THE INSTITUTE FOR SYSTEMS RESEARCH

ISR TECHNICAL REPORT 2007-20

Seeking the Neural Basis of Auditory Perception: A Study in Phoneme Confusions

Lee Zhang

Seeking the Neural Basis of Auditory Perception: A Study in Phoneme Confusions

Lee Zhang

Mentors: Dr. Shihab A. Shamma, Dr. Jonathan B. Fritz, and Mr. Nima Mesgarani

Faculty Advisor: Dr. S.K. Gupta

ABSTRACT

Recent studies concerning phoneme representation and classification suggest neural responses in the primary auditory cortex of ferrets are "sufficiently rich to encode and discriminate phoneme classes, and that humans and animals may build upon the same general acoustic representations to learn boundaries for categorical and robust sound classification."[1] This paper further explores phoneme discrimination— specifically perceptual confusion among plosives /p/, /t/ and /k/ and fricatives /s/ and /ʃ/— in ferrets, the ability for the animals to generalize across different speakers, and also the behavioral training procedure used to test the sensory and perceptual abilities of the animals.

INTRODUCTION

Perceptual Confusions

There are multiple levels of representation in mapping sound to meaning. Distinctive features, the smallest units of speech with acoustic interpretation, "form the basic inventory characterizing the sounds of all languages."[2] Coordinated bundles of these distinctive features constitute phonemes, which are sequenced to be the building blocks of words.[3] Humans "reliably identify many phonemes and discriminate them categorically, despite considerable natural variability across speakers."[4] But despite their expertise humans do confuse phonemes, especially in unusual or noisy contexts.

Miller and Nicely recognized perceptual confusions are often far from random, and suggested advances can be made in communication and the understanding of speech perception by studying the kinds of errors that occur.[5] 16 common English consonants (/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/ and /n/) were presented before the vowel /a/ (as in f*a*ther) and masked with various levels of frequency distortion and noise. Listeners "were forced to guess at every sound and a count was made of all the different errors that resulted when one sound was confused with another."[6] Results were displayed in tables referred to as confusion matrices (Figure 1).

Study of error distribution suggested phonemes that share some acoustic features tend to be more confusable than those that do not. For example, while there were 38 recorded instances of /p/-/t/ confusion and 88 recorded instances of /p/-/k/ confusion, the plosive /p/ was seldom confused with fricatives /f/, /θ/, /s/ and /ʃ/ and never confused with nasals /m/ and /n/.

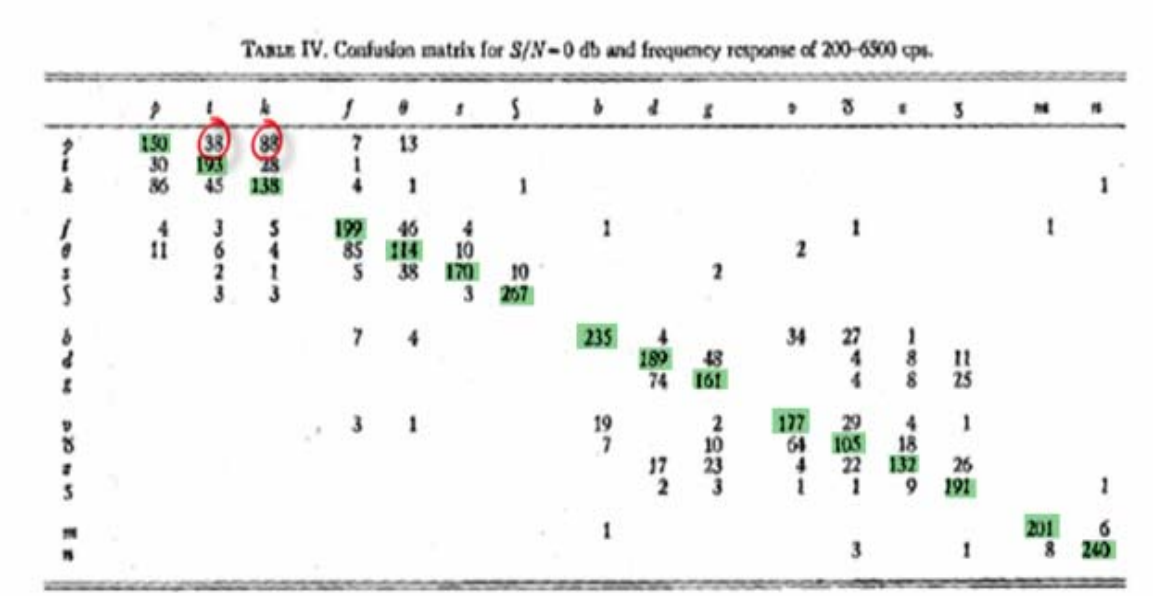


TABLE IV. Confusion matrix for $S/N = 0$ db and frequency response of 200-6500 cps.

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	130	38	88	7	13											
t	30	193	28	1												
k	86	45	138	4	1		1									1
f	4	3	5	199	46	4		1				1			1	
θ	11	6	4	85	114	10					2					
s		2	1	5	38	170	10			2						
ʃ		3	3			3	207									
b				7	4			235	4		34	27	1			
d									189	48		4	8	11		
g									74	161		4	8	25		
v				3	1			19		2	177	29	4	1		
ð								7		10	64	105	18			
z									17	23	4	22	132	26		
ʒ									2	3	1	1	9	191		1
m								1							201	6
n												3		1	8	240

Figure 1: Confusion Matrix

This table displays the data collected for speech-to-noise ratio (SNR) of 0 db and frequency response of 200-6500 cps. The syllables that were spoken are indicated by the consonants listed vertically in the first column on the left; the syllables that were written by the listener are listed horizontally across the top of the table. The number in each cell is the frequency that each stimulus-response pair was observed. Correct responses, listed along the main diagonal, are highlighted in green.[7]

Analysis of the results focused on voicing, nasality, affrication, duration, and place of articulation— articulatory features of speech production that characterize different phonemes and are presumably discriminated by the listener.

Neurophysiological Basis

Recently, focus has shifted to the neurophysiological basis of understanding language. But despite decades of research, the functional neuroanatomy of speech perception remains difficult to characterize. Hickok and Poepeel suggested speech is special— that "lexical items have some representational property that sets them apart from other auditory information."[8]

The implicit goal of speech perception studies is to understand sublexical stages (such as syllable discrimination) in the process of speech recognition.[9] Trained animals have been shown to discriminate phoneme pairs categorically and distinguish phonetic acoustic features, suggesting suggest speech perception may not be unique to humans. Steinschneider's investigation the neural mechanisms underlying speech perception and perceptual confusions focused on voice onset time— an articulatory parameter measuring "the interval between consonant release (onset) and the start of rhythmic vocal cord vibrations (voicing)."[10] The model findings in monkeys revealed a "characteristic pattern of activity" similar to speech-evoked response pattern recorded directly from human auditory cortex.[11]

Recent studies suggest the primary auditory cortex responses in ferrets are "sufficiently rich to encode and discriminate phoneme classes."[12] Mesgarani's study on the encoding of consonants focused how place of articulation, manner of articulation and voicing are encoded in the neuron population in the primary auditory cortex of ferrets. Analysis of the acoustic similarity among the phonemes at the level of auditory spectrograms reflected fundamental similarities to human and neural confusion matrices.[13]

In this current study, we set out to compare perceptual confusions in ferrets with that in humans in hopes of obtaining better understanding of the neural representation of complex patterns. Since trained animals have been shown to "discriminate phoneme pairs categorically and to generalize to novel situations"[14] we tested the animals using three different speakers, to see if they could generalize and develop and abstract representation of sounds independent of frequency.

We wanted our results to be comparable to those recorded by Miller and Nicely, but reduced the range of speech-to-noise ratios tested to from -18, -12, -6, 0, +6 and +12 db to 0, +3, +6, +9 and 100 db (with 100 db corresponding to clear speech) in order to simply the task. We began with a comparison of a smaller set of phonemes consisting of three plosives (/p/, /t/ and /k/) and two fricatives (/s/ and /ʃ/).

METHODS

All experimental procedures were in accord with National Institutes of Health Guidelines and approved by the University of Maryland Animal Care and Use Committee. Before working with the animals, I received online training through the American Association for Laboratory Animal Science Learning Library and in-facility training at Central Animal Resources Facilities with licensed veterinarian Dr. Hall. I passed a core of specific on-line courses required of all animal research personnel and additional courses addressing the specific needs of the research, and attended a discussion reviewing documentation requirements, facility standard operating procedures, nutrition, husbandry, handling, zoonotic diseases and occupational health. Training records can be located at the UM Department of Laboratory Animal Care office.

Behavioral Training

Two adult ferrets were trained to lick water from a reward spout during the presentation of reference sounds and to stop licking after the presentation of target sounds. The animals were trained on this two-choice task using conditioned avoidance. The psychophysical procedure, as outlined by Heffner and Heffner, involves

> training an animal to make steady contact with a reward spout in order to receive food or water and then pairing a stimulus with mild electric shock delivered through the spout. The animal quickly learns to avoid the shock by breaking contact with the spout whenever it detects the stimulus. The breaking of contact with the spout is then used to indicate that the animal detected the stimulus.[15]

The ferrets were trained twice daily. Each day, one of the five phonemes (/p/, /t/, /k/, /s/, or /ʃ/) was randomly fixed as the reference. In each trial a random number (1-6) of reference sounds were presented followed, by a target sound. (The exception being sham trials, during which the reference sound was again presented.). The animal licked water from the spout while listening to the sequence of reference sounds, and learned to stop licking after the presentation of a target sound or receive a mild shock. Each trial occurred in a random level of speech-to-noise of either 0, +3, +6, +9 or 100 db.

Stimuli

My mentor, Nima Mesgarani, began training the animals using a recording of a single female speaker presenting each of the five phonemes before the vowel /a/. The animals performed reasonably well on this paradigm. However the database only provided one recorded sound sample for each phoneme from the selected speaker.

Humans "reliably identify many phonemes and discriminate them categorically, despite considerable natural variability across speakers."[16] As a child matures "there is an increase in vocal-tract length, and as a result, the formant frequencies of the vowels decrease."[17] Humans are able to recognize that the sounds produced by men, women and children saying a given phoneme are indeed the same, despite differences in the

waveforms. In order to study the animals' ability to normalize the phonemes across speakers, the experiment was then conducted using a recording of a single male speaker presenting each of the five phonemes. The animals perform reasonably well on this paradigm, but were still only exposed to a single recorded sound sample for each phoneme from the selected speaker.

The current paradigm features a recording of my voice. I presented 16 common English consonants (/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/ and /n/) before the vowel /a/. Although we used English phonemes, we decided to record also using tones.

Mandarin Chinese is a tonal language consisting of four tones (not including the neutral tone.). By saying "ma" in different tones one can ask, "Did mother scold the horse?"

媽 罵 馬 嗎?

(mā mà mǎ ma?)[18]

The following table[19] illustrates tone markings above "ma" and describes how each tone is vocalized:

1st	mā	High and level.
2nd	má	Starts medium in tone, then rises to the top.
3rd	mǎ	Starts low, dips to the bottom, then rises toward the top.
4th	mà	Starts at the top, then falls sharp and strong to the bottom.

We recorded five samples for each phoneme to simulate the natural variations in everyday speech and normalized the duration of each sample for the test phonemes. Each

reference and target sound was chosen randomly from among the five recorded samples for each phoneme.

The animals were trained using the fourth tone. In future studies, it may be worthwhile to train the animals on the other three tones and record the reactions to the different inflections.

RESULTS

The animal licked water from the spout while listening to the sequence of reference sounds, and learned to stop licking after the presentation of a target sound or receive a mild shock. A contact switch connected between the spout and the cage floor detected the animal's contact with the reward spout, and a computer recorded whether or not the animal was in contact with the spout immediately before the shock was delivered (see Figure 2).

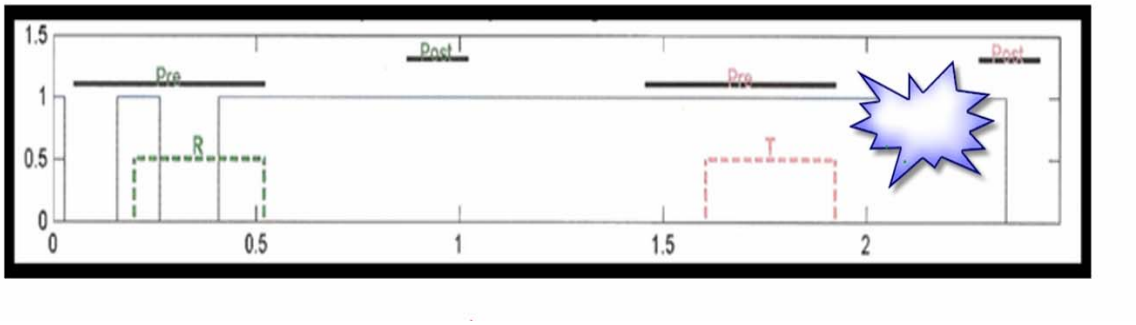


Figure 2: Trial Presentation

This figure shows the time frame for a trial during which a target phoneme was presented after one reference phoneme. The dotted green line labeled with a capital R indicates the period during which the reference phoneme was presented; the dotted red line labeled with a capital T indicates the period during which the target phoneme was presented. The solid black lines span pre- and post-stimulus windows. During this trail, the animal did not pull away after the presentation of the target phoneme and so received a mild shock.

Breaking contact after the presentation of a target sound were recorded as hits while failure to do so were recorded as a misses. False alarms, breaking contact in the absence of a warning stimulus, were obtained by determining the animals during safe trial, intervals when a stimulus could have been, but was not, presented.[20] The animals' performances were recorded in blocks of 22 trials (see Figures 3 and 4).

Trial	SF	HR	DR	H#	SZ	HR	SN	Shm	bSR	bHR	bDR
22	79	35	28	7/20	0		0	2	79	35	28
44	82	33	27	13/39	0		1	4	84	32	27
66	85	30	25	17/57	0		3	6	91	22	20
88	84	30	25	23/77	0		3	8	79	30	24
110	82	28	23	27/95	20		5	10	77	22	17
116	83	28	23	28/99	33		6	11	75	24	18

Figure 3: Data Analysis

This table displays the data collected for the animal Saturn on June 12th, 2007. The first column on the left indicates the trial block. The second column indicates the safe rate (SF), the third column indicates the hit rate (HR), and the fourth column indicates the discrimination rate (DR) for each block the animal performed for. This is one of the first sessions conducted under the current paradigm, and the animal's performance hovers around chance (25%).

Trial	SF	HR	DR	H#	SZ	HR	SN	Shm	bSR	bHR	bDR
22	72	63	45	12/19	100		1	2	72	63	45
44	77	59	45	23/39	100		1	4	82	55	45
66	76	53	40	31/59	100		1	6	74	40	30
88	77	49	38	39/79	100		1	8	82	40	33
96	78	48	37	40/84	100		2	9	83	42	35

Figure 4: Data Analysis

This table displays the data collected for the animal on July 25th, 2007. After six weeks of training the animal's discrimination rates (45-45-40-38-37) have risen well above chance.

The shock administered was adaptively adjusted based on the animal's performance. It was important to find the lowest level that would produce reliable avoidance— too low a level would result in a low hit rate while too high a level would result in a high false alarm rate.[21] More explanations concerning the data collected can be found in the appendix (see Figures 5-10).

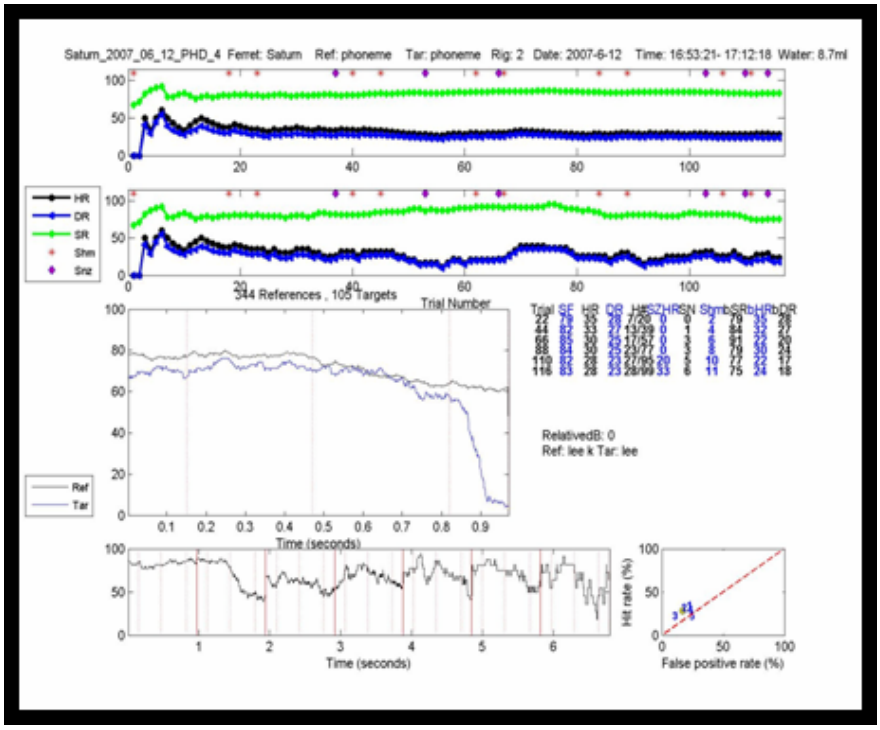
The data was then sorted, and unusable trial blocks for which the animal performed at a discrimination rate below 30% were discarded. Although the animals performed better on certain phonemes than others, there were at least 32 useable trials for each phoneme for the animal Zim and at least 48 useable trials for each phoneme for the animal Saturn.

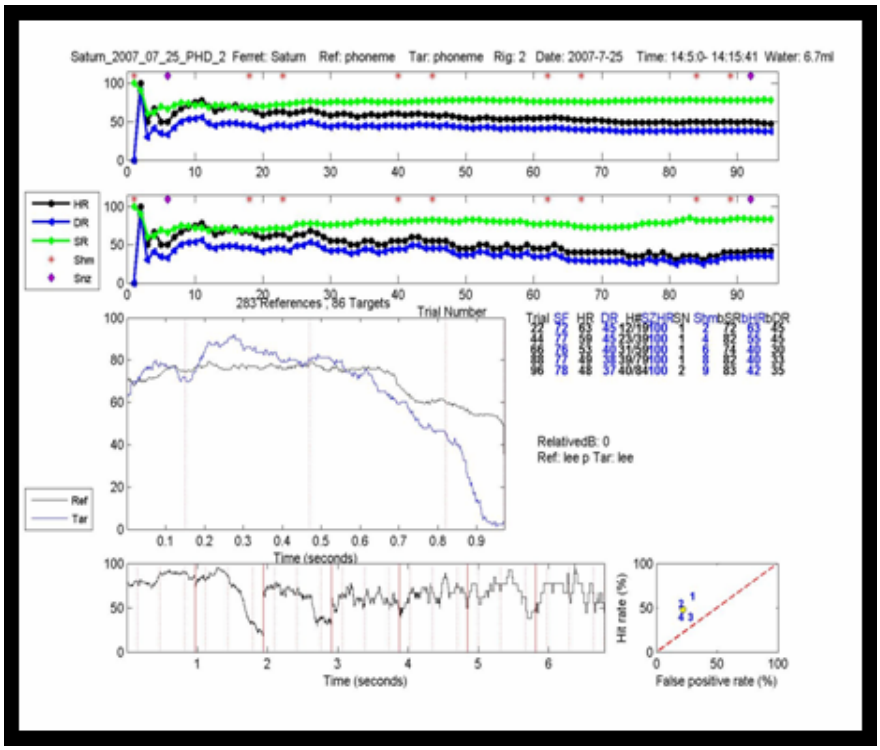
DISCUSSION

Humans have the amazing abilities to reliably identify and discriminate phonemes categorically, to generalize across speakers, to become specialists in their language. Understanding speech perception has been a topic of investigation for more than 130 years.

The "implicit goal of speech perception studies is to understand sublexical stages in the process of speech recognition (auditory comprehension)."[22] From a biological perspective, research on the neural processes supporting speech perception gives new insight to the neural representation of complex patterns and speech processing. From a practical perspective, speech research has applications in revealing additional strategies to improve automated speech identification systems and speech recognition for hearing- and language-impaired listeners.

APPENDIX





Figures 5 and 6: Data Collection

The data gathered during each training session were organized into diagrams like the two seen above. Figure 5 (on top) was collected from the animal Saturn on June 12th, 2007, one of the first training sessions on this paradigm. Figure 6 (on bottom) was collected from the same animal five weeks later on July 25th, 2007.

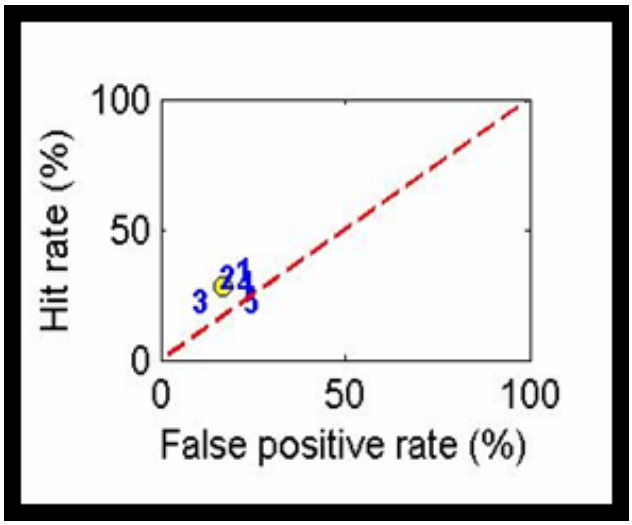


Figure 7: Hit Rate, False Positive Analysis

This graph, taken from Figure 5, gives an overview of the data collected from the June 12th training session. The blue numbers show the animal's performance for each trial block in relation to false positive rate and hit rate. Notice how all the numbers are clustered around the dotted red, an indicator of chance performance.

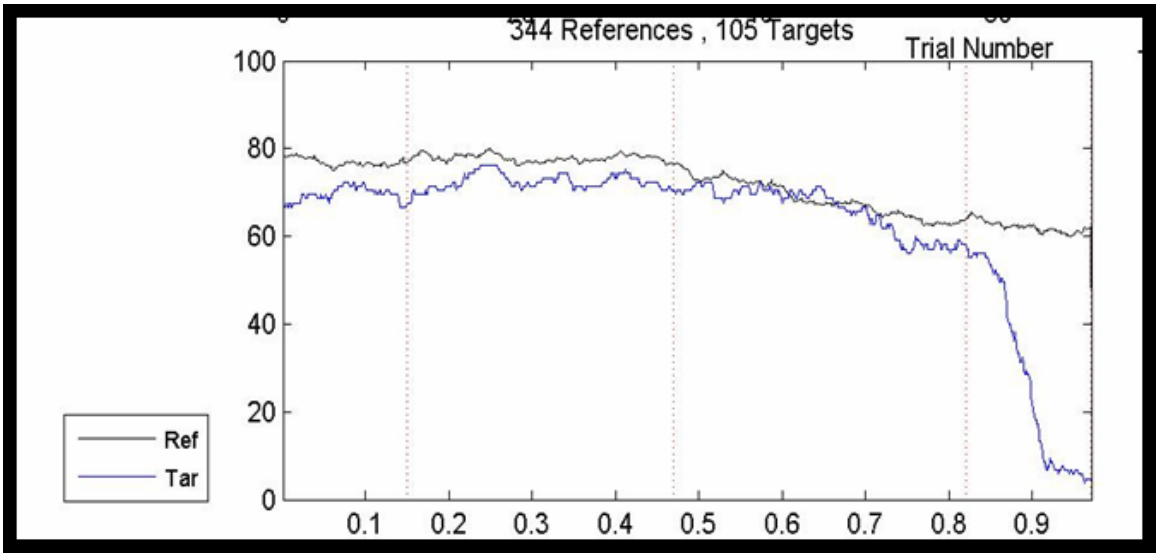


Figure 8: Lick Behavior Analysis

This graph, also taken from Figure 5, describes the lick behavior data collected from the June 12th training session. The black Ref line indicates licking behavior before, during and after reference sounds; the blue Tar line indicates licking behavior before, during and after target sounds. During this training session, the animal responded similarly to reference sounds and target sound. The animal did pull away after the presentation of target sounds, and there is little difference between the two lines until .85 seconds, well into the shock window.

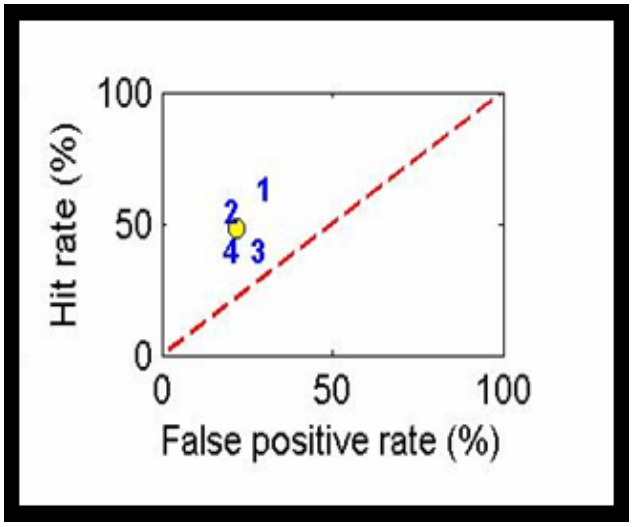


Figure 9: Hit Rate, False Positive Analysis

This graph, taken from Figure 6, gives an overview of the data collected from the July 25th, training session. The numbers, especially representing trial blocks 1 and 2, have distanced from the dotted red line, indicating above chance performances.

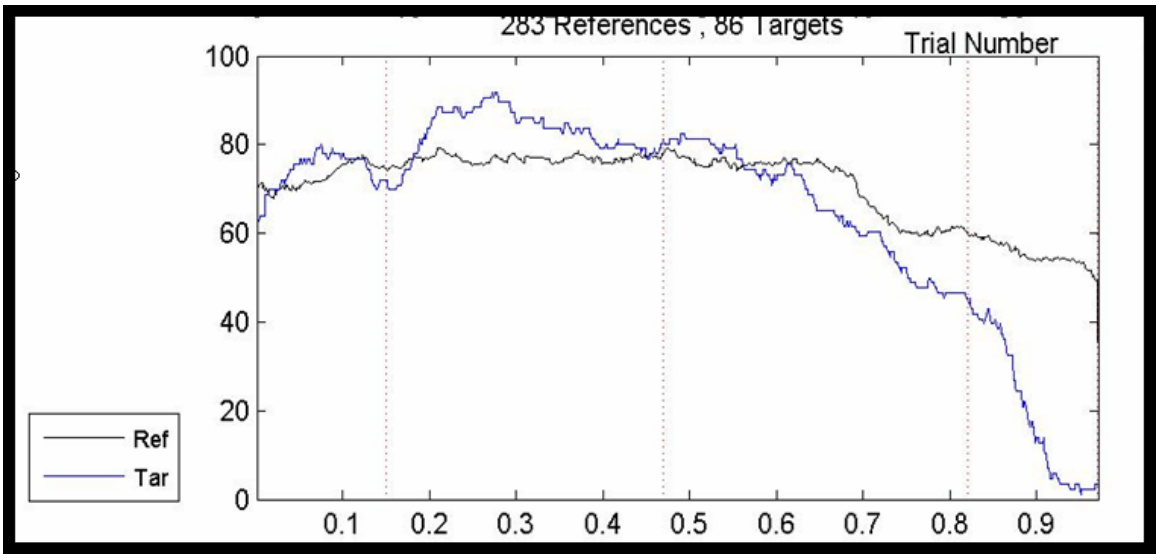


Figure 10: Lick Behavior Analysis

This graph, also taken from Figure 6, describes the lick behavior data collected from the July 25th, training session. Although there are slight differences between the two lines, the black Ref line and blue Tar do not begin to distance from one another more noticeably until around .6 seconds. There is an especially sharp drop (from just above 60 to just above 40) around .85 seconds, just before the shock is administered.

[1] Mesgarani, N., David, S.V., Fritz, J.B. and Shamma, S.A. Phoneme representation and classification in primary auditory cortex.
[2] Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature* 8, 393-402 (2007).
[3] Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature* 8, 393-402 (2007).
[4] Mesgarani, N., David, S.V., Fritz, J.B. and Shamma, S.A. Phoneme representation and classification in primary auditory cortex.
[5] Miller, G.A. and P.E. Nicely. An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America* 27, 338-352 (1955).
[6] Miller, G.A. and P.E. Nicely. An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America* 27, 338-352 (1955).
[7] Miller, G.A. and P.E. Nicely. An Analysis of Perceptual Confusions Among Some English Consonants. *The Journal of the Acoustical Society of America* 27, 338-352 (1955).
[8] Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature* 8, 393-402 (2007).
[9] Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature* 8, 393-402 (2007).
[10] Lisker L. and Abramson, A.S. A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 384-422 (1961).
[11] Steinschneider, M., Volkov, I.O., Fishman, Y.I., Oya, H., Arezzo, J.C. and Howard, M. A. Intracortical Reponses in Human and Monkey Primary Auditory Cortex Support a Temporal Processing Mechanism for Encoding of the Voice Onset Time and Phonetic Parameter. *Cerebral Cortex* 15, 170-186 (2004).
[12] Mesgarani, N., David, S.V., Fritz, J.B. and Shamma, S.A. Phoneme representation and classification in primary auditory cortex.
[13] Mesgarani, N., David, S.V., Fritz, J.B. and Shamma, S.A. Phoneme representation and classification in primary auditory cortex.
[14] Mesgarani, N., David, S.V., Fritz, J.B. and Shamma, S.A. Phoneme representation and classification in primary auditory cortex.
[15] Heffner, H.E. and Heffner, R.S. Conditioned Avoidance. in *Methods in Comparative Psychoacoustics* (eds. Klump, G.M. *et al*.) 79-94 (Basel: Birkhauser Verlag, 1995).
[16] Mesgarani, N., David, S.V., Fritz, J.B. and Shamma, S.A. Phoneme representation and classification in primary auditory cortex.
[17] Smith, D.R., Patterson, R.D., and Turner, R. The processing and perception of size information. *The Journal of the Acoustical Society of America* 177, 305-318 (2004).
[18] "The Tones of Mandarin Chinese." The Chinese Outpost. 8 August 2007 <http://www.chinese-outpost.com>.
[19] "The Tones of Mandarin Chinese." The Chinese Outpost. 8 August 2007 <http://www.chinese-outpost.com>.
[20] Heffner, H.E. and Heffner, R.S. Conditioned Avoidance. in *Methods in Comparative Psychoacoustics* (eds. Klump, G.M. *et al*.) 79-94 (Basel: Birkhauser Verlag, 1995).
[21] Heffner, H.E. and Heffner, R.S. Conditioned Avoidance. in *Methods in Comparative Psychoacoustics* (eds. Klump, G.M. *et al*.) 79-94 (Basel: Birkhauser Verlag, 1995).
[22] Hickok, G. and Poeppel, D. The cortical organization of speech processing. *Nature* 8, 393-402 (2007).