AD HOC PHONETIC CATEGORIZATION AND PREDICTION

Ryan Rhodes[1], Chao Han[1], Arild Hestvik[1]

[1]University of Delaware
robot@udel.edu, hanchao@udel.edu, hestvik@udel.edu

ABSTRACT

We conducted two experiments using the 'varying standards' oddball mismatch negativity (MMN) paradigm to test whether the auditory cortex can create and use ad hoc phonetic representations to make predictions about a stream of varying speech sounds. Experiment 1 used stimuli from a /dæ/-/tæ/ continuum, comparing 'high-T' standards (average VOT: 80ms) and 'low-T' standards (average VOT: 65ms) to a 15ms deviant. Only a main MMN effect was observed, suggesting that the brain failed to generate an ad hoc phonetic representation.

To test this, we shifted the VOTs of all stimuli up by 35ms so that all stimuli (standards and deviant) were in the same category (/t/) – requiring ad hoc categorization to discriminate standard from deviant. We again find an MMN but no difference between 'high' and 'low' conditions. We conclude that the auditory cortex can generate ad hoc phonetic representations, but these representations do not retain detailed phonetic information.

Keywords: MMN, speech perception, VOT

1. INTRODUCTION

Speech sounds must be modeled by the mind at several levels of representation. These range from the acoustic to the phonetic to the phonological. The ultimate linguistic goal of the auditory perceptual system is to assign sounds to categories in order to parse words from the speech stream. Because words are stored as sequences of phonemes, sounds must be mapped from acoustic properties to categorical representations.

Phonetic categories group sounds according to their acoustic properties, such as VOT or formant values. When listeners categorize sounds on an acoustic continuum, an S-shaped probability curve emerges. Sounds on the continuum are grouped into distinct phonetic categories with only a small area of uncertainty between them. This type of categorical perception involves simply sorting sounds into categories based on their phonetic properties.

Phonological representations are also categorical, but lack category-internal structure. A phonological category does not distinguish between a /t/ of 60 ms and a /t/ of 90 ms. Phonological processes – such as assimilation, syllabification, and stress assignment – apply at the level of the phonological category, targeting all members of a category. Every token of /t/ is treated in exactly the same way: simply as a member of the category /t/.

Several studies have investigated auditory perception using the mismatch negativity (MMN) response, an automatic response to any change in auditory stimulation generated by the auditory cortex [1]. The appearance of an infrequent deviant causes a 'surprise' response, which manifests as a negative deflection relative to the response to the standard stimulus. In order to generate this surprise response, the auditory cortex must generate a 'memory trace' of the standard stimuli, which it then uses to predict upcoming sounds.

The varying standards MMN paradigm, in which the standards consist of a randomized set of related sounds, is claimed to enforce phonological representations [2]–[7]. The variance among the standards potentially constrains the type of viable memory trace representation that can be generated, and the predictions that can be made.

Here we conduct two studies to determine whether, when making predictions about a varying stream of speech sounds, the auditory perceptual system generates phonetic representations in an ad hoc manner, or whether it uses phonological category representations, stored in long-term memory.

2. EXPERIMENT 1

2.1 Methods

2.1.1 Participants

46 participants were recruited (18 male). All participants were undergraduates at the University of Delaware, native speakers of English, and reported no history of speech or hearing impairment. The average age of participants was 22.5 (SD = 4.6). Participants were compensated either with $20 or extra credit in a linguistics course.

2.1.2 Stimuli and design

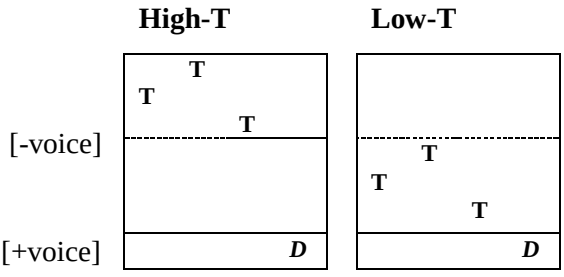
The stimuli were a sequence of synthesized CV syllables composed of an alveolar stop and the

vowel [æ]. Each syllable had a duration of 290 ms. The stimuli were adapted from Hestvik and Durvasula [3], and were generated via Klatt Synthesizer to exactly reconstruct the stimuli used in Phillips et al. [2]. The onset consonant of the deviant stimulus had a VOT value of 15 ms. The consonant onset of the standard stimuli had VOT values of 60, 65, 70, 75, 80, and 85 ms.

The experiment consisted of 3 blocks, corresponding to 3 conditions: Low-T, High-T, and Control. Deviants from each experimental block (Low-T, High-T) would be compared with the deviant from the Control block to establish a main MMN effect. The two experimental-control differences would then be compared to observe a potential phonetic "distance effect". Each block contained 1000 trials. For both Low-T and High-T blocks, the 1000 trials consisted of 900 standards (90%) and 100 deviants (10%). The /tæ/ stimuli in the Low-T condition had an onset [t] with a VOT value of 60, 65, or 70 ms. The /tæ/ stimuli in the High-T condition had an onset [t] with a VOT value of 75, 80, or 85 ms. The oddball /dæ/ in both Low-T and High-T conditions had an onset [d] with a VOT value of 15 ms.

The presentation of trials in the High-T and Low-T conditions was pseudorandomized with at least 3 standards between every 2 deviants. The inter-stimuli interval (ISI) in each condition randomly varied from 410 to 600 ms.

Figure 1: Illustration of standards and deviant in the High-T and Low-T conditions.



Following Jacobsen & Schröger [8], [9], we used a Control condition in which the target /dæ/ appeared in a randomized sequence of equi-probable varying sounds. Because the target /dæ/ is identical to the token used as a deviant in the High-T and Low-T conditions, we refer to it as a "deviant" appearing among "random standards". Stimuli in the Control condition consisted of synthesized syllables with onset VOTs varying by increments of 5 ms (5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 ms). Each stimulus, including the 15 ms token, appeared with equal probability.

The main MMN effect will be generated as a difference between the brain response to the deviant in the experimental blocks (High-T, Low-T) and the same 15 ms token in the Control block. The brain response to the deviant in the experimental block is a function of prediction error and surprise – the auditory perceptual system generates a prediction about upcoming sounds, which is violated by the appearance of the deviant. In the control condition, there is no pattern and no prediction can be made.

By comparing the deviant stimulus to itself in the random control condition, we control for inherent differences in brain response between the standard and deviant stimuli and ensure that the expected MMNs come from the memory comparison. The deviant in the control condition will serve as a control to compute the mismatch effect in lieu of the standard stimuli.

2.1.3 Procedure

The brain responses were recorded in a passive listening EEG paradigm. Participants sat in a sound-attenuating booth and watched a silent movie while stimuli were presented by two free field speakers. Participants were told to watch the film and not pay attention to the sounds. The order of the High-T, Low-T, and Control blocks was randomized for each participant. The entire recording session took approximately one hour.

2.1.4 Data acquisition and analysis

Continuous EEG data were recorded from 128 electrodes in an elastic net (Geodesic Hydrocel 128) and was sampled at 250 Hz. Electrode impedances were lowered to below 50 kΩ.

After acquisition, the data were passed through a 0.3 Hz FIR high-pass filter. The continuous EEG were then segmented into epochs of 1000 ms, with a 200 ms pre-stimulus period. The segmented data were baseline corrected based on the mean voltage of the 200 ms pre-stimulus period. The data were then submitted to an automated process of eyeblink subtraction using ICA with the ERP PCA toolkit, artifact correction, and bad channel replacement. The remaining trials were averaged into 6 cells: High-T-deviants, High-T-standards, Low-T-deviants, Low-T-standards, Control-deviants, and Control-standards. The data were then re-referenced to linked mastoids and 40Hz low-pass filtered.

We used a principal components analysis (PCA) to determine the electrode regions and time windows for ERP analysis. PCA provides a more objective way of selecting time windows and electrode regions for analysis than visual inspection [10], [11]. The

PCA decomposes the temporal and spatial dimensions into a linear combination of a smaller set of abstract ERP factors based on covariance patterns among time points and electrode sets. The PCA can tease apart the underlying contributions of the factors to the summed scalp activity. For the input to the PCA, we used two difference waveforms (Low-T minus Control and High-T minus Control) to identify the main mismatch effect of the Low-T condition and the High-T condition.

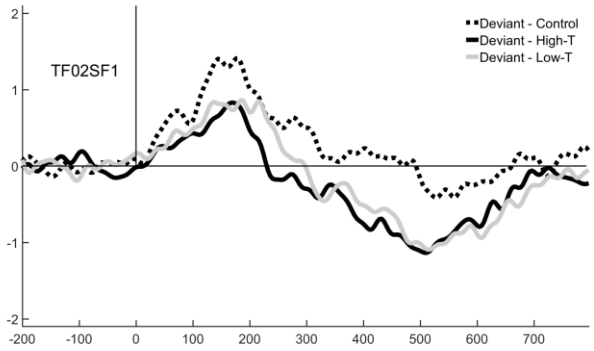
2.2 Results

After pre-processing, 9 participants' data were excluded due to having either more than 10 percent of bad channels or more than 25% bad trials. Of the remaining 37 participants, 8 participants showed a reliable mismatch effect with a positive polarity. We excluded these participants from further analysis, leaving us with 29 total participants.

The PCA generated 23 temporal factors and 6 spatial factors. Of these, two temporospatial factors, each accounting for greater than 5% of the total variance, had a temporal and spatial distribution consistent with an MMN effect. TF02 peaked at 272 ms, and TF04 peaked at 492 ms. Both had a fronto-central scalp distribution. Although MMNs in early time windows are more typical, late MMNs (sometimes referred to as Late Discriminatory Negativity) have been reported in several studies [3], [12]–[14].

A significant mismatch effect was found for both experimental conditions relative to control in both early and late time windows, but the difference between High-T and Low-T conditions was not significant in either time window.

Figure 2: Comparison of brain response to deviant tokens in High-T, Low-T, and Control conditions for Experiment 1.



In the early time window (244-308 ms), High-T vs Control was significant [t(28) = 2.6; p = 0.015], and Low-T vs Control was also significant [t(28) = 2.06; p = .049]. High-T vs Low-T was not significant [t(28) = -1.03; p = 0.311]. In the late time window (480-516 ms), High-T vs Control was highly significant [t(28) = 3.62; p = 0.001], and Low-T vs Control was also highly significant [t(28) = 3.61; p = 0.001], but High-T vs Low-T again was not significant [t(28) = -0.334; p = 0.741].

2.3 Discussion

In both time windows a general mismatch effect was found, indicating that participants discriminated between standards and deviant in the passive listening procedure. The lack of difference in brain response to deviant in the High-T condition and Low-T condition indicates that the High-T and Low-T standards were represented equivalently across the two conditions. There was no phonetic "distance effect" – rather, all tokens in the category /t/ were treated identically. This is evidence for the claim that the auditory perceptual system does not generate ad hoc phonetic representations when phonological representations are available and sufficient to generate a viable prediction.

However, there remains the possibility that phonetic distance is contributing to the overall amplitude of the MMN effect, but that this contribution is overshadowed by the much larger contribution of phonological category (cf. Sharma and Dorman [15]). To test for this possibility, we conducted Experiment 2, in which the VOTs of all stimuli have been uniformly increased so that the deviant is no longer in the voiced category. With the contrast no longer being across-category, we can eliminate phonological category differences as a contributor to the mismatch amplitude.

Also, because the deviant does not fall into an oppositional category relative to the standards, the only way to discriminate deviant from standards is to generate an ad hoc phonetic representation of both. In this way, Experiment 2 will serve as a strict test of the auditory perceptual system's ability to generate ad hoc phonetic representations.

3. EXPERIMENT 2

3.1 Methods

3.1.1 Participants

21 participants were recruited (2 male)[1]. All participants were undergraduates at the University of Delaware, native speakers of English, with no history of speech or hearing impairment. The average age of participants was 20.9 (SD = 2.7). Participants were compensated either with $20 or extra credit in a linguistics course.

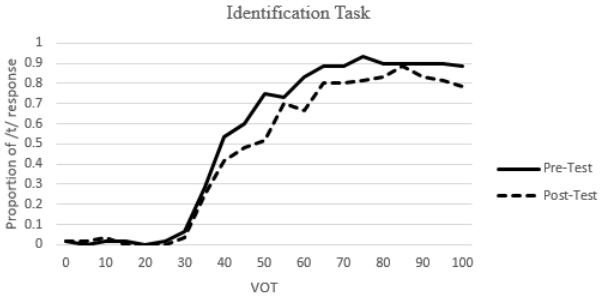
3.1.2 Stimuli and design

Experiment 2 used similar stimuli as experiment 1. New syllables were synthesized that corresponded to the stimuli from experiment 1, with the VOT shifted up by 35 ms. This resulted in a deviant of 50 ms and standards of 95, 100, 105, 110, 115, and 120 ms. The varying standards of the Control condition were also shifted up to the new range.

The experiment was preceded by an identification pre-test and followed by an identical post-test. The pre- and post-test was meant to establish two things: (1) that the 50 ms deviant was not being perceived as a member of the /d/ category; and (2) that exposure to the high VOT values of the standards during the passive listening EEG procedure would not shift the perceptual boundary separating voiced from voiceless.

A threshold analysis of the identification data found a median boundary value of 42.1 ms (SD = 14.4) for the pre-test and 49.6 ms (SD = 22.4) for the post-test. A t-test found the difference between the pre- and post-test was not significant [t(9) = 1.39; p = .197]. This indicates that the 50 ms deviant was not perceived as voiced either before or during the passive EEG procedure, although it is at or near the perceptual boundary for many participants.

Figure 3: Proportion of /t/ responses by VOT in the identification pre- and post-test.



3.1.3 Data acquisition and analysis

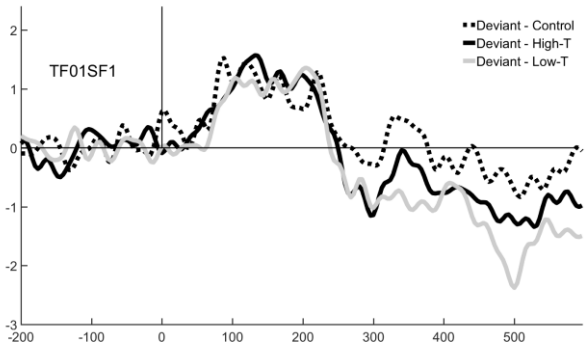
The data were recorded and processed exactly as in Experiment 1, with two minor changes: the raw EEG data were 0.1Hz high-pass filtered and segmented into 800 ms epochs with a 200 ms pre-stimulus baseline period. The high-pass filter was changed to avoid generating illusory ERP effects [16]. The data were again subject to a PCA to select time and spatial regions for analysis.

3.2 Results

After processing the data, 6 participants' data were excluded due to having either more than 10 percent of bad channels or more than 25% bad trials. Of the remaining 15 participants, 4 participants showed a reliable mismatch effect with a positive polarity, and one participant showed no discrimination in the pre-test. These participants were also excluded, leaving 10 total participants.

The PCA picked out 18 temporal and 5 spatial factors, of which one accounted for greater than 5% of the total variance and had a distribution consistent with an MMN effect: TF01SF1. Again we found a significant difference between High-T and Control [t(9) = 2.45; p = .037] and a significant difference between Low-T and Control [t(9) = 2.40; p = .04], but no significant difference between High-T and Low-T [t(9) = .304; p = .768].

Figure 4: Comparison of brain response to deviant tokens in High-T, Low-T, and Control conditions for Experiment 2.



3.3 Discussion

The results of Experiment 2 mirror the results of Experiment 1. This indicates that participants successfully discriminated between standards and deviant, even when the deviant was not a clear member of an opposing phonological category. Because the auditory system cannot rely on phonological category assignment to differentiate the standards and deviant in this case, it must be sorting the two distant types into phonetic representations generated in an ad hoc manner. There are no pre-defined categories that can group high-VOT exemplars of /t/ to the exclusion of a marginal/boundary /t/. Rather, the successful discrimination is evidence that representations are being generated on the fly.

However, the absence of a distance effect suggests that these ad hoc representations are not fully specified – they do not contain detailed information about VOT. Whether the failure to measure a distance effect is due to an absence of specified phonetic information or simply the system's (lack of) sensitivity to the magnitude of the difference will require further study.

4. REFERENCES

[1] K. Alho. 1995. “Cerebral generators of mismatch megativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes,” *Ear Hear.*, vol. 16, pp. 38–51.

[2] C. Phillips, T. Pellathy, a Marantz, E. Yellin, K. Wexler, D. Poeppel, M. McGinnis, and T. Roberts. 2000. “Auditory cortex accesses phonological categories: an MEG mismatch study.,” *J. Cogn. Neurosci.*, vol. 12, no. 6, pp. 1038–1055.

[3] A. Hestvik and K. Durvasula. 2016. “Brain & Language Neurobiological evidence for voicing underspecification in English,” vol. 152, pp. 28–43.

[4] G. Dehaene-Lambertz and M. Pena. 2001. “Electrophysiological evidence for automatic phonetic processing in neonates,” *Cogn. Neurosci. Neuropsychol.*, vol. 12, no. 14, pp. 3155–3158.

[5] A. Shestakova, C. A. E. Brattico, M. Huotilainen, V. Galunov, A. Soloviev, M. Sams, R. J. Ilmoniemi, and R. Nìitìnen. 2002. “Abstract phoneme representations in the left temporal cortex : magnetic mismatch negativity study,” *Cogn. Neurosci. Neuropsychol.*, vol. 13, no. 0, pp. 1–5.

[6] T. Jacobsen and E. Schröger. 2004. “Pre-attentive perception of vowel phonemes from variable speech stimuli,” *Psychophysiology*, vol. 41, pp. 654–659.

[7] C. Eulitz and A. Lahiri. 2004. “Neurobiological Evidence for Abstract Phonological Representations in the Mental Lexicon during Speech Recognition,” pp. 577–583.

[8] T. Jacobsen and E. Schröger. 2003. “Measuring duration mismatch negativity,” *Clin. Neurophysiol.*, vol. 114, pp. 1133–1143.

[9] T. Jacobsen and E. Schröger. 2001. “Is there pre-attentive memory-based comparison of pitch?,” pp. 723–727.

[10] J. Dien. 2012. “Applying principal components analysis to event-related potentials: A tutorial,” *Dev. Neuropsychol.*, vol. 37, no. 6, pp. 497–517.

[11] J. Dien, W. Khoe, and G. R. Mangun. 2007. “Evaluation of PCA and ICA of simulated ERPs: Promax vs. infomax rotations,” *Hum. Brain Mapp.*, vol. 28, no. 8, pp. 742–763.

[12] M. Cheour, R. Ceponiene, A. Lehtokoski, A. Luuk, J. Allik, K. Alho, and R. Näätänen. 1998. “Development of language-specific phoneme representations in the infant brain.,” *Nat. Neurosci.*, vol. 1, no. 5, pp. 351–353.

[13] H. Datta, V. L. Shafer, M. L. Morr, D. Kurtzberg, and R. G. Schwartz. 2010. “Electrophysiological indices of discrimination of long-duration, phonetically similar vowels in children with typical and atypical language development,” *J. Speech, Lang. Hear. Res.*, vol. 53, pp. 757–778.

[14] V. L. Shafer, M. L. Morr, H. Datta, D. Kurtzberg, and R. G. Schwartz. 2005. “Neurophysiological indexes of speech processing deficits in children with specific language impairment,” *J ournal Cogn. Neurosci.*, vol. 17, no. 7, pp. 1168–1180.

[15] A. Sharma and M. F. Dorman. 1999. “Cortical auditory evoked potential correlates of categorical perception of voice-onset time.,” *J. Acoust. Soc. Am.*, vol. 106, no. 2, pp. 1078–83.

[16] D. Tanner, K. Morgan-short, and S. J. Luck. 2015. “How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition,” *Psychophysiology*, vol. 52, pp. 997–1009.

[1]More participants are being recruited, and more data is currently being collected.