Similarity Between Vowels Influences Response Execution in Word Identification

Jason D. Zevin¹, Thomas A. Farmer¹²

¹Sackler Institute for Developmental Psychobiology, Weill-Cornell Medical College, New York
²Department of Psychology, Cornell University, Ithaca, NY
jdz2001@med.cornell.edu, taf22@cornell.edu

Abstract
Speech categories vary in their similarity to one another, with consequences for phenomena ranging from patterns of confusability under noisy listening conditions to phonological processes described as operating on relatively abstract representations. Studies of inter-category similarity are typically conducted by studying error rates in noise, or by using metalinguistic similarity judgments. Here, we exploited the continuous and non-ballistic properties of computer mouse movements to investigate, in a fine-grained manner, the inter-category similarity of speech sounds. Participants listened to recordings of naturally produced words (“pin,” “pen,” and “pan”) and used a computer mouse to select the matching stimulus from an array of two pictures. The same participants also performed a dissimilarity judgment task. Both tasks revealed evidence for graded effects of inter-category similarity, albeit in the context of strongly categorical classification. Multidimensional scaling of mean dissimilarity judgments revealed a very tight distribution of stimuli from the same category in perceptual space and, at a larger scale, differences between pairs of categories in their similarity to one another. Although accuracy in the word identification task was nearly perfect, arm-movement dynamics revealed attraction toward the alternative response, scaled in magnitude to the perceived similarity of the two categories.

Index Terms: human speech perception, graded category similarity

1. Introduction
Speech processing has both discrete and continuous attributes. On the one hand, the categories of a language’s phonetic inventory clearly shape the perception of speech sounds by speakers of that language [1], such that many different physical instantiations can be treated as roughly “the same” for the purposes of linguistic contrast. On the other hand, there is a wealth of evidence that intra-category variability is not “discarded” as described in early accounts of categorical perception. Rather, within-category differences in similarity clearly play a role in phonetic categorization [2], [3], and lexical processing [4],[5],[6]. In short, how “typical” a stimulus is as a member of its category can drive subtle effects on performance in a wide range of tasks.

The difference between tasks in which performance reflects strongly categorical responses and tasks in which graded stimulus properties play a larger role can be fairly characterized as one of granularity: When the outcomes of binary decisions are used to describe the data, responses appear more categorical than when response latency [4], lexical priming [5], or eye-movements [6] are examined.

Interestingly, this type of fine-grained measure has rarely been applied to studies of similarity between categories, despite the importance of inter-category similarity to a wide range of disciplines from automated speech recognition [7] to formal phonology [8] to second-language speech perception [9]. Miller and Nicely’s classic study of confusability of speech sounds [10] employed a free identification paradigm, in which subjects were asked to identify nonsense syllables. In low-noise conditions, few errors occurred, but by adding enough noise to the stimuli to produce errors, they were able to observe orderly, non-random patterns of errors revealing the relative perceptual similarity of speech categories.

Although this approach is clearly productive [11], and obviously relevant to how speech communication is accomplished outside the sound-proofed rooms typical of the speech lab, it depends on pushing participants to generate errors, leaving open the possibility that it is telling us more about which stimulus properties are most susceptible to occlusion by white noise than which properties are most important for distinguishing sounds from one another under relatively clear listening conditions. An alternative approach has been to ask participants directly how similar they perceive stimuli to be, and to use these pair-wise similarity judgments, combined with multidimensional scaling analyses, to "map" the "perceptual space" [12],[13]. This approach has the advantage of permitting analysis of error-free performance, and provides a rich description of the similarity of multiple stimulus classes that can be used to compare groups that differ in language experience [14]. The primary data in these studies, however, comes from metalinguistic judgments of stimulus similarity that may not reflect the processes that subserve speech perception under more ecological task goals.

Applying finer-grained techniques to studies of inter-category perceptual similarity should, in principle, permit a richer description of how speech sounds are classified without the drawbacks of introducing noise to induce errors or requiring metalinguistic judgments. For example, data from an eye-tracking study using a four-alternative choice paradigm have revealed graded patterns of similarity for sets of English speech sounds under conditions that produced essentially error-free performance in the choice data [15]. In that study, we found that although categorization responses were perfectly consistent, patterns of eye-movements to distract stimuli before response execution were not randomly distributed, but depended on the phonetic similarity of the target and distractor.

A newer technique, mouse-tracking [16] (see also [17],[18]) permits us to observe the influence of confusability on the dynamics of response execution itself. Although individual saccadic eye movements can occasionally show some curvature [19], individual movements of the arm and hand can show quite dramatic curvature [20] which can be interpreted as the dynamic blending of two mutually exclusive motor commands [21]. Additionally, whereas eye-movement data allow for approximately 2-3 data points (saccades) per second, “mouse-tracking” yields somewhere between 30 and 60 data points per second. In light of the ability to record...
many data points per second, along with their ability to curve mid-flight in response to continuous dynamic competition between target and distractor. Mouse movements have the unique ability to illuminate the spatio-temporal dynamics of the categorization process itself, not only the product of it.

In this study, we presented native English speakers with a three-way vowel contrast in order to examine the influence of perceptual similarity on categorization, using arm-movements as a dependent measure. We predict that the degree of competition revealed by arm movements toward distractor stimuli will be directly related to the perceived similarity of pair-wise lexical contrasts, as measured in metalinguistic dissimilarity judgments.

2. Method

2.1. Participants

Ten right-handed native-English speaking adults with normal hearing (M=26.3 years, SD=5.54) participated in this study.

2.2. Materials

The stimuli were presented using Macromedia Director MX, and mouse movements were recorded at an average sampling rate of 40 Hz. The display resolution was set to 1024 x 768. Spoken target stimuli (pin, pen, and pan) were digitized using a SONY minidisc recorder, with each stimulus saved in a monaural 44,100 Hz WAV file with 10ms before and after the stimulus word using Praat, with 10ms rise/fall time to prevent back-ground noise from mimicking a consonantal burst. For each word, five different recordings were selected so that for each stimulus type, variation in fundamental frequency, harmonicity, word duration, and overall amplitude were matched in order to minimize the influence of the idiosyncrasies of any particular recording on performance. The same stimuli were used in both tasks.

For the mouse-tracking portion of the study, images representing each critical item were prepared from public domain images found on the internet, resized to 2” (in the longest dimension) at a resolution of 72 dpi in Photoshop. On any given trial, one of the three target words were heard while viewing a scene that contained both a picture of the item denoted by the target word and a picture of one of the two remaining items (the distractor). One object appeared in the top left-hand corner of the display and the other appeared in the right-hand corner (with the center of the image occurring at approximately 1.15” from its respective side of the display and 1.5” from the top). Additionally, at the bottom center of the display, a “Start” box approximately 1.5” x .33” appeared, along with a vertical transparent rectangle (.5” X 1”) centered directly above the start box.

2.3. Procedure

Participants were asked to make themselves comfortable in front of the computer screen, adjusting the mouse to a location on the right-hand side that suited them. All participants first completed the mouse-tracking portion of the study, followed by the dissimilarity judgment task.

2.3.1. Mouse-Tracking Task

At the onset of each trial, participants were presented with a black screen. After a 500ms delay, the target and the distractor appeared (at the top-left or top-right), each surrounded by a 2” x 2” light grey square. After another 500ms delay (used to provide a brief preview of the objects appearing in each location), a small box appeared in between the target and the distractor, containing the instructions “Click here to begin the next trial.” Immediately after clicking the initiation box, a “Start” button appeared at the bottom center of the display. Participants then clicked the “Start” button and moved the mouse up through the transparent rectangle. When the mouse entered the rectangle, the sound file containing the target word was cued, signaling to participants the correct object on which to click. After clicking on either one of the two objects in the display, the display disappeared, thus ending the trial.

Stimulus presentation was arranged for each of the three possible target-distractor pairings (pin/pen; pin/pan; pen/pan), so that for each participant, each speech stimulus type occurred as both the correct and incorrect response, and with the corresponding images occurring in both possible locations. This set-up produced 60 experimental trials in a completely counter-balanced within-subject design: 3 (target-distractor pairing) X 2 (target 1 versus target 2) X 2 (target-left versus target-right) X 5 (spoken target-word token). Presentation order was randomized for each participant.

2.3.1. Dissimilarity Judgment Task

At the beginning of each trial, a start-box appeared at the bottom center of the screen, and four boxes appeared at the top of the screen. Each box contained a number between 1 and 4, signifying the degree to which the two stimuli to be heard sounded dissimilar, and was accompanied by a verbal description (1=Exactly the Same; 2=Somewhat Similar; 3=Somewhat Different; 4=Completely Different). Upon clicking the start-box, it disappeared and the sound-file for the first target word began after a 500-ms delay, followed by the second target word. The stimulus onset asynchrony was 750ms, producing an approximately 500ms inter-stimulus interval. After the onset of the second target word, participants were able to click one of the four boxes to denote their dissimilarity rating, thus ending the trial.

3. Results

3.1. Mouse-Tracking

3.1.1. Data Screening and Coding

Mouse movements were recorded for the entire duration of each trial, starting with the trial-initiating click on the Start button. Due to of the large number of possible trajectory shapes, the x,y coordinates for each trajectory were plotted in order to detect the presence of any aberrant movements. A trajectory was considered valid and submitted to further analysis if a) the correct target referent was selected, and b) there were no heavily sporadic movements (such as a loop) embedded within the trajectory. This screening procedure resulted in the exclusion of 10 trials (all of which were trials where the distractor object was clicked), accounting for less than 2% of all experimental trials.

Given the large potential for trial-by-trial variability in trajectory-initiation time and trial duration, all analyzable trajectories were time-normalized to 101 time-steps by a procedure originally described in [16]. All trajectories were aligned so that their first observation point corresponded to (0, 0). Then, across 101 normalized time-steps, the corresponding x,y coordinates were computed using linear interpolation.

3.1.2. Trajectory Divergence Analyses

Overall target-selection accuracy was high across each condition, although subtle by-pair differences emerged.
Participants were perfect for pin/pan pairs, whereas pin/pen elicited 2 of the ten errors, and pen/pan pairings elicited the remaining 8 errors. Due to the high accuracy rates, inferential statistics on accuracy were not possible. These rates mirror the general trend of the trajectory-divergence data, as illustrated in Figure 1. The amount of curvature toward the distractor object (left-ward movement for right-located targets, and vice versa) was largest for the pen/pan pairings, smallest for the pin/pan pairings, and intermediate for pin/pen. In order to produce a by-condition index of competition strength for each vowel contrast, across time, all left-ward trajectories were flipped, making them right-branching, by flipping the sign associated with the x-coordinates. The flipped (originally left) and original right-branching trajectories were then averaged, per each possible target-distractor pairing.

![Normalized trajectories indexing average pair-wise competition strength for each possible vowel contrast](image)

Figure 1: Normalized trajectories indexing average pair-wise competition strength for each possible vowel contrast

To determine whether any pair-wise divergences observed across the trajectories from each vowel contrast were statistically reliable, we first conducted a series of paired-sample t-tests. As per [16], we viewed the x-coordinates of the trajectories to be a strong index of spatial attraction toward either the competitor or the target object. As such, the t-tests were conducted across the x-coordinates of each possible vowel contrast, separately, at each of the 101 time-steps. In order to avoid the increased probability of a Type-1 error associated with multiple comparisons, and in keeping with Bootstrap simulations of such multiple t-tests on mouse-trajectories [17], an observed divergence was not considered significant unless the coordinates between the two vowel-contrast conditions diverged significantly ($p < .05$) for at least eight consecutive time-steps.

For the x-coordinate comparisons between the pin/pen and the pin/pan contrasts, a significant divergence occurred from time-steps 77-93 (all $t$'s $> 2.32$, $p$'s $.05$). At each point-wise comparison, x-coordinates were further right for the pin/pan comparison, indicating that it elicited less attraction to the competitor (left-placed) object than the pin/pen comparison, across time. Additionally, the x-coordinates elicited by the pen/pan contrast were significantly closer to the competitor (left-ward) object than they were for the pin/pan vowel contrast from time-steps 57-86 (all $t$'s $> 2.37$, $p$'s $.05$), or, for the pin/pen contrast (marginally significant divergence from time-steps 57-86, all $p$'s $.098$).

### 3.1.3. Analyses of Maximum Deviation Values

In order to extract a trial-by-trial index of the magnitude of spatial attraction toward the incorrect (competitor) object, we calculated the maximum deviation value for each trial. First, we imposed a straight line from each trajectory’s starting point to its endpoint, and then extracted the single largest point-wise Euclidean coordinate divergence from the straight line. A 3 (target-distractor Pair) X 2 (Location: Right versus Left) repeated-measures ANOVA on the MD values revealed no main effect of Location, and no Location X Pair interaction, $F$s $< 1.03$, indicating that there was no differential response to a target when it occurred on the left versus the right-side. Consistent with the analyses reported above, however, there was a main effect of target-distractor pairing, $F$(2,18)=10.00, $p<.0005$, $MSE=117.2$. Corroborating the t-test divergence analyses above, a significant difference was elicited in the magnitude of attraction toward the competitor object between the pin/pen ($M=234.35$, $SD=20.5$) and pin/pan ($M=214.1$, $SD=21.52$) contrasts, $t$(9)=2.37, $p<.05$, between pin/pan and pen/pan ($M=262.31$, $SD=43.05$), $t$(9)=4.61, $p<.01$, and a marginally significant difference between pin/pen and pen/pan, $t$(9)=2.16, $p=0.05$.

### 3.2. Dissimilarity Judgment Data

Rated dissimilarity for each contrast was analyzed directly in one-way ANOVAs, with subjects (F1) or items (F2) as the random variables, and pair-type (pin/pen, pen/pan, and pin/pan) as the IV, revealing a significant effect of pair-type, $F$(2,18)=21.02, $F$(2,72)=93.78, both $p$'s $.001$. The greatest dissimilarity was observed between pin and pan, and the least between pen and pan. Planned pair-wise comparisons revealed reliable differences between the pen/pan pair ($Dissim=3.00$) and the pen/pin pair ($Dissim=3.38$), $t$(9)=3.91, $t$(48)=9.25, $p<.005$, and between pin/pen and pin/pan ($Dissim=3.61$), $t$(9)=2.45, $t$(48)=3.55, $p<.05$.

![Perceptual distance between all stimuli for all subjects.](image)

Figure 2: Perceptual distance between all stimuli for all subjects.

Average dissimilarity judgments for each subject, for all pairs of items, were then arranged into distance matrices with zeroes on the identity diagonal, and a multidimensional scaling solution was identified using isometric MDS with two dimensions. Figure 2 shows the two-dimensional MDS solution based on the dissimilarity scores from all subjects for all items, along with the mean perceptual distance between categories. Of interest in this plot is the different degree of scatter within categories for the stimulus classes. A one-way ANOVA revealed a main effect of stimulus type on within-category distance, $F$(2,18)=4.91, $p<.05$. The effect is driven by the high value for “pen,” (1.02) relative to “pan” (0.60), $t$(9)=2.84, $p<.05$ and “pin” (0.64), $t$(9)=1.98, $p=.08$. “Pin” and “pan” did not differ significantly from one another, $t<1$.

### 4. Discussion

We found evidence for an influence of graded similarity between speech sound categories on both metalinguistic dissimilarity judgments and an on-line word identification task. At the same time, it is critical to note that performance in both tasks was largely driven by stimulus category. The multidimensional scaling analysis of the dissimilarity judgment data revealed relatively tight clustering of stimuli...
from the same category in perceptual space. At a larger scale, however, orderly differences in similarity between categories were also observed.

**Figure 3:** Acoustic measures of the steady-state vowel portions of the stimuli.

Graded differences in height influenced the perceptual similarity space, such that the one non-adjacent pair ("pin" and "pan") were rated as most dissimilar to one another. For adjacent pairs, there was a large difference between the higher ("pin" and "pen") and the lower pair ("pen" and "pan"), with the lower pair rated as much more similar overall. These findings are consistent with earlier studies applying MDS to the perceptual similarity of vowels [13].

Arm movement data from the decision task were in line with the perceptual similarity data, with the greatest competition observed in the contrast judged to be most similar. It is possible that the differential role of duration in the pair-wise contrasts may contribute to the observed influence of similarity on temporal response dynamics. All three stimuli were monophthongal front vowels differing from one another on the phonetic dimension of "height," which is typically described as corresponding to the first formant (e.g. [22]). As shown in Figure 3, however, there are clear spectral cues to height for contrasts between /I/ and the other two types, the values for /E/ and /ae/ are overlapping on F1. Conversely, duration is contrastive for /ae/ and /E/, whereas /I/ and /E/ have overlapping distributions. Future studies will use synthetic stimuli contrasting on either F1 or duration to tease apart the influence of each to online processing.

The correspondence between the perceptual similarity space and the strength of attraction to distracting stimuli suggests that despite the clear difference in the ecological validity of the tasks, they are influenced in a similar way by graded aspects of inter-category similarity.

**5. Acknowledgments**

This research was supported by NIH grant R01-DC007694 to JZ and a Delores Z. Liebmann Fellowship to TF. We thank Bruce McCandliss for input on the experimental design.

**6. References**


