4-2001

A novel experimental paradigm for study of incidental learning of sound categories

Zelska Buturovic
Carnegie Mellon University

A novel experimental paradigm for study of incidental learning of sound categories

Zeljka Buturovic

Adviser: Lori L. Holt

Submitted in partial fulfillment of Senior Honors Thesis
Carnegie Mellon University
April 18, 2001

Introduction

An early and fundamental task that any novice in the language environment (be that an infant learning her first language or an adult learning his second one) must accomplish is to be able to differentiate between linguistically-relevant and linguistically-irrelevant information. When we look at the speech as a physical signal, we see that its variability depends on the message that is being conveyed but also on extra-linguistic influences such as which speaker is talking, how quickly that speaker is speaking, and even what emotional state the speaker is in presently. Each of these sources of variability may be important for speech processing, but for a listener to understand the linguistic message it is crucial that he attunes to variability in speech signal that is relevant for the meaning and discounts variability that is not. Essentially, this is a task of categorization.

Although categorization appears to be necessary to speech processing, the specific characteristics of speech categorization depend on the language environment of the learner. The exact same physical variability in speech signal can be linguistically functional in one language but serve no function in another one. This has become known as a task of phonetic categorization.

Quite a bit is already understood about phonetic categorization. Recent demonstrations of language-appropriate categorical behavior in pre-linguistic infants (e.g., Kuhl et al., 1992; Polka & Werker, 1994), methodological advances in examining the interior structure of phonetic categories (e.g., Volaitis & Miller, 1992; Johnson et al., 1993; Iverson & Kuhl, 1995; Aaltonen et al., 1997; Tremblay et al., 1997; Cheour et al., 1998; Lotto et al., 1998), and demonstrations of the malleability of adults' language-specific categorical behavior (e.g., Logan et al., 1991; Lively et al., 1994; Flege et al.,

1997; McClelland, 1998) have led to a vital sub-field in speech perception concerned with describing the function of categories in development and maintenance of language-appropriate perception.

Some exciting new research suggests that listeners may parcel speech variability with quite general learning mechanisms. For example, new data and computational models in other arenas of language acquisition (e.g., Brent & Cartwright, 1996; Saffran et al., 1996), suggest general learning mechanisms may play an important role in language learning. In addition, there is already some preliminary evidence that useful techniques for assisting second language learners in categorizing sounds of the second language can be developed from general principles of perceptual organization (McCandliss et al., 1999; McClelland et al., in press).

If phonetic categorization really is related to the general learning and categorization capacities of humans, then it may be well-informed by the large literature that already exists for visual categorization. However, speech categories are notoriously difficult to define in terms of any particular number of attributes, features, or cues. Usually, no particular attribute is either necessary or sufficient for a given acoustic segment to belong to a particular phonetic category. This sort of category is quite different from the categories typically studied in visual categorization. In order for these kinds of categories to develop, at least some ability to learn from the correlations among attributes is probably necessary. It is now becoming apparent that humans are remarkably tuned to this kind of information (e.g., Pitt & McQueen, 1998; Aslin et al., 1999; Saffran et al., 1999). However, categorization of auditory signals is not yet well understood.

Another major difference in phonetic categorization versus visual categorization is that visual categories typically have been studied by examining categorization of novel stimuli with methods that rely upon explicit feedback to train participants to label stimuli. This is unlikely to be the manner by which phonetic categories are learned, so it would be very desirable to have a means by which to examine categorization without reliance on explicit training paradigms.

The specific goal of this experiment is to develop a method for studying auditory categorization without feedback. In the broadest sense, this project aims to develop a set of complex non-speech auditory stimuli and train human adults to categorize these sounds. The structure of these categories mimics that typical of phonetic categories and the training procedures attempt to more closely model the means by which phonetic categories might be learned. By using novel auditory stimuli in our experiment, we have a means to study development of auditory categories in adult listeners despite their extensive auditory experience with natural signals like human speech. Well-controlled laboratory experience with structured novel sound distributions allows us to gauge consequent changes in sound perception. Avoiding explicit training in sound categorization tasks could bring unique advantage in ecological validity to this method.

The experiment described here implements this goal by perfectly correlating novel non-speech sound stimuli with the spatial location of visual stimuli presented on a computer monitor. Conceptually, a novel auditory stimulus space that varied in two acoustic dimensions was "overlaid" upon the two spatial dimensions of the computer monitor. In this way, each sound corresponded to a particular location on a computer monitor. Participants were not informed of this perfect correlation. Rather, they heard the

sound as a "warning" that a visual stimulus would soon appear. Participants were under instructions to detect the identity of the visual stimulus as an 'X' or an 'O' as quickly and as accurately as possible. The identity of the visual stimulus varied randomly, with 'Xs' and 'Os' equally likely for any spatial location, but with spatial location entirely determined by the identity of the sound.

The sounds were drawn from one of two underlying distributions in the two-dimensional acoustic stimulus space. The location of the visual stimulus on the computer monitor was perfectly predicted by the acoustic characteristics of the sound. Though the sound quality varied from trial to trial, the subject was not instructed that the sound and visual stimulus have any relationship whatsoever. On any given trial, the visual stimulus presented could be either an 'X' or an 'O'. If participants are sensitive to the underlying relationship between sound and visual stimulus, then reaction time for identification of the visual stimuli should be faster than for control conditions where sound and visual stimulus are uncorrelated. If participants pick up on the regularity underlying this task, this paradigm may serve as a novel means of implicitly "training" listeners to categorize sounds with various underlying statistical distributions.

The feasibility of this kind of learning is motivated by the evidence from visual categorization which suggests that adults are capable of categorization under conditions without labels or error correction (e.g., Edmonds & Evans, 1966; Fried, 1979). Most impressively, Fried and Holyoak (1984) demonstrated that observers form visual categories with no feedback, no knowledge of the number of categories, no instructions to categorize and even no knowledge that they are involved in a categorization task.

These categorization data appear to reflect a tendency of humans to parcel responses in a manner that reflects the underlying structure of input distributions.

If we demonstrate that adults are indeed capable of learning associations between auditory and visual stimuli, many questions in auditory categorization can be addressed. For example, one could reliably address listeners' ability to discriminate sounds by changes in reaction times in recognition of (essentially irrelevant) visual stimuli. Furthermore, this method might be adapted to address development of such ubiquitous phenomena as categorical perception, perceptual magnets, hyperspace effects and even usefulness of infant directed speech ("motherese") in acquisition of speech categories.

Methods

Participants

Seventeen undergraduate students from Carnegie Mellon University participated in exchange for psychology course credit.

Stimuli

Acoustic Stimulus Set. Novel nonspeech stimuli were synthesized to create a 2-dimensional stimulus space. The stimuli that inhabit this space were sculpted from 300-millisecond (ms) bursts of white (Gaussian) noise. White noise, as opposed to single tones or collections of several tones, has a continuous uniform spectrum that is, in its complexity, similar to speech though it sounds nothing like speech. Using digital signal processing, energy was eliminated from two parts of the spectrum by bandstop filtering across two 300-Hz-wide bands. This created two frequency "notches" in the noise. These

notches are referred to as NF1 and NF2, respectively (for Negative First Formant and Negative Second Formant, to make obvious the relation to vowel stimuli).[1] NF1 and NF2 frequencies were logarithmically scaled to produce perceptually uniform increments. Each stimulus had a 10 ms linear onset and offset amplitude ramp. All stimuli were RMS matched in overall amplitude and saved digitally. Figure 1 illustrates the waveform and spectrogram for each of two stimuli drawn from this space.

Each auditory stimulus had unique NF1 and NF2 frequency values, creating a 2-dimensional stimulus space. This stimulus space was sampled to create a stimulus set for the experiment. Sampling was done to create two identifiable categories of sound. Figure 2 displays the configuration of these categories. The distribution of the stimuli within each category was not uniform. Rather, exemplars from the centroid of the category distribution were presented more frequently than exemplars on the outskirts of the category:

The distribution of stimulus presentation is shown in Figure 3. Stimuli from the inner ring (shown in green) were presented 13 times each, for a total of 52 stimuli (or 31% of the trials), stimuli from the second ring (purple) were presented four times each, for a total of 48 stimuli (or in 28% of the trials), stimuli from the third ring (blue) were presented two times each, for a total of 40 stimuli (24% of the trials), while stimuli from the outer rings (red) were presented one times each, for a total of 28 stimuli (i.e. in 17% percent of the trials). Thus, within one block, listeners heard 168 stimuli in randomized order from each of the two categories, i.e. there were 2*168=336 trials per block. The whole experiment consisted of 3 blocks, each of 336 trials for a total of 1008 trials.

[1] In the purest definition of the term, this is an abuse of the term "formant" (which does not mean simply a spectral prominence, but is essentially an articulatory term). However, we feel that this indiscretion is compensated for by the mnemonic value added by the terms.

Successive blocks were separated by a short break. The total duration of the experiment was about an hour.

Visual Stimulus Set. For each auditory stimulus, two visual stimuli were created. The visual stimuli were related to the auditory stimuli by an implicit mapping from the 2-dimensional NF1xNF2 stimulus space to the spatial layout of the computer monitor used in the experiment. Visual stimuli with dimensions of 38x38 pixels were placed on a larger white image created to fill the 760x760 pixels centered white screen of an iMac 1024x768 monitor. This centered white screen was divided into 20*20=400 visual stimuli (38x38 pixels). A single visual stimulus was a 38x38 white square that had either a black 'O' or a black 'X' in its center. The placement of this 38x38 pixel square in the larger 760x760 white image could vary across any of 128 different positions on the screen. The position of the visual image on the screen was perfectly correlated with the auditory stimulus (experimental condition) or uncorrelated with auditory stimulus (control condition). For example, in experimental condition the auditory stimulus with an NF1 notch at 431 Hz (i.e. 281-581 Hz notch) and NF2 notch at 1345 (i.e. 1195-1495 Hz notch) was always followed by a visual image for which the bottom left corner was positioned at 118x550 pixels on the iMac screen. In the control condition, the same visual stimuli were used, but there was no correlation between them and auditory stimuli that preceded them. "X" and "O" were equally probable across trials.

Procedure

The experimental procedure was executed using Psyscope (Cohen, MacWhinney, Flatt, & Provost, 1993). Upon arrival to the laboratory, participants were instructed that they would be responding to visual stimuli as quickly and as accurately as possible. They

were further instructed that the beginning of a trial would be signified by the appearance of a crosshair in the center of the screen. Participants were warned that shortly afterwards, they would hear a warning sound. This sound was described to indicate the immediate appearance of an image flash on the screen. Participants' task was explained to be to indicate whether the visual stimulus was "X" or "O" as quickly and accurately as possibly by pressing appropriate button on the response box. Each participant was randomly assigned to either experimental or control condition.

Each trial started with 500 ms long exposure of a "+" centered in the middle of the screen. The fixation point remained on the screen throughout presentation of the warning sound. Immediately thereafter, one of the visual stimuli was presented for 200 ms (its position determined as described above, and its identity as an 'X' or an 'O' randomly determined). Response time and accuracy was recorded. Reaction time was measured from the onset of the visual stimulus to the button press.

Results

Our analyses included seven subjects in the experimental condition and nine subjects in the control condition. The average reaction time for each subject was calculated as an average over trials for which the participant correctly identified the visual stimulus. Outliers greater than two standard deviations from the average reaction time for subjects in a particular block were removed. This rule of thumb was violated occasionally when distribution of the data was extremely skewed and the rule suggested an unnatural split.

These results were then analyzed according to the block order (first, second or third) and type of visual stimuli presented (X vs. O). Results for one of the subjects in the experimental condition were discarded because his mean reaction time was two standard deviations above that of the other subjects in experimental group (his average reaction time was 516.016 while average reaction time in experimental group was 358.775 (59.015)).

There was no significant difference in response times across groups of subjects and or block position (1 vs. 2 vs. 3) depending on the kind of visual stimuli presented ('X' vs. O). However, there still was a trend for subjects to respond to X stimuli more quickly than O stimuli. These descriptive statistics are presented in Table 1.

For the rest of the analysis mean reaction time (averaged across 'X' and O) was used for all subjects.

A one-way analysis of variance showed no significant effect of experimental manipulation (mean reaction times for two groups across two blocks are 358.775 (59.0151) for experimental and 338.658 (42.662) for control group, $p > 0.18$, ns). Likewise, there was no significant effect when the blocks were analyzed separately, as shown in Table II.

Analysis of differences between blocks showed that there was no significant change in subjects' speed as they went through the experiment. These results are presented in Table III.

Errors

In reaction time experiments there is a trade-off between accuracy and speed. For example, it is possible that a small effect may be overridden by individual differences in

response strategy (accuracy versus speed) in a small sample. Therefore, we decided to look at the error rates (proportion of incorrect responses) of subjects in both groups. We did not find any significant difference in performance between the two groups, as can be seen in Table IV.

The same subject that was excluded from reaction time analysis was excluded from error analysis, this time due to no error (which is again an aberration compared to other subjects).

General Discussion

The purpose of our study was to examine whether participants may learn a correlation between auditory and visual stimuli without explicit feedback. The rationale behind this approach was to develop a method that ultimately might be implemented to examine a wide range of phenomena in auditory categorization, without reliance on explicit labeling typical of categorization studies.

Participants heard a warning sound on each trial. Immediately thereafter, a visual stimulus was flashed and participants responded with the identity of the stimulus as quickly and as accurately as possible. Unknown to the participants, the characteristics of the warning sounds were perfectly associated with the location of the visual stimuli on the screen. We expected our participants to master this correlation through their participation in the experiment, without explicit instruction or feedback. We predicted that, as a result of this learning, their reaction time to identification of the visual stimuli would decrease. Participants in the control condition responded in the same paradigm, except that the stimuli they saw and heard had no underlying correlation. As such, their

performance served as a comparison against which we compared performance of the experimental group.

Our preliminary results did not confirm this hypothesis. Performance of both experimental and control group subjects remained the same throughout the experiment, with no difference between the two groups. Although our sample size was very small, we know that lack of effect is not due to differences in error rates (for example, one could argue that although subjects in experimental group were equally quickly, the are more accurate than subjects in control group), because those, too, were equal in two groups. Nevertheless, there was considerable variability in individual reaction times in both groups. Therefore, any claim that there is no effect of experimental condition would be presumptuous at this point. Reaction time experiments often produce rather small, but reliable, average differences between control and experimental groups (e.g., Gonnerman, 1999). The small number of participants in this study may have masked our ability to observe what might be a rather small effect of learning. We intend to pursue this hypothesis by examining further participants.

Although there were no significant differences in subjects' responses to two kinds of visual stimuli, it seemed that a trend emerged in subjects to respond to "X" stimuli more quickly. This might be due to something intrinsic about that kind of visual stimuli, but also to the fact that labels on the response box were not counterbalanced, meaning that "O" button was on the left side for all subjects. However, this is unlikely to be related to lack of effect between experimental and control group, as the response box was the same in both. One plausible explanation for this effect would be, perhaps, that most of the subjects are right-handed, and were therefore quicker in pressing right (X) button.

As we believe that the method that we are proposing could open a new chapter in studies of general perceptual mechanisms and sound perception in particular, we would like to have more conclusive results about it. Therefore, new subjects are being recruited for the follow-up experiment that will have larger sample size and counterbalanced response-box labels (we will also record preferable hand for our subjects).

References

PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments, and Computers, 25(2), 257-271.

Patricia K. Kuhl, Karen A. Williams, Francisco Lacerda, Kenneth N. Stevens et-al. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255, 606-608.

Linda Polka, Janet F. Werker (1994). Developmental Changes in Perception of Nonnative Vowel Contrasts. Journal of Experimental Psychology 20 (20), 412-435.

L. E. Volaitis, J. L. Miller (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. Journal of Acoustical Society of America, 92, 723-735.

K. Johnson, E. Flemming, R. Wright (1993). The hyperspace effect: Phonetic targets are hyperarticulated. Language, 69, 505-528.

P. Iverson, P. K. Kuhl (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. Journal of the Acoustical Society of America, 97, 553-562.

O. Aaltonen, O. Eerola, Å. Hellström, E. Uusipaikka, A. Heikki Lang (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. Journal of the Acoustical Society of America, 101, 1090-1105.

Kelly Tremblay, Nina Kraus, Thomas D. Carrell, Therese McGee (1997). Central auditory system plasticity: Generalization to novel stimuli following listening training. Journal of the Acoustical Society of America, 102 (6), 3762-3773.

M. Cheour, R. Ceponiene, A. Lehtokoski, A. Luuk, J. Allik, K. Alho, R. Näätänen (1998). Development of language-specific phoneme representations in the infant brain. Nature Neuroscience, 1, 351-353.

Lotto, A. J. , Kluender, K. R., Holt, L. L. (1998). Depolarizing the perceptual magnet effect. Journal of the Acoustical Society of America, 103 (6), 3648-3655.

J. S. Logan, S. E. Lively, D. B. Pisoni (1991). Training Japanese listeners to identify /r/ and /l/: A first report. Journal of the Acoustical Society of America, 89, 874-886.

S. E. Lively, D. B. Pisoni, R. A. Yamada, Y. Tohkura, T. Yamada (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. Journal of the Acoustical Society of America, 96, 2076-2087.

J. E. Flege, O. S. Bohn, S. Jang (1997). Effects of experience on non-native speakers' production and perception of English vowels. Journal of the Acoustical Society of America, 25, 437-470.

J. R. Saffran, E. L. Newport, R. N. Aslin (1996). Word Segmentation: The Role of Distributional Cues. Journal of Memory and Language, 35, 606-621.

J. L. McClelland, A. Thomas, B. D. McCandliss, J. A. Fiez (in press). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. Cognitive Disorders: The Neurocomputational Perspective, Oxford University Press.

M. A. Pitt, J. M. McQueen (1998). Is compensation for coarticulation mediated by the lexicon? Journal of Memory and Language, 39, 374-370.

J. R. Saffran, E. K. Johnson, R. N. Aslin, E. L. Newport (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70, 27-52.

Lisbeth S. Fried, Keith J. Holyoak (1984). Induction of Category Distributions: A Framework for Classification Learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 234-257.

Figure Captions

Figure 1. The top row illustrates waveforms for two stimuli drawn from the 2-dimensional acoustic stimulus space used in this experiment. The bottom row shows spectrograms for the same two stimuli. The blue lines in the spectrograms illustrate the notches denoted NF1 and NF2.

Figure 2. A 2-dimensional acoustic stimulus space is defined by NF1 and NF2. Each 'X' on the figure represents a single sound stimulus. Note that there are two distinct "categories" in the acoustic space.

Figure 3. The 2-dimensional acoustic stimulus space was not evenly sampled in the experiment. Here, color depicts the number of times each stimulus was presented, with centroid stimuli presented more often than stimuli on the outskirts of the distribution. Specific presentation details are described in the text.

Figure 4. Average reaction times across blocks for control and experimental groups.

Figure 1

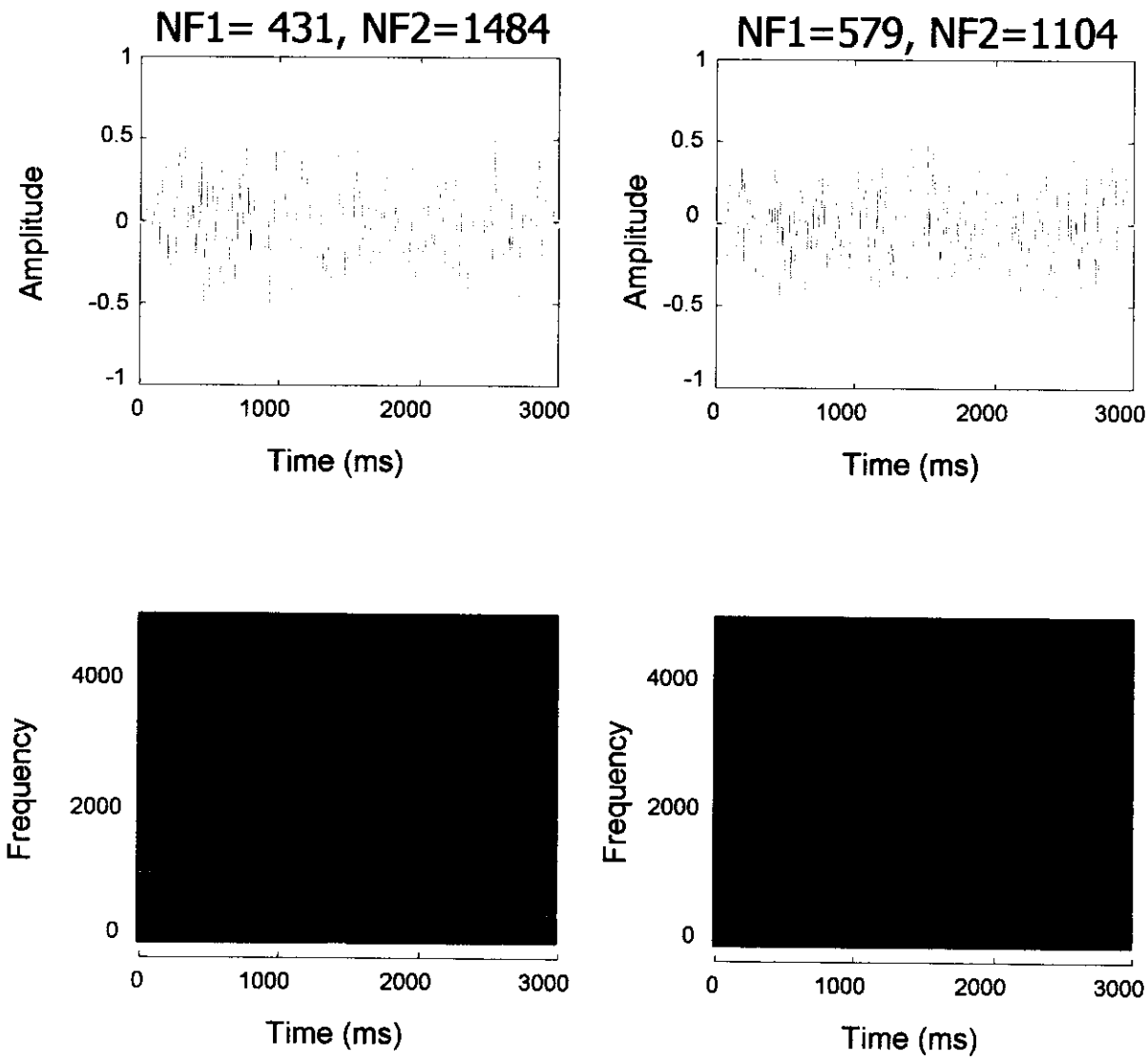
NF1= 431, NF2=1484
NF1=579, NF2=1104
Amplitude
Time (ms)
Frequency
1
0.5
0
-0.5
-1
4000
2000
0
1000
2000
3000

Figure 2

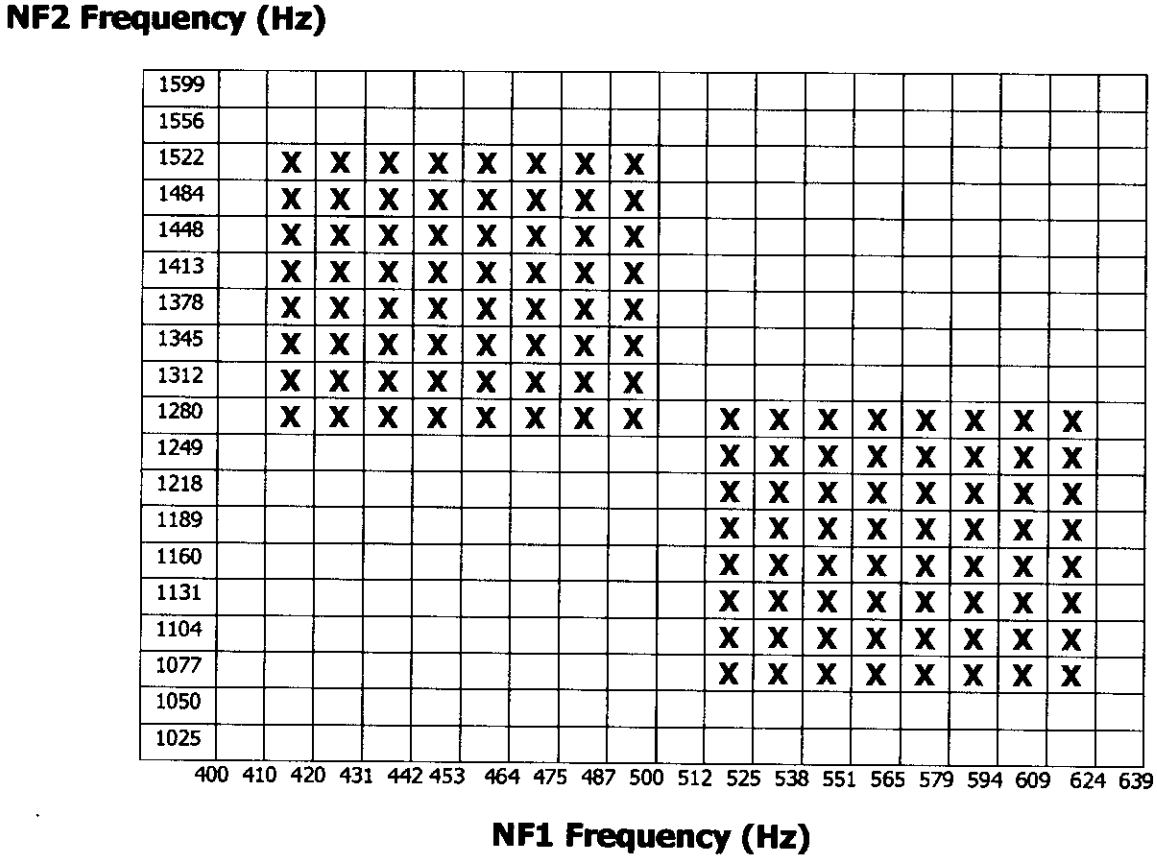
NF2 Frequency (Hz)
1599
1556
1522
1484
1448
1413
1378
1345
1312
1280
1249
1218
1189
1160
1131
1104
1077
1050
1025
400 410 420 431 442 453 464 475 487 500 512 525 538 551 565 579 594 609 624 639
NF1 Frequency (Hz)

Figure 3

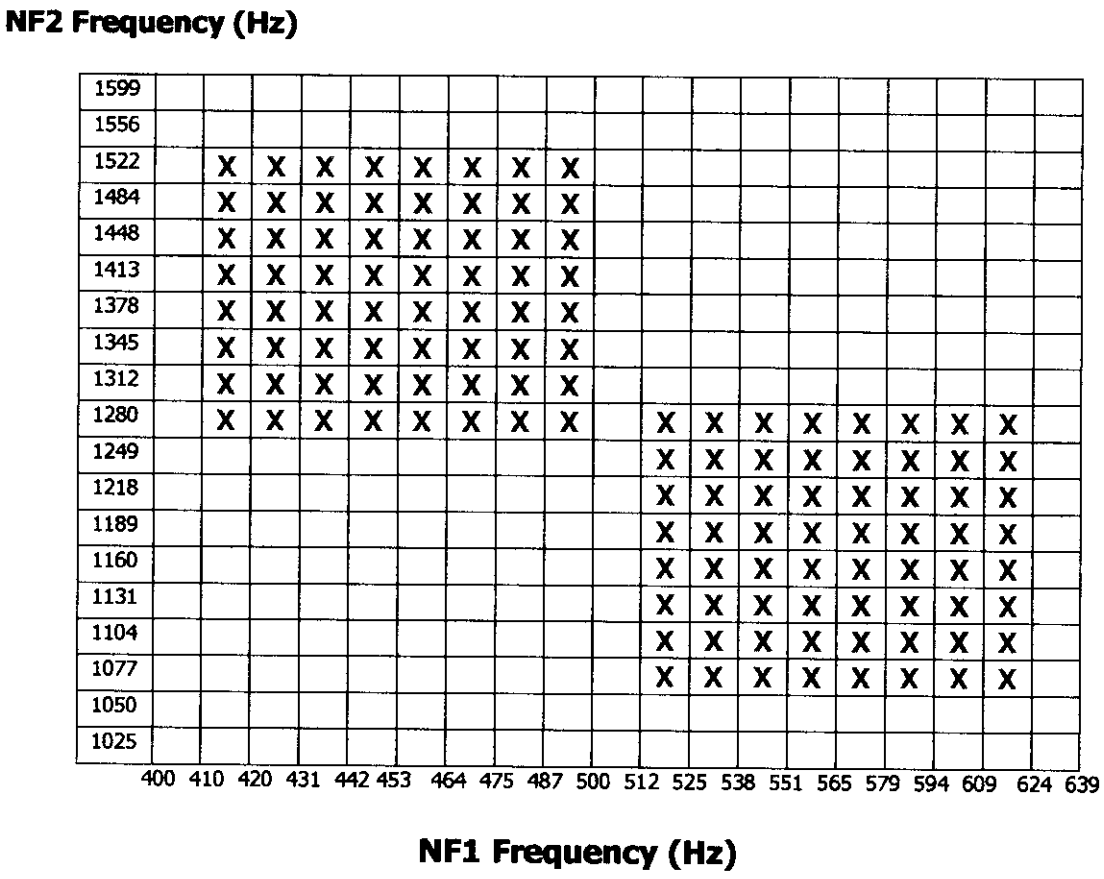
NF2 Frequency (Hz)
1599
1556
1522
1484
1448
1413
1378
1345
1312
1280
1249
1218
1189
1160
1131
1104
1077
1050
1025
400 410 420 431 442 453 464 475 487 500 512 525 538 551 565 579 594 609 624 639
NF1 Frequency (Hz)

Figure 4

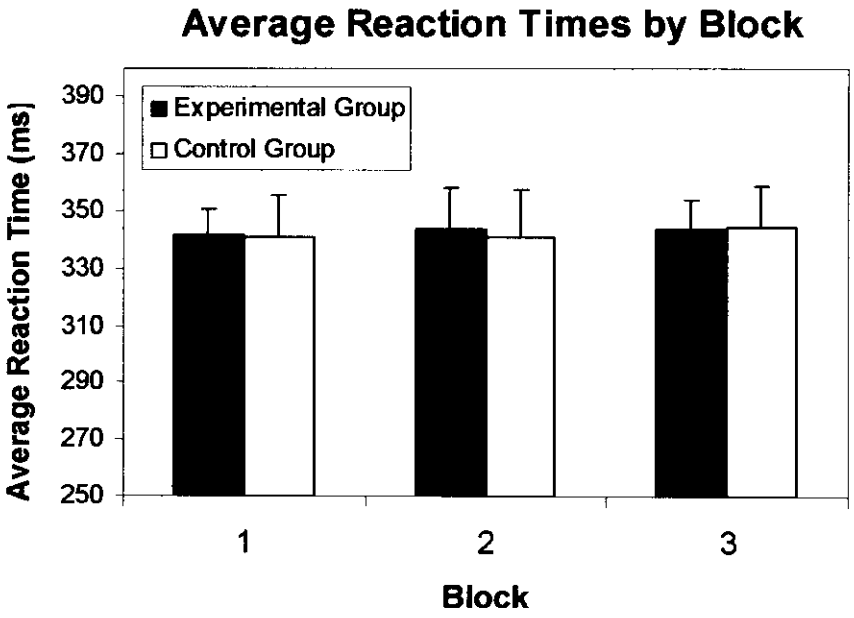
Average Reaction Times by Block
Experimental Group
Control Group
Average Reaction Time (ms)
390
370
350
330
310
290
270
250
1
2
3
Block

Table I

Experimental Group

	BLOCK 1	BLOCK2	BLOCK3
O	350.911 22.439	356.776 43.201	347.822 30.374
X	332.447 27.469	332.198 34.267	340.169 32.120

Control Group

	BLOCK1	BLOCK2	BLOCK3
O	343.363 50.45604	346.088 58.94829	337.854 53.15732
X	337.852 42.08809	335.721 44.07267	331.178 38.29953

Table II

	BLOCK1	BLOCK2	BLOCK3
Experimental group	**341.594** 24.196	**344.094** 37.630	**343.533** 28.866
Control group	**340.7073** 44.912	**340.858** 51.131	**334.410** 43.960
p-value	**>0.96**	**>0.89**	**>0.63**

Table III

	BLOCK1	BLOCK2	BLOCK3
Reaction Time	**341.095** 36.196	**342.274** 44.311	**338.059** 37.726
p-value	>0.93		

Table IV

	BLOCK1	BLOCK2	BLOCK3
Experimental Group	**19.857** 9.155	**23.714** 15.305	**26.667** 21.970
Control Group	**16.667** 13.509	**13.556** 9.606	**13.556** 10.525
p-value	**>0.58**	**>0.16**	**>0.22**