WITHIN CATEGORY PHONETIC VARIABILITY AFFECTS PERCEPTUAL UNCERTAINTY

Meghan Clayards, Richard N. Aslin, Michael K. Tanenhaus, Robert A. Jacobs

Dept. of Brain and Cognitive Sciences, University of Rochester
mclayards@bcs.rochester.edu, aslin@cvs.rochester.edu, mtan@bcs.rochester.edu, robbie@cvs.rochester.edu

ABSTRACT

We explored a mechanism for adjustments in the perceptual weighting of multiple probabilistic cues in speech. Subjects heard words that varied along a voice onset time (VOT) continuum (eg. "beach" to "peach") while performing a two alternative forced choice task (2AFC). For one group the VOT values that they heard came from distributions with wide variance (wide group) around the category prototype and for the other group they came from distributions with narrow variance (narrow group). The slope of the labeling response curve was shallower for the wide group indicating greater perceptual uncertainty. This suggests listeners are sensitive to the reliability of an acoustic cue when making category judgments and can rapidly adjust cue-weights in response to cue-reliability.

Keywords: speech perception, cue weighting, variability, Bayesian perception.

1. INTRODUCTION

The speech signal is made up of multiple probabilistic cues. The variability of these cues within speech categories has long been known [6][14] and may result from differences in vocal tract size and shape across speakers, speaking rate and style, as well as interactions with the preceding and following phonetic context. Even within a single speaker and speaking style, these phonetic cues show considerable variability across exemplars and are roughly normally distributed [8]. Variability of cues has traditionally been considered a problem for models of speech perception and word recognition because phonemes and words were viewed as having one prototypical target to which a given token was compared. In contrast, recent theories have argued that variability is in fact informative and utilized by comprehenders to make accurate judgments about lexical items [10] [11] as well as by language learners to infer relationships between speech cues and perceptual categories [7][8][9]. In this paper we further investigate the role of variability in guiding comprehenders' interpretation of acoustic cues in speech. In particular, we ask whether the degree of variability in a phonetic cue leads listeners to rapidly adjust their cue-weights to reflect the reliability of this cue.

A class of models which is increasingly being applied to perception in many domains and at multiple levels is Bayesian inference models. In these models, decisions made about perceptual information are guided by basic principles which are mathematically grounded [3][4]. One is to acknowledge that the world provides only probabilistic information which is inherently ambiguous at any given time. A second is that decisions should be made using all the available information. This includes prior knowledge as well as the current estimates available to the perceptual system. Crucially, each of these information sources is evaluated and weighted according to how useful it is. One major advantage of these models is that they are computationally explicit and implementable at the neuronal level [2]. A second important advantage for present purposes is that they provide straightforward predictions about how cues should be evaluated. In particular, the precision, or amount of certainty about the world that a particular cue provides is inversely proportional to the variance of that cue in the world. This has clear implications for theories of speech perception coping with inherent variability in the speech signal.

In testing these predictions we assume that the information available to the comprehender includes knowledge of the distribution of a cue in the world. This knowledge must be built up through experience with prior input which could in principle (and for many people does in practice) vary over the course of a lifetime. If

comprehenders evaluate a cue according to the variability in its distribution along a given acoustic dimension, they should be sensitive to changes in this distribution, even if they occur in the short term. Shifts in category boundary for voice onset times (VOT) have already been observed with short term exposure [1]. In this paper we explicitly manipulate the distribution of VOTs that comprehenders are exposed to while hearing words beginning with bilabial stops. Our prediction is that the degree of confidence that listeners place in VOT as a cue to voicing contrasts is proportional to the variance of the distributions to which they are exposed (i.e., high variance – low confidence). We will manipulate symmetrically the distributions that both the voiced and voiceless stops were drawn from, producing two different bimodal distributions over VOT (Fig.1).

We evaluated the confidence of comprehenders' judgments about a lexical contrast by 2 alternative forced choice (2AFC) labeling responses. If responses are generated by a Bayesian decision process which models the uncertainty of the underlying distribution, then the slope of the labeling response curve should become shallower with increasing uncertainty.

2. METHOD

Subjects were 20 monolingual native English speaking students from the University of Rochester with no known hearing problems. Each participated in two sessions on consecutive days, lasting approximately one hour.

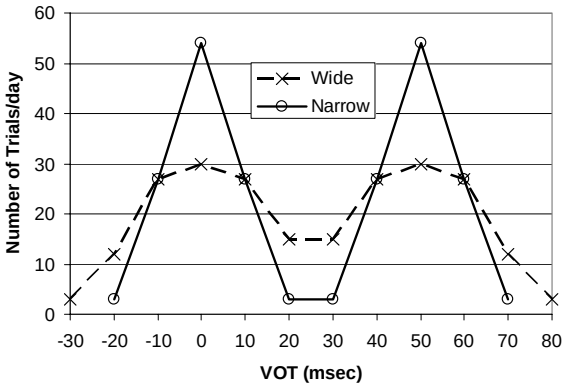
2.1.1. Stimuli

Stimuli were created using the Klattworks [12] interface to the Klatt 1980 synthesizer [5] and modeled after natural tokens. Three VOT continua were created who's endpoints corresponded to "peach"-"beach", "peak"-"beak", and "peas"-"bees" respectively. Negative VOT values were created by adding voicing before the stop burst, while positive VOT values were created by replacing 10msec of voicing after the stop burst with aspiration to create a continuum from -30 to 80 msec of VOT. Six filler items were also synthesized. These were "lace"-"race", "lake"-"rake" and "lei"-"ray".

2.1.2. Distribution

All subjects heard the same stimuli along the VOT continuum. Subjects were divided into two groups according to the *number of times* they heard each VOT token as shown in Figure 1. Half of the subjects heard stimuli taken from a bimodal distribution with variance of 8msec around each mean (Narrow group) and half heard stimuli taken from a bimodal distribution with variance of 12msec around each mean (Wide group). Both bimodal distributions had means centered at 0msec and 50msec with a hypothetical category boundary at 25msec (Fig. 1). The means and category boundary were chosen to match natural distributions measured by Lisker & Abrahmson [6] as well as categorization performance in piloting.

Figure 1: The distribution of tokens for the Wide and Narrow groups.



2.1.3. Procedure

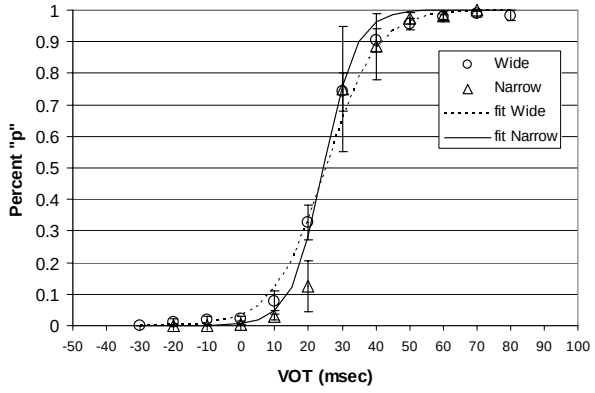
On each day subjects heard 456 trials (228 test and 228 filler). On each trial a word was presented over headphones. Subjects were instructed to choose the correct visual referent by clicking on a picture on the computer screen. The visual display consisted of four pictures: a pair of /p/-/b/ items and a pair of filler /l/-/r/ items. On half of the trials subjects heard one of the filler items. On the other half of the trials they heard a test item taken from one of the two distributions above. Order of trials was randomized. All tokens from the distribution were heard on each day.

3. RESULTS

Data from all critical trials on both days were analyzed together. Our key prediction was that the slope of the categorization function would be shallower for the Wide group than for the Narrow group. To assess this we performed a logistic regression on the categorization data. The data and fits for each group are seen in Fig. 2. For this analysis we considered the task to be two

alternative forced choice (trials where subjects clicked on filler items were discarded). A logistic regression was used in order to accurately model the binomial nature of the data. The model included VOT and condition (Wide vs. Narrow) as predictors as well as an interaction term. As is standard, the means of the mian effects were removed out of the interaction term to reduce colinearity in the model (this lead to VIFs <1.2, so that confounds due to colineratity are unlikely). As expected VOT was a significant predictor of response (coefficient's Wald Z = 28.15, p<.0001). Condition was not a significant predictor with colinearity removed (coefficient's Wald Z = -1.12, p=0.26). The interaction of VOT and condition was a highly significant predictor (coefficient's Wald Z = -9.03, p<.0001). The slope of the VOT effect was less steep for the Wide group than for the Narrow group. In other words, as predicted, comprehenders responding to a wide distribution exhibited more uncertainty about category membership than those responding to a narrow distribution.

Figure 2: Probability of /p/ responses and best fit logistic regression lines.



A mixed effects linear regression analysis performed on each side of the continuum found an effect of VOT on looks to the competitor when it

4. CONCLUSIONS

We found that the distribution of cues that comprehenders responded to influenced how certain they were about which word they were hearing. When responding to stimuli from a wide distribution their 2AFC responses were less categorical (shallower labeling slopes) than when responding to stimuli from a narrow distribution.

This evidence provides support for the theory that when listening to speech, comprehenders are continually assessing the quality of the cues in the speech stream. This gives us a straightforward mechanism for understanding how multiple speech cues are weighted, how second language learners re-tune their perceptual system to accommodate new patterns of cue distribution, as well as how infants learn which cues to attend to in their native language. Importantly, it brings mechanisms known to guide perception in other domains into the domain of speech perception.

5. REFERENCES

[1] Clarke, C.M., Luce, P.A. 2005. Perceptual adaptation to speaker characteristics. In: C.T. McLennan, P.A. Luce, G. Mauner, J. Charles-Luce (eds.) *U at Buffalo Working Papers on Language and Perception* 2, 362-366.

[2] Denever, S., Latham, P.E., Pouget, A. 1999. Reading population codes; a neural implementation of ideal observers. *Nature Neuroscience* 2(8), 740-745.

[3] Ernst, M.O., Banks, M.S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429-433.

[4] Ernst, M.O., Bulthoff, H.H. 2004. Merging the senses into a robust percept. *TRENDS in Cognitive Sciences* 8(4), 162-169.

[5] Klatt, D. 1980. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* 67, 971-995.

[6] Lisker, L., Abrahmson, A. 1964. Cross language study of voicing in initial stops. *Word* 20, 384-482.

[7] Maye, J., Weker, J., Gerken, L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, B101-B11.

[8] Maye, J., Gerken, L. 2000. Learning phonemes without minimal pairs. In: S. C. Howell, S. A. Fish, T. Keith-Lucas (eds), *Proc. of 24th BU Conference on Language Development*. Somerville, MA: Cascadilla Press, 522–533.

[9] Maye, J., Weiss, D., Aslin, R.N. In Press. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science.*

[10] McMurray, B., Tanenhaus, M., Aslin, R. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86(2), B33-B42.

[11] McMurray, B., Tanenhaus, M., Aslin, R.N. 2006, June. Garden-path phenomena in spoken word recognition: Gradient sensitivity to continuous acoustic detail facilitates ambiguity resolution. Proceedings *151st Acoust. Soc. Am.* Providence, RI. 119, Issue 5, p. 3443

[12] McMurray, B. In preparation. KlattWorks: A [somewhat] new systematic approach to formant-based speech synthesis for empirical research.

[13] Newman, R., Clouse, S.A., Burnham, J.L. 2001. The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am.* 109(3), 1181-1196.

[14] Peterson, G. E., Barney, H. L. 1952. Control methods used in a study of vowels. J. Acoust. Soc. Am. 24, 175–184.