INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION: EVIDENCE FROM VISUAL ANALOGUE SCALING AND EYE-TRACKING

Eun Jong Kong & Jan Edwards

Waisman Center, University of Wisconsin-Madison, Wisconsin, USA
ekong@wisc.edu; jedwards2@wisc.edu

ABSTRACT

This study investigated whether there were individual differences in within-category sensitivity to the voicing contrast and, if so, whether these differences were related to listeners' differential sensitivity to different acoustic cues. To do this, we conducted two speech perception experiments with English-speaking adults: visual analogue scaling (VAS) and anticipatory eye movement (AEM). Stimuli were a 30-item /ta/ to /da/ continuum, which systematically varied both VOT and *f0*. We found evidence of gradient sensitivity to within-category fine phonetic detail for both tasks. Consistent with previous research, we also found that listeners were more sensitive to changes in VOT than to changes in *f0* for both tasks. Listeners who had a gradient response pattern on the VAS task showed evidence of sensitivity to *f0* on the AEM task, while listeners who had a categorical response pattern on the VAS task did not. This result suggests that there are individual differences in responses to subphonemic detail and that these differences may be systematically related to sensitivity to different acoustic cues.

Keywords: speech perception, categorical perception, stop voicing contrast, eye-tracking

1. INTRODUCTION

One of the foundational results of our understanding of human speech perception is categorical perception [4]. Categorical perception of speech describes the finding that listeners seemingly cannot discriminate between items that they identify as being in the same phoneme category. Between-category discrimination is superior to within-category discrimination, even if the acoustic differences between the stimuli are identical in both cases. Researchers originally interpreted categorical perception as evidence that listeners discard subphonemic acoustic variation and attend only to higher-level categorical representations of speech sounds in perception.

However, over approximately the last two decades, a large body of research has shown that listeners also pay attention to lower-level phonetic detail in processing speech sounds. For example, listeners pay attention to lexically irrelevant information, such as differences between talkers and perform better in a single-speaker condition as compared to a multiple-speaker condition across a variety of experimental paradigms [3].

Furthermore, it has been shown that categorical perception of speech is specific to a particular set of experimental paradigms [8], while studies using different paradigms have observed listener sensitivity to fine phonetic detail. For example, Munson and colleagues used Visual Analog Scaling (VAS) to show that naïve listeners are able to discern fine-grained differences in children's productions of different consonants [7]. Furthermore, listener judgments for different consonant contrasts were well correlated with the acoustic characteristics that differentiated the two members of the contrast. On-line measures of speech perception such as eye-tracking have also provided evidence that listeners are sensitive to subtle phonetic details. McMurray et al. found that for adult listeners, the proportion of looks to a pictured object was sensitive to changes in within-category VOT values of the object name [5].

Because early studies of speech perception assumed that listeners did not attend to subphonemic acoustic variation, little or no attention focused on individual differences in listeners' performance. Even today, most studies of speech perception continue to report only group results. However, as it has become clear that fine-grained phonetic detail is utilized in speech processing, some researchers have investigated whether there are individual differences in how closely listeners attend to subphonemic detail [9]. For example, Zhao [9] found that listeners with better frequency discrimination performed better on a statistical learning task than listeners with poorer frequency discrimination. In this paper, we examine whether there are individual differences

in within-category sensitivity to the voicing contrast and, if so, whether these differences are related to listeners' differential sensitivity to different acoustic cues.

2. METHODS

2.1. Stimuli

The stimuli were synthetic CV syllables that were constructed to make a continuum from /ta/ to /da/. They were synthesized using words (*tot* and *dot*) produced by a Wisconsin adult male speaker. We selected one token of /da/ and we systematically varied both VOT and *f0* to create the continuum. VOT values were manipulated by excising a portion of the burst release/aspiration from /ta/ and pasting it before the voicing onset of the /da/ token. Six different log-scale steps of VOT were included: 9ms (original VOT of /da/), 13ms, 19ms, 28ms, 40ms, and 59ms. At each VOT step, we replaced the original *f0* value during the vowel with one of five different *f0* values: 98Hz, 106Hz, 114Hz, 122Hz, and 130Hz. A total of 30 different stimuli (6-step VOT × 5-step *f0*) were created.

2.2. Experimental tasks and procedure

There were two experimental tasks: 1) visual analogue scaling (VAS) and 2) anticipatory eye movement (AEM). For each task, all listeners heard the same 30 stimuli items three times in random order. The two tasks were counter-balanced, so approximately 50% of the participants did the AEM experiment first and the VAS experiment second, while the other 50% had the reverse order.

In a VAS rating task, individuals are asked to scale a psychophysical parameter by indicating their percept on an idealized visual display. In our VAS task, an arrow was displayed on the computer monitor immediately after each stimulus was played. One end of the arrow was labeled as the 'd' sound and the other end of the arrow was labeled as the 't' sound. Listeners were instructed to click anywhere on the arrow, based on their judgment of how close the stimulus was to either /da/ or /ta/.

In the AEM task, we used the SMART-T program to implement an anticipatory eye movement paradigm using the Tobii 2150 eye-tracker [6]. Listeners were conditioned to make anticipatory looks to either the left or the right side of a Y-shaped pipe based on whether the sound they heard was more similar to /t/ or /d/. Six training trials preceded the experimental trials. For all trials, when an auditory stimulus item was played, a picture (representing either *doggie* or *taco*, words which contain a syllable-initial /da/ or /ta/) appeared at the bottom of the pipe and moved slowly through the pipe. The picture would exit on either the left or right side of the pipe. The /da/ stimuli were paired with the picture of doggie and consistently appeared on one side of the Y-shaped pipe, while the /ta/ stimuli were paired with the picture of taco and consistently appeared on the other side of the pipe. Ambiguous stimulus items were played twice and appeared once on each side of the pipe. The Y-shaped pipe was transparent at the beginning of the training trials so participants could see the path of the moving picture. The pipe gradually became opaque during the 6 training trials. During the experimental trials, the pipe was opaque and the participants had to anticipate on which side the picture would appear based on whether they perceived the auditory stimulus to be /d/ or /t/. The participants were given no other instructions beyond "look at the computer screen."

2.3. Subjects

The participants were 24 English-speaking adults with no reported history of speech, language, or hearing problems. All participants were female undergraduate students at the University of Wisconsin-Madison and received course credit for their participation.

2.4. Analysis

For the VAS task, we transformed the pixels of the click locations along the arrow into generalized logit values. Lower logit values indicate more /d/-like tokens and higher logit values indicate more /t/-like tokens. We used these logit-transformed values as the dependent variable and VOT and *f0* as the independent variables in order to examine the influence of these acoustic measures on perception of the /t/-/d/ contrast. We also examined individual differences in gradiency of response by comparing histograms of click locations for each individual listener.

For the AEM task, we grouped observations into a series of temporal bins (50ms bins) within each listener's trials of the same stimuli and calculated the empirical logit of looks to /d/ in each bin [1]. We analyzed listeners' responses to the first two presentations of each stimulus. We constructed a mixed-effects model with the estimated logit value of looking to /d/ as the

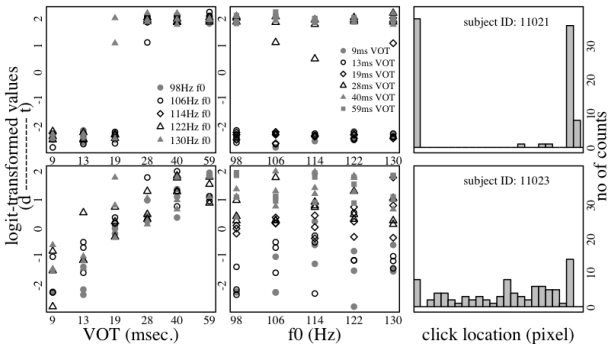
dependent measure and time (temporal bins), VOT, and *f0* as the independent variables.

3. RESULTS

3.1. Visual analogue scaling

Fig. 1 shows the results of the VAS task for two representative subjects. In the left and middle panels, the logit-transformed values are plotted as a function of VOT and *f0*, respectively. It can be observed that perception of the contrast between /t/ and /d/ was strongly influenced by VOT, but not by *f0*. The two left panels of Fig. 1 show a strong relationship between VOT and the logit-transformed values; stimuli with longer VOT were consistently rated as more /t/-like. In contrast, the middle two panels show that stimuli with higher *f0* were not consistently rated as more /t/-like.

Figure 1: Two individual subjects' response patterns for the VAS task: logit-transformed values as a function of VOT (left) and of *f0* (middle), and histograms of click locations (right).



Importantly, there were clear individual differences in gradiency of response, as shown by the histograms of click locations in the rightmost panels of Fig. 1. Some listeners, such as the subject in the top right panel, judged the stimuli categorically, by choosing responses mostly at the two endpoints of /t/ and /d/. By contrast, other listeners, such as the subject in the bottom right panel, judged the stimuli in a much more gradient manner, by choosing responses across the entire VAS scale. About 50% of the participants were distributed approximately equally between these two extreme groups, with 6 in the categorical group and 7 in the gradient group. The remaining 11 participants did not fit clearly into either group based on click location histograms.

3.2. Anticipatory eye movement

Fig. 2 shows the change in the estimated logit values of looking to /d/ over time as a function of different VOT and *f0* conditions. This figure plots the estimated slopes from the mixed-effects model. The results for the AEM task were mostly similar to the results for the VAS task: listeners were much more sensitive to changes in VOT than to changes in *f0*. Across the different *f0* conditions (bottom panels), stimuli with shorter VOT values consistently resulted in more looks to /d/ in each temporal bin than stimuli with longer VOT values. However, the AEM task provides a more sensitive online measure of perception than the VAS task. With this measure, we identified one condition in which listeners were also sensitive to changes in *f0*. In the 19ms VOT condition (middle top panel), listeners looked more to /d/ over time as *f0* decreased. That is, listeners attended to the *f0* cue to voicing when the VOT cue was ambiguous.

Figure 2: Logit values of looking to /d/ over time (temporal bins) for all speakers, separated by VOT conditions (top) and by *f0* conditions (bottom).

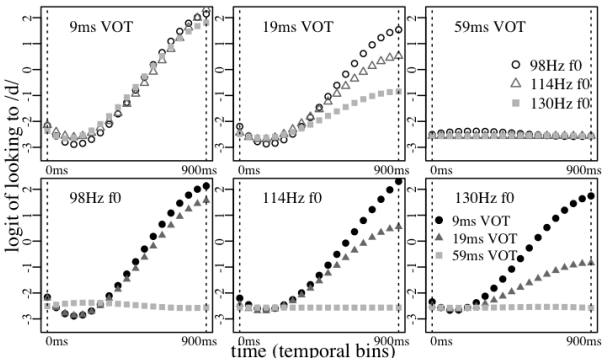
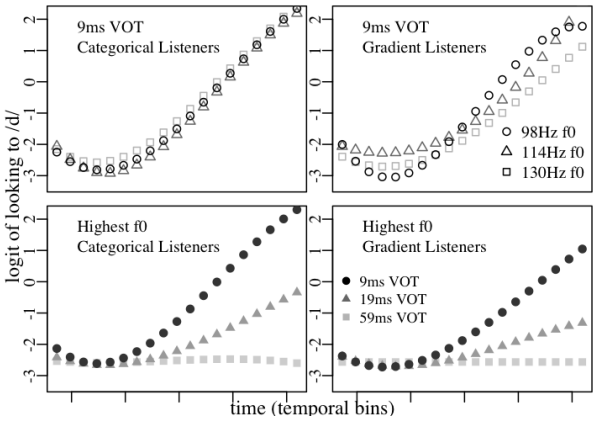


Fig. 3 shows the trajectories of the logit values of looking to /d/ separately for the categorical and gradient listener groups, as identified by the VAS results. In the 9ms VOT condition, the categorical group (top left panel) shows no sensitivity to *f0* with the trajectories from the three different *f0* conditions lying virtually on top of each other. For the gradient group (top right panel), the slopes of the trajectories indicate an influence of *f0*. The slopes became shallower, indicating fewer looks to /d/ over time, when the *f0* cue was in conflict with the percept of voicing. Similar results are illustrated in the bottom two panels, where the high *f0* is consistent with a voiceless stop. In both the 9ms and 19ms VOT conditions, the slopes of the categorical group were influenced only by VOT. However, the slopes of the gradient group for these two VOT conditions were shallower than those of the categorical group because of the conflicting cue provided by the high *f0*.

Figure 3: Logit values of looking to /d/ over time (temporal bins) as a function of VOT and *f0*, separated by listener groups: categorical (left) and gradient (right).



4. DISCUSSION & CONCLUSION

This study was designed to examine whether there were individual differences in how sensitive listeners were to within-category differences for the stop voicing contrast. As in many previous studies, we found that most listeners were able to perceive the contrast between /d/ and /t/ gradiently rather than categorically, given an appropriate task. More importantly, we found that there were individual differences in the perception of within-category differences. About 25% of our listeners perceived the stimuli categorically, even on a VAS task that was designed to encourage gradient perception. Another 25% of our listeners had a gradient pattern of response on the VAS task. We found that only this gradient listener group was sensitive to changes in *f0* on the AEM task.

Further research is needed to understand these individual differences. First, it needs to be confirmed that these individual differences in speech perception are consistent: do we find that subjects in the categorical and gradient groups remain in these same groups across experimental sessions? If this turns out to be the case, then we need to investigate whether there are subject-level characteristics that are consistently associated with these individual differences. In this study, we measured forward and backward digit span of all participants, as backward digit span is considered to be a valid measure of working memory capacity [2]. While the mean backward digit span was greater for the gradient listener group than for the categorical listener group, this difference was not significant because of the small number of subjects in the two subgroups and the fact that the digit spans of the two groups overlapped almost completely. In principle, we might predict that individuals with greater working memory capacity would be more able to hold multiple acoustic cues in memory. Thus, a larger *n* and additional measures of auditory working memory would be useful in future research on these individual differences in speech perception.

To conclude, this study is important in two respects. First, it illustrates how an online measure such as eye-tracking can provide information about aspects of speech perception that cannot be detected by offline measures. Second, it provides evidence that there are differences in how sensitive individuals are to fine phonetic detail and suggests that these differences may be related to different patterns of attention to acoustic cues.

5. ACKNOWLEDGMENTS

Work supported by NSF grant BCS0729040.

6. REFERENCES

[1] Barr, D.J. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59, 457-474.

[2] Daneman, M., Merickle, P.M. 1996. Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review* 3, 422-433.

[3] Goldinger, S.D., Pisoni, D.B., Logan, J.S. 1991. On the nature of talker variability effects in recall of spoken word lists. *J. Exp. Psych. [Learn. Mem. Cog.]* 17, 152-162.

[4] Liberman, A.M., Harris, K.S., Kinney, J.A., Lane, H. 1961. The discrimination of relative onset-time of components of certain speech and non-speech patterns. *J. Exp. Psychol.* 61, 379.

[5] McMurray, B., Tanenhaus, M., Aslin, R., Spivey, M. 2003. Probabilistic constraint satisfaction at the lexical/phonetic interface. *J. Psycholing. Res.* 32, 77-97.

[6] Mohinish, S. Wen, J., White, K., Aslin, R. Accepted. SMART-T: A system for novel fully automated anticipatory eye-tracking paradigm. *Behav. Res. Meth*.

[7] Munson, B., Edwards, J., Schellinger, S., Beckman, M.E., Meyer, M. 2010. Deconstructing phonetic transcription. *Clin. Linguist. Phon.* 24, 245-260

[8] Schouten, B., Gerrits, E., van Hessen, A. 2003. The end of categorical perception as we know it. *Speech Commun.* 41(1), 71-80.

[9] Zhao, Y. 2009. *Statistical Inference in the Learning of Novel Phonetic Categories*. Ph.D. Thesis, Stanford University.