, & D'Imperio, 1999; dialect, Niedzielski, 1999).

Finally, I found that linguistic cue distributions are also informative about socio-indexical

variables themselves. A simple ideal observer model can recognize a talker's gender and dialect (based on their productions of vowels) and age (based on VOT) using the distributions of the relevant cues from other talkers. This suggests a potential unification of sociolinguistic and cognitive approaches to speech perception, wherein both linguistic and socio-indexical judgements are treated as inference under uncertainty, based on the same knowledge about how different types of talkers produce linguistic cue distributions.

Conclusion

The ideal adapter framework presents a solution to one of the oldest puzzles in research on human speech perception: how is it possible at all given the amount of variability between individual talkers? In the tradition of ideal observer or rational analysis approaches to cognition (Anderson, 1990; Marr, 1982), this framework lays out the computational problem of speech perception in a variable but structured world. This framework provides a unified perspective on a large and often apparently contradictory literature on how listeners cope with talker variability, leads naturally to implemented computational models that provide good descriptions of human speech perception behavior, and opens up new directions for research on speech perception and perception in general. My ongoing work is focused on exploring these directions and in understanding how the computational principles of the ideal adapter might be implemented in neural mechanism (both through computational modeling and through functional neuroimaging).

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544. doi:10.1121/1.1528172

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi:10.1016/j.cognition.2008.04.004

Clopper, C. G., Pisoni, D. B., & Jong, K. J. de. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, *118*(3), 1661. doi:10.1121/1.2000774

Creel, S. C., & Bregman, M. R. (2011). How Talker Identity Relates to Language Processing. *Language and Linguistics Compass*, *5*(5), 190–204. doi:10.1111/j.1749-818X.2011.00276.x

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: the role of talker variation in lexical access. *Cognition*, *106*(2), 633–64.

doi:10.1016/j.cognition.2007.03.013

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*(1), 99–109. doi:10.1016/0010-0285(73)90006-6

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–38.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–82. doi:10.1037/a0017196

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183. doi:10.1037/0278-7393.22.5.1166

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. doi:10.1515/ling.2010.027

Heald, S. L. M., & Nusbaum, H. C. (2015). Variability in vowel production within and between days. *PLoS ONE*, *10*(9). doi:10.1371/journal.pone.0136791

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. doi:10.1006/jpho.1999.0100

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. doi:10.1037/a0038695

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*, *23*(3), 678–691. doi:10.3758/s13423-015-0943-z

Kleinschmidt, D. F., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 599–604). Austin, TX: Talk; Cognitive Science Society.

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, *97*(5), 3155–64. doi:10.1152/jn.00086.2007

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–78. doi:10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–8.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. doi:10.1016/j.jml.2006.07.010

Labov, W., Ash, S., & Boberg, C. (2005). *The Atlas of North American English* (p. 318).

doi:10.1515/9783110206838

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–61.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt; Co., Inc.

Miller, J. L. (2001). Mapping from acoustic signal to phonetic category: Internal category structure, context effects and speeded categorisation. *Language and Cognitive Processes*, *16*(5-6), 683–690. doi:10.1080/01690960143000065

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing* (PhD thesis No. December). University of Iowa.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. doi:10.1177/0261927X99018001005

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–95. doi:10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–76.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning Additional Languages as Hierarchical Probabilistic Inference: Insights From First Language Processing. *Language Learning*, *e-pub ahea*. doi:10.1111/lang.12168

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–28.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(2), 539–55. doi:10.1037/a0034409

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651–666. doi:10.1037/0096-1523.23.3.651

Samuel, A. G. (2001). Knowing a Word Affects the Fundamental Perception of The Sounds Within it. *Psychological Science*, *12*(4), 348–351. doi:10.1111/1467-9280.00364

Vroomen, J., Linden, S. van, Gelder, B. de, & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–7. doi:10.1016/j.neuropsychologia.2006.01.031

Webster, M. A., Werner, J. S., & Field, D. J. (2005). Adaptation and the Phenomenology of

Perception. In C. Clifford & G. Rhodes (Eds.), *Fitting the mind to the world: Adaptation and after-effects in high-level vision (advances in visual cognition)* (Vol. 2, pp. 241–277). Oxford University Press.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. doi:10.1016/j.jml.2015.08.003