J. Neural Eng. 15 (2018) 055003 (13pp) https://doi.org/10.1088/1741-2552/aac663

Word recognition: re-thinking prosthetic vision evaluation

Shui'Er Han[1,2], **Cheng Qiu**[1,3], **Kassandra R Lee**[1], **Jae-Hyun Jung**[1] **and Eli Peli**[1]

[1] Department of Ophthalmology, The Schepens Eye Research Institute, Massachusetts Eye and Ear, Harvard Medical School, 20 Staniford Street, Boston, MA 02114-2500, United States of America
[2] School of Psychology, University of Sydney, Sydney, Australia
[3] Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States of America

E-mail: eli_peli@meei.harvard.edu

Received 5 February 2018, revised 30 April 2018
Accepted for publication 21 May 2018
Published 7 August 2018

Abstract

Objective. Evaluations of vision prostheses and sensory substitution devices have frequently relied on repeated training and then testing with the same small set of items. These multiple forced-choice tasks produced above chance performance in blind users, but it is unclear if the observed performance represents restoration of vision that transfers to novel, untrained items. *Approach*. Here, we tested the generalizability of the forced-choice paradigm on discrimination of low-resolution word images. Extensive visual training was conducted with the same 10 words used in previous BrainPort tongue stimulation studies. The performance on these 10 words and an additional 50 words was measured before and after the training sessions. *Main results*. The results revealed minimal performance improvement with the untrained words, demonstrating instead pattern discrimination limited mostly to the trained words. *Significance*. These findings highlight the need to reconsider current evaluation practices, in particular, the use of forced-choice paradigms with a few highly trained items. While appropriate for measuring the performance thresholds in acuity or contrast sensitivity of a functioning visual system, performance on such tasks cannot be taken to indicate restored spatial pattern vision.

Keywords: prosthetic vision evaluation, vision rehabilitation, pattern discrimination, object recognition

S Supplementary material for this article is available online

Introduction

Vision prostheses and sensory substitution devices (SSDs) are designed to provide blind users visual restoration or to establish a novel sense of vision. Vision prostheses can be implanted in the retina, relaying visual information through electrical stimulation of retinal neurons (Margalit *et al* 2002). Current retinal prostheses include the Argus II (Second Sight, Sylmar, CA) (Ahuja and Behrend 2013), the Alpha IMS (Retina Implant AG, Reutlingen, Germany) (Stingl *et al* 2015) and AMS (Stingl *et al* 2017), and the Bionic Eye Retinal Prosthesis (Bionic Eye Technologies Inc., Ithaca, NY) (Rizzo *et al* 2003) that is under development. Other visual, non-retinal prostheses, transmit electrical stimulation to the optic nerve or the early visual cortex (Fernandes *et al* 2012). Non-invasive SSDs provide non visual input through tactile sensations on the skin (Meers and Ward 2005, Ortiz *et al* 2011) or tongue (Grant *et al* 2016), or through auditory stimulations (Capelle *et al* 1998, Cronly-Dillon *et al* 1999, Hanneton *et al* 2010, Striem-Amit *et al* 2012a, 2012b, Stiles and Shimojo 2015).

Existing visual prosthetic systems provide only rudimentary capabilities (Shaw 2016) with substantial limitations; the main limitation is low spatial resolution. Retinal implants such as the Argus II has 10 × 6 electrodes (Ahuja and Behrend 2013) and the Alpha AMS has an array of 40 × 40 electrodes (Stingl *et al* 2017). Non-invasive options like the BrainPort

V100 SSD (Wicab, Middleton, WI) provide electrical stimulation to the tongue via 20 × 20 electrodes (Nau *et al* 2013). Some of these devices use a high resolution camera that provides a flexible field of view with zooming, but the low resolution of these devices limits the useful field of view (Jung *et al* 2015). Moreover, though as many as 10 perceived stimulus levels were reported in a tongue stimulation study (Lozano *et al* 2009), the number of meaningfully different stimulus levels (dynamic range) of these devices is highly limited, often binary (on and off) or just 3 to 4 levels (Chouvardas *et al* 2008, Stingl *et al* 2017). With these severe limitations, it is important to carefully design the evaluation process of these devices (Rizzo and Ayton 2014) so that we can obtain an accurate representation of their utility in real-world applications and guide further developments.

The primary purpose of evaluation is to determine if a device provides *pattern vision* (Caspi and Zivotofsky 2015) with similar spatial and temporal characteristics to the normal human visual system. A normal visual system is capable of learning spatial patterns and it can generalize from training to novel stimuli and contexts. Prior evaluation studies of visual prostheses tended to measure performance using multiple forced-choice tasks, during which subjects were required to discriminate among a few heavily pre-trained objects (De Neve 2011, Zrenner *et al* 2011, Humayun *et al* 2012, Ahuja and Behrend 2013, da Cruz *et al* 2013a, Bionic Vision Australia 2014, Nau *et al* 2014a, Stingl *et al* 2015, 2017, Grant *et al* 2016). Test objects were also typically presented over a clutter-free and high-contrast background, such as a banana on top of a black velvet cloth (Ahuja and Behrend 2013, Nau *et al* 2014a, 2014b, Stingl *et al* 2015, 2017, Edwards *et al* 2018). Under this paradigm, users of prosthetic systems can learn to develop strategies such as head motor scanning to achieve successful task performance, even in the explicit absence of visual spatial pattern data (Caspi *et al* 2009, Dorn *et al* 2013, Caspi and Zivotofsky 2015). These studies have demonstrated an ability to discriminate a few objects (e.g Ahuja and Behrend (2013) and Nau *et al* (2014a)) or word images (e.g Grant *et al* (2016)) after extensive training, but it remains unclear if the demonstrated capabilities transfer to novel stimuli or contexts. Therefore, it is necessary and important to determine what can and cannot be learned using repeated multiple-choice testing and training. This would provide a more accurate understanding of the extent of pattern vision capabilities in prosthetic vision systems.

Relying on multiple-choice tasks as an evaluation method is problematic because it, implicitly and sometimes explicitly, equates high task performance with the attainment of functioning pattern vision. For example, previous studies claimed successful ‘recognition’ or ‘identification’ of object and word stimuli with the Argus II retinal implant (da Cruz *et al* 2013a, 2013b), the Alpha IMS (Stingl *et al* 2015), and the BrainPort SSD (Nau *et al* 2014a). These conclusions, however, were based on performance measured with a few heavily trained items without testing post-training performance with novel stimuli. For example, Alpha IMS/AMS implant patients were trained with the same 4 objects over the course of a year in up to 7 follow-up visits (Stingl *et al* 2015). Designating performance in these studies as object/word ‘recognition’ is a misnomer, because recognition requires connecting a previously encountered stimulus with a new encounter of the same stimulus (Wilson 1995), such as in a new context or environment (DiCarlo *et al* 2012). It would be more appropriate to refer to performance on such repeated forced-choice tasks as discrimination. ‘Identifying’ a few highly trained items could be achieved through distinguishing low-level features, as directly demonstrated by the brightness-only mode in Caspi and Zivotofsky (2015), where the only available information was the brightness and size of the stimulation (no spatial details, such as orientation, were provided). Training *discrimination* may not generalize, even for the same objects in new contexts. For example, it has been shown that training *discrimination* among synthetic speech samples is more likely to enhance sensitivity to differences within the trained set rather than lead to transfer to natural speech perception (Jamieson and Morosan 1986). The difference between recognition and discrimination is not a mere matter of semantics or terminology, but rather an essential difference in the nature and requirements of the task. For further discussion on the need for distinction between visual recognition and discrimination, see van Meeteren (1995) and other chapters in the same book section.

Another example of the importance of distinction between discrimination and object recognition is illustrated in Nau *et al* (2014a). In that study, subjects were trained to use the BrainPort device to manually reach for a target object among a set of 4 items (softball, coffee mug, plastic banana, and a highlighter marker) placed on a uniform contrasting background. These same 4 objects were later used to evaluate performance in 5 testing sessions, each comprised of 20 trials conducted over the course of a year (Grant *et al* 2016). As the subjects were trained extensively with these 4 objects, the task was really a discrimination task among the 4 stimuli. Such testing provides little evidence that the response is reflective of pattern vision, or that the device has any properties of a functioning pattern vision system for object recognition.

Similarly, the use of forced-choice paradigms can give a false impression of object recognition with visual-to-auditory SSDs (Cronly-Dillon *et al* 1999, Auvray *et al* 2007, Striem-Amit *et al* 2012a). For example in Striem-Amit *et al* (2012a), low-level geometric features were explicitly coded with specific tone sequences. This allowed one subject to associate which tone sequence coded for a circular shape (interpreted in the forced-choice task as an open-mouth surprise) and which sequence represented a straight line (representing teeth and interpreted as a smiling face), as shown by their Supplementary movie (stacks.iop.org/JNE/15/055003/mmedia). Since only three facial expressions (anger, surprise, and smiling) were tested, discrimination among the different tone sequences would be sufficient for the subject to perform the task. Simple 3 tone discrimination could also explain the video’s demonstration that the subject was able to distinguish similar facial expressions acted by a different individual, therefore not demonstrating a transfer, as it is implied. With the small number of distinct facial expressions tested, high performance could be obtained without relation to the visual characteristics presented (in fact only two-tone patterns had to be discriminated from all others). This forced-choice testing paradigm does not prove or even suggest a transfer of the ability beyond the simple tone discrimination of the three tone sequences. The results cannot be interpreted as successful identification of facial expressions and they do not necessarily represent pattern vision capability.

Other studies have also evaluated vision prosthetic devices and SSDs with multiple-choice letter and word 'recognition' (da Cruz *et al* 2013a, Nau *et al* 2014a, Grant *et al* 2016). The BrainPort multiple-choice word recognition task via tongue stimulation is a good example of how low-level discrimination strategies could be applied to successful multiple-choice task performance. As the subjects were trained repeatedly and then tested with the same 10 words (Nau *et al* 2014a), they likely learned that there were 3 three-letter words, 4 four-letter words, and 3 five-letter words (figure 1(b)). Hence, without any other visual details, subjects could increase their chances of making a correct guess from 10% to 25%–30% by using just the word length as a cue. The word length could be determined using scanning if only one electrode of the 400 was active. In addition, the unique distribution of letters with ascenders and descenders within each word length class (figure 1(b)) could be used to further improve pattern discrimination. Subjects could combine the word length cue with the ascender/descender cues to make a correct word discrimination, requiring no recognition of any other letter or visual details (see figure 1(a)). Here, we tested the ability to effectively train visual pattern discrimination with such a multiple-choice task and asked if such training and improved performance of the trained words could produce generalizable outcomes with a fully functional human vision system, let alone via tongue stimulation.

Methods

In the pre-training test, we presented normally-sighted subjects with low resolution images (figure 1(a)) of the 10 trained words from the BrainPort studies (Grant *et al* 2016, Nau *et al* 2014a) and additional 50 words (untrained words). Subjects' recognition of the 60 words was measured, and no feedback was given in the pre-training test. We then trained the subjects on the 10 words and provided feedback about the accuracy of their responses and the identity of the presented word. Following the training, we again measured recognition accuracy of the same set of 60 words presented in the pre-training test.

Word selection

In addition to the 10 words used in previous BrainPort studies (Grant *et al* 2016, Nau *et al* 2014a), 50 commonly used, lower case words were selected from the english lexicon project (ELP) database (Balota *et al* 2007). Matching the proportions of word lengths in the set of 10 trained words, 40% of the 50 words had four letters, 30% had 3 letters and 30% were five-letters long. In addition, 10% of the 50 words had no ascender and/or descender features, similar to that of the set of 10 trained words (where only one word, 'moon' had no ascenders or descenders). Frequently encountered words were chosen, and the selection was based on the hyperspace analogue to language (HAL) frequency norms (Lund and Burgess 1996) provided by the ELP. The 50 words were selected using these norms, and word frequency was matched across the different word length groups. Specifically, the average log-transformed HAL norm was 11 ± 0.7 for three-letter words, 10.8 ± 0.6 for four-letter words, and 10.8 ± 0.4 for five-letter words.

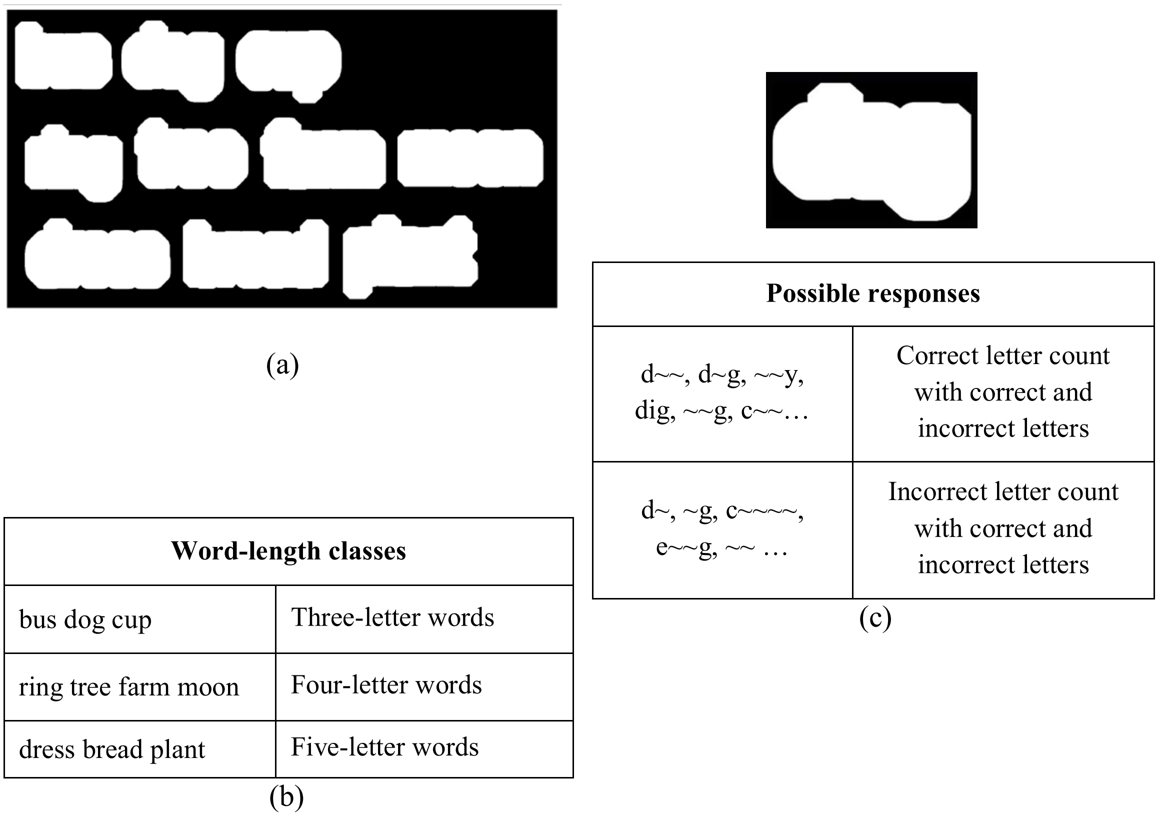
Simulated prosthetic word images

Lowercase words were set to Arial Narrow font[4] with a font size of 100. The words were converted to binary images (white text on black background). We then performed morphological dilation on the binary images in MATLAB (MathWorks, Natick, MA), using the iterative binary dilation method with a disk-shaped structuring element measuring 84 pixels in radius (Gonzalez *et al* 2009). The dilated and degraded words were displayed on an LCD monitor (ViewSonic 1920 × 1080 pixels, 23 inches) placed 60 cm away from the subject. With this distance and display resolution, the height of a degraded lowercase 'e' was 3.2° visual angle. Examples of the degraded word images are presented in figure 1(a) and the dimensions of the 60 word images are summarized in table 1. Welch two samples t-tests of the word length showed that four-lettered words were significantly wider than three-letters $t(39.8) = 3.7$, $p < 0.001$, but narrower than five-lettered words, $t(36.6) = 5.6, p < 0.00001$.

Procedure

Each trial began with an 800 ms central fixation cross that indicated the center of the word image. Twenty subjects (normally sighted or corrected-to-normal) were required to recognize, to the best of their ability, an English word made up of lower case letters and type their best guess of the word or its components. Subjects were asked to guess the number of letters represented by the word image and type that number of letters or symbols: using lowercase letters to represent recognizable letters of the word image, substituting the tilde symbol, ~, for letters they could not recognize (figure 1(b)). The use of image features such as ascender and descender was not explicitly mentioned in the instructions. Thus, the subjects' attention was explicitly directed to the word length cue (number of letters) but only implicitly to the ascender/descender cue. The maximum presentation duration for each word image was 1 min, however, subjects were instructed to begin typing as soon as they

[4] We opted for a different font type from previous BrainPort evaluation studies (Grant *et al* 2016, Nau *et al* 2014a). This was because a larger structuring element radius was required to generate the word envelope of the previously used, wider Century Gothic font, and that process obscured the ascender and descender features. Our approximation of the word envelope resembled the definition adopted by Bouma (1971), who studied the effect of ascenders and descenders on letters discrimination. He used the smallest enclosing polygon around a lower case letter without indentations (i.e. ignoring the inner gaps of letters *n*, *h* and *c*).



(a)

Word-length classes	
bus dog cup	Three-letter words
ring tree farm moon	Four-letter words
dress bread plant	Five-letter words

(b)

Possible responses	
d~~, d~g, ~~y, dig, ~~g, c~~…	Correct letter count with correct and incorrect letters
d~, ~g, c~~~~, e~~g, ~~ …	Incorrect letter count with correct and incorrect letters

(c)

Figure 1. Low-resolution representations of the 10 words used in this study. (a) Simulated word envelopes that show how easily discriminable the 10 word patterns are through the conjunction of word length and the locations of ascenders and descenders. (b) The three lengths classes of the 10 words. (c) Examples of possible responses to the low-resolution representation of the word *dog*. Subjects responded to each word image by typing a word or a string of letters. The tilde symbol, ~, was used to indicate positions of letters that were not recognizable.

Table 1. Image dimensions in visual angle (degree) for the set of 60 degraded words, sorted by word type and dimension of measurement.

Word type	Dimension	Degree of visual angle (*mean* ± SD)
Three-letters	Width	5.3° ± 0.6°
Four-letters	Width	6.1° ± 0.8°
Five-letters	Width	7.6° ± 0.8°
Ascender/ descender	Height	4° ± 0.3°
Non-ascender/ descender	Height	3.2° ± 0°

were ready to respond. Immediately as they began typing a response the word image would disappear from the screen, so subjects were instructed not to type their response until they had taken time to view the whole word. Each subject completed three experimental blocks; pre-training test, training, and post-training test.

Prior to the experimental blocks, subjects familiarized with the task demands by completing a practice block consisting of three word images that were not included in the experiment. The experiment then began with the pre-training test, where a total of 60 word images (including the 10 words to be trained from BrainPort studies) were presented in a random order. No feedback as to the accuracy of the response was given. Similar to previous BrainPort studies (Grant *et al* 2016), subjects repeatedly viewed the 10 words in the training sessions. A total of 15 repeats were conducted for a total of 150 trials. Feedback was provided during the training phase. The subjects were informed of the correct word presented following each incorrect response. The post-training test was then administered using the same 60 word images presented, as in the pre-training test (no feedback). The accuracy of subject responses, the words (or strings) keyed, and the time required to make a response were recorded for each trial.

Statistical analyses

Each subject's proportion of accurate whole-word responses (i.e. having same spelling as the stimulus presented) was first computed for each of the two word groups. The computed proportion was then evaluated against two types of predetermined success rates. The first type of success rate was used in previous BrainPort studies (Grant *et al* 2016), defined as correctly identifying at least 6 out of 10 words (or 14 out of the first and final 25 trials during the training sessions). This value was determined by setting the chance level at 55%, the midpoint between the presumed chance level of 10% (i.e. 1 out of 10 words) and 100% perfect recognition. The second type of success rate took into account word length cues, defining chance performance at 30% accuracy (i.e. average chance

Table 2. Statistical results comparing pre- and post-training recognition rates.

(a) Two-way repeated measures ANOVA (significant effects are **bolded**)			
Factor	***F***	**(df_1, df_2)**	***p***
Test (pre/post-training)	**200**	**(1,19)**	**<0.0001**
Word group (trained/untrained)	**165**	**(1,19)**	**<0.0001**
Test (pre/post-training) × word group	**171**	**(1,19)**	**<0.0001**
(b) Follow-up student's t-tests (significant effects are **bolded**)			
Comparison	***t***	***df***	**Holm–Bonferroni corrected *p***
Pre-/post-training (10 words)	**14.2**	**19**	**<0.0001**
Pre-/post-training (50 words)	**6.9**	**19**	**<0.0001**
Post-training improvements (10 and 50 words)	**13.1**	**19**	**<0.0001**

performance of all three word classes). Therefore, the mid-point between 30% and 100% would be 65%, which would require correctly typing 7 out of 10 words (or 17 out of the first and final 25 trials during the training sessions).

For untrained words, the success rate was defined as correctly typing more than 50% of the 50 words, which was the mid-point between the chance level approximately 0%[5] and perfect recognition. The same performance goal was used for both trained and untrained words, and it represented the minimum percentage of subjects who were required to achieve the individual success rates for each word group. We first computed the proportions of subjects who achieved success in each word group, and then compared the proportions with the one-sided, lower 97.5% Agresti-Coull confidence limit. The performance goal is considered achieved when the data's lower confidence limit is greater than 50%.

The accuracy of responses was also evaluated using parametric analyses. We first normalized each subject's data by computing the percentage of correct responses (i.e. recognition rate) within each word group. A two-way repeated measures ANOVA (pre- and post-training session × trained and untrained word group) was used to compute the effect of training on the recognition of trained and untrained words. The time course of training performance was also plotted using the accuracy rates of 6 consecutive bins (25 trials each, 150 trials in total). The progress of training was then statistically evaluated with a one-way repeated measures ANOVA. We also sorted the 60 words into words with (90%) and without (10%) ascender and descender features. As with word groups, we computed the percentage of correct responses within each feature group for each individual subject. Three-way repeated measures ANOVAs (test (pre/post-training) × feature group (Asc/Dsc/Non Asc/Dsc) × word group (trained/untrained)) were then performed on the resultant dataset to statistically calculate the influence of these features on word recognition/discrimination accuracy, reaction times (RTs), and word length estimation. Follow-up pairwise comparisons conducted after all parametric analyses were subjected to Holm–Bonferroni correction where necessary.

Subjects

Twenty normal or corrected-to-normal sighted subjects (15–51 years of age, 5 females) participated in the study. All subjects gave written informed consent; two subjects were under 18 years of age and parental consent was obtained for those. Experiments accorded with the Declaration of Helsinki and the protocol and written consent were approved by the Massachusetts Eye and Ear Human Studies Committee.

Results

Overall performance between pre- and post-training sessions

Success rates and performance goals. Post-training, all of the twenty subjects were able to recognize 6 or more of the 10 trained words, resulting in a 97.5% lower one-sided bound of 81%, thus achieving both performance goals set in prior studies. Only 14 of the 20 subjects were able to recognize 7 or more of the 10 trained words, producing a 97.5% lower limit of 47.9%, thus falling slightly short of the higher performance goal we selected. In contrast, when presented with the untrained words for the second time in the post-training session, none of the subjects were able to recognize more than 50% of the untrained words. As a result, the performance of the untrained words produced a lower bound of −0.03%, failing to meet the second performance goal.

Overall recognition rate. A two-way repeated measures ANOVA (test pre/post-training × word group (trained/untrained)) was conducted on the dataset, and the results of the statistical analyses are summarized in table 2. Test session and word group had significant main effects on recognition rates. Both factors also interacted significantly, characterized by a greater increase in post-training accuracy for the 10 trained words than the 50 untrained words. Specifically, average accuracies on the 10 trained words increased from 13.5% ± 14.2% (maximum = 50%) in pre-training session to 82% ± 17% (maximum = 100%) in the post-training session. The performance also improved for the other 50 words, but average accuracies went from 14% ± 8.4% (maximum = 30%) to only 24.1% ± 8.5% (maximum = 36%) after training, this modest effect was statistically significant (table 2(b)). These effects are shown in figure 2(a).

[5] The chance level was set to 0% for the untrained words because subjects had no prior experience with these word images and were not given feedback for the respective word identities. Even if subjects learnt during training that there are three, four, and five letters words, the chance level remains very close to 0% (e.g. probability of about 26^{-3} for recognizing a three-lettered untrained short word accurately).

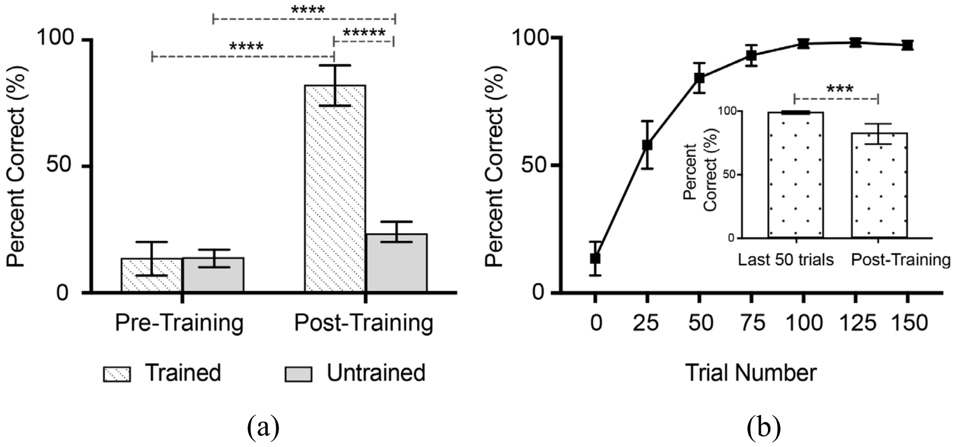


Figure 2. Effect of training on word 'recognition'. (a) Training of the 10 words produced significantly higher percent correct for both the trained and untrained words, and the improvement was substantially and significantly larger for the trained words. (b) Discrimination performance with the 10 words improved rapidly and plateaued after the first 75 trials. Percent correct of the 10 trained words was higher in the final 50 trials of the training sessions than the post-training sessions (inset). Asterisks denote statistical significance after Holm–Bonferroni correction ($\ast\ast\ast$: $p < 0.001$, $\ast\ast\ast\ast$: $p < 0.0001$, $\ast\ast\ast\ast\ast$: $p < 0.00001$). Error bars indicate 95% confidence intervals of the mean.

Performance during the training session

Figure 2(b) shows the effect of training, computed by binning the accuracy rate in every 25 consecutive trials of the training sessions. Starting from an average of 13.5% ± 14.2% in the pre-training test session (shown in both figures 2(a) and (b)), the accuracy rate of the 10 words increased rapidly to 58% ± 19.8% in the first 25 trials of the training sessions, eventually averaging at 97.5% ± 2.4% for the last 50 trials. These trends were statistically evaluated with a one-way repeated-measures ANOVA, which showed that subjects were making significant progress throughout the training sessions, $F(5,95) = 63.2$, $p < 0.0001$. However, performance on the 10 words fell significantly to an accuracy of 82% ± 17% in the post-training test, where the 10 words were mixed with the 50 untrained words, $t(19) = 4.1$, $p < 0.001$.

We also evaluated training performance with pre-determined success rates and performance goals. During the first 25 trials of the training sessions, 11 out of the 20 participants were able to make correct responses in at least 14 trials, producing a one-sided, lower 97.5% Agresti-Coull confidence limit of 39%. With extensive training, all of the participants were able to respond correctly for at least 14 of the final 25 training trials, giving rise to a lower bound of 81%. Taking the use of word length cues into consideration, we re-computed the proportion of subjects who had met the second success goal (i.e. more than 17 correct responses out the first and final 25 trials of the training session). Only 5 participants were able to achieve the success goal in the first 25 trials, resulting in a lower bound of 18%. After training, all participants had more than 17 out of the final 25 trials correct, producing a lower bound of 81%.

Effect of ascender and descender features

Effect of word features on accuracy. Figure 3 shows the effect of ascender and descender (henceforth Asc/Dsc) features on percent correct (10 words in figure 3(a) and the other 50 words in figure 3(b)). A three-way repeated measures ANOVA (2 feature group (Asc/Dsc / Non Asc/Dsc) × 2 test (pre/post-training) × 2 word group (trained/untrained)) was conducted on the dataset, resulting in a significant interaction between feature group and word group (table 3). That is, collapsed across pre- and post-training sessions, Asc/Dsc features produced statistically higher accuracy rates in the other 50 words, but not when subjects had extensive experience with the 10 trained words. As shown in figure 3(b), untrained words with Asc/Dsc were generally responded to with greater accuracy than words without, producing respective averages of 14.8% ± 9.1% and 7% ± 9.8% in the pre-training session, and 25.3% ± 8.9% and 13% ± 16.3% in the post-training test session. Further underscoring the importance of these features, significant post-training improvements of the 50 untrained words were only observed in those with Asc/Dsc features. In contrast, Asc/Dsc features did not facilitate discrimination of the trained words. Pre-training recognition accuracy rates were not different between words with and without Asc/Dsc features, 13.3% ± 14.2% and 15% ± 36.6% respectively. After extensive training, subjects were more accurate in discriminating the only trained word without Asc/Dsc (i.e. *moon*), scoring an average of 95% ± 22.4% as compared to 80.6% ± 18.3% for words with Asc/Dsc.

Effect of word features on reaction time. Figure 4 presents the effect of Asc/Dsc features on reaction time (RT). To include sufficient data for the analysis, both inaccurate and

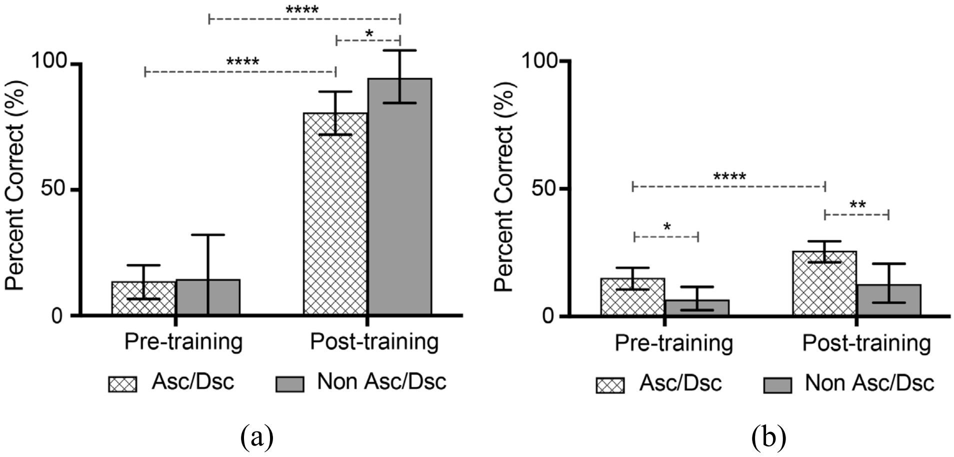


Figure 3. Effect of Asc/Dsc features on the percent correct of (a) the 10 trained words and (b) the 50 untrained words. Regardless of word feature, subjects were better at recognizing trained words post training. There was also a slight advantage for the only one non-Asc/Dsc trained word (*moon*). Only untrained words with Asc/Dsc features had a significant increase in accuracy after training. Asterisks denote statistical significance after Holm–Bonferroni correction ($*$: $p < 0.05$, $**$: $p < 0.01$, $***$: $p < 0.001$, $****$: $p < 0.0001$). Error bars represent 95% confidence intervals about the mean.

Table 3. Statistical results comparing accuracies for words with and without Asc/Dsc.

(a) Three-way repeated measures ANOVA (significant effects are **bolded**)			
Factor	***F***	**(*df*$_1$, *df*$_2$)**	***p***
Test (pre/post-training)[a]	**205.8**	**(1,19)**	<**0.0001**
Word group (trained/untrained)[a]	**147**	**(1,19)**	<**0.0001**
Feature group (Asc/Dsc/Non Asc/Dsc)	0.2	(1,19)	=0.66
Test × Word group[a]	**162.7**	**(1,19)**	<**0.0001**
Feature group × Test	0.5	(1,19)	=0.48
Feature group × Word group	**13.5**	**(1,19)**	<**0.01**
Feature group × Word group × Test	1.8	(1,19)	=0.20
(b) Follow-up Student's t-tests comparing feature groups within each test session (significant effects are **bolded**)			
Session	***t***	***df***	**Holm–Bonferonni corrected *p***
Pre-training (50 words)	**2.8**	**19**	<**0.05**
Post-training (50 words)	**3.4**	**19**	<**0.01**
Pre-training (10 words)	0.2	19	=0.84
Post-training (10 words)	**2.5**	**19**	<**0.05**
(c) Follow-up Student's t-tests comparing pre-/post training accuracies for each feature group (significant effects are **bolded**).			
Feature group	***t***	***df***	**Holm–Bonferonni corrected *p***
Asc/Dsc (in 50 words)	**6.9**	**19**	<**0.0001**
Without Asc/Dsc (in 50 words)	1.5	19	=0.16
Asc/Dsc (in 10 words)	**12.4**	**19**	<**0.0001**
Without Asc/Dsc (in 10 words)	**8.7**	**19**	<**0.0001**

[a] These results duplicated the corresponding results reported in table 2(a) with the two-way ANOVA.

accurate word recognition responses were included in the dataset. A three-way repeated measures ANOVA (2 feature group (Asc/Dsc / Non Asc/Dsc) × 2 test (pre/post-training) × 2 word group (trained/untrained)) was conducted, and the results are summarized in table 4. As with recognition accuracy, we found no significant effect of feature group. Feature group neither interacted with word group nor test. There was also no significant three-way interaction among all three factors. Collapsed across feature group, however, we found significant main effects of test and word group on RTs. Both factors also interacted significantly. Specifically, comparable RTs were obtained in the pre- and post-training sessions for untrained words (16.2s ± 8s to 12.5s ± 8.7s respectively). RTs for the 10 words, on the other hand, were significantly reduced from 18.9s ± 12.7s to 2.8s ± 1.7s from the pre- to post-training sessions.

Effect of Asc/Dsc features on the estimation of letter count. Asc/Dsc features indicate the letters' locations and may be useful in the estimation of letter count. A three-way repeated measures ANOVA (2 feature group (Asc/Dsc / Non Asc/Dsc) × 2 test (pre/post-training) × 2 word group) (trained/untrained) was conducted and the results are summarized in table 5. We found significant main effects of test and word group, but not feature group. Word feature neither

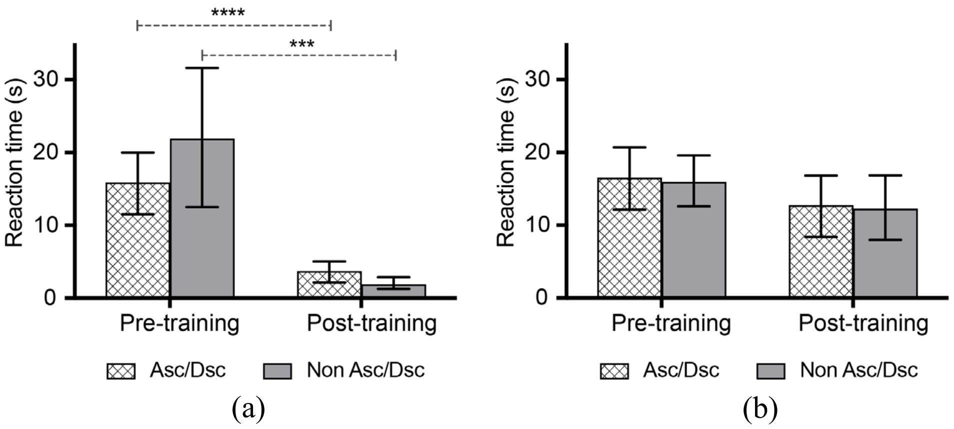


Figure 4. Effect of Asc/Dsc features on RTs for the (a) 10 trained and (b) 50 untrained words. After training with the 10 words, subjects responded faster to the trained words. The reduction of response time was not significant in the untrained words, indicating that the reduction in the time to respond was specific to the trained stimuli. Asterisks represent significance level after Holm–Bonferroni correction ($***$: $p < 0.001$, $****$: $p < 0.0001$). Error bars represent 95% confidence intervals about the mean.

Table 4. Statistical results comparing RTs for words with and without Asc/Dsc.

(a) Three-way repeated measures ANOVA (significant effects are **bolded**)

Factor	F	(df_1, df_2)	p
Test (pre/post-training)	**29.3**	**(1,19)**	**<0.001**
Word group (trained/untrained)	**14.7**	**(1,19)**	**<0.01**
Feature group (Asc/Dsc / Non Asc/Dsc)	1.4	(1,19)	=0.25
Training × Word group	**18.2**	**(1,19)**	**<0.001**
Feature group × Test	2.2	(1,19)	=0.15
Feature group × Word group	1.2	(1,19)	=0.29
Feature group × Word group × Test	2.9	(1,19)	=0.10

(b) Follow-up Student's t-tests comparing pre-/post-training RTs for each word group (significant effects are **bolded**)

Word group	t	df	Holm–Bonferonni corrected p
50 untrained words	1.9	19	=0.08
10 trained words	**6.2**	**19**	**<0.0001**

interacted significantly with word group nor with training, but there was a significant three-way interaction among test, word group and feature group. As shown in figure 5, the presence of Asc/Dsc features resulted in poorer post-training performance for the 10 words (86.1% ± 13.4% compared to 100% ± 0%) but more accurate letter count estimation for the other 50 words (73.2% ± 6.8% compared to 56% ± 19%). Note, however, there was only one non-Asc/Dsc word (*moon*) in the 10 words group. Post-training performance also increased significantly for both feature groups for the 10 trained words. This was not observed for the other 50 words, where performance was only significantly improved in the presence of Asc/Dsc features. Results of these statistical analyses are summarized in table 5.

Discussion

We demonstrated that subjects performed significantly and substantially better when discriminating a few highly trained low-resolution word shapes: correctly discriminating 82% of the 10 words improving from an initial 14% on the same words before training. Subjects' discrimination accuracy was even higher at the end of the training session (97.5%), during which the 10 words were not mixed with the 50 untrained words. However, the magnitude of performance improvements in trained words did not transfer to the recognition of untrained low-resolution words, where there was a mere 10% increase in correct responses. The poorer performance for untrained words was found despite the task's explicit and implicit directions to the available low level cues, such as the length of the words and the presence of ascenders and descenders, respectively. Performance on untrained words with Asc/Dsc was consistently higher than words without these features (figure 3(b)). These features were also useful for estimating letter count (figure 5), suggesting that these low level cues may account for the small but statistically significant improvement noted in recognizing the untrained words. Importantly, even with a normal visual system capable of learning and transfer (normally sighted subjects were tested), repeated multiple-choice training did not result in learning transfer.

These findings support our contention that low level discrimination can produce dramatic improvements in a multiple-choice task without generalization to novel objects,

Table 5. Statistical results comparing letter count estimation for words with and without Asc/Dsc.

(a) Three-way repeated measures ANOVA (significant effects are **bolded**)

Factor	F	(df_1, df_2)	p
Test (pre/post-training)	**72.2**	**(1,19)**	**<0.0001**
Word group (trained / untrained)	**13.7**	**(1,19)**	**<0.01**
Feature group (Asc/Dsc / Non Asc/Dsc)	1.4	(1,19)	=0.25
Training × Word group	**39.4**	**(1,19)**	**<0.0001**
Feature group × Test	0.9	(1,19)	=0.37
Feature group × Word group	2.2	(1,19)	=0.15
Feature group × Word group × Test	**9.8**	**(1,19)**	**<0.01**

(b) Follow-up two-repeated measures ANOVA (pre-/post-training × feature group (significant effects are **bolded**)

10 words				**50 words**			
Factor	F	df	p	Factor	F	df	p
Test (pre/post-training)	**70**	**19**	**<0.0001**	**Test**	**5.3**	**19**	**<0.05**
Feature group (Asc/Dsc / Non Asc/Dsc)	0.002	19	=0.96	**Feature group**	**4.7**	**19**	**<0.05**
Test × Feature group	**4.7**	**19**	**<0.05**	**Test × Feature group**	**9.7**	**19**	**<0.01**

(c) Follow-up Student's t-tests (significant effects are **bolded**)

10 words				**50 words**			
Comparison	t	df	**Holm–Bonferroni corrected** p	**Comparison**	t	df	**Holm–Bonferroni corrected** p
Pre-/post-training (Asc/Dsc)	**7.4**	**19**	**<0.00001**	**Pre-/post-training (Asc/Dsc)**	**4.7**	**19**	**<0.001**
Pre-/post-training (Without Asc/Dsc)	**5.3**	**19**	**<0.0001**	Pre-/post-training (Without Asc/Dsc)	0.2	19	=0.83
Pre-training (with, without Asc/Dsc)	1.2	19	=0.24	Pre-training (with, without Asc/Dsc)	0.3	19	=0.75
Post-training (with, without Asc/Dsc)	**4.6**	**19**	**<0.001**	**Post-training (with, without Asc/Dsc)**	**3.8**	**19**	**<0.01**

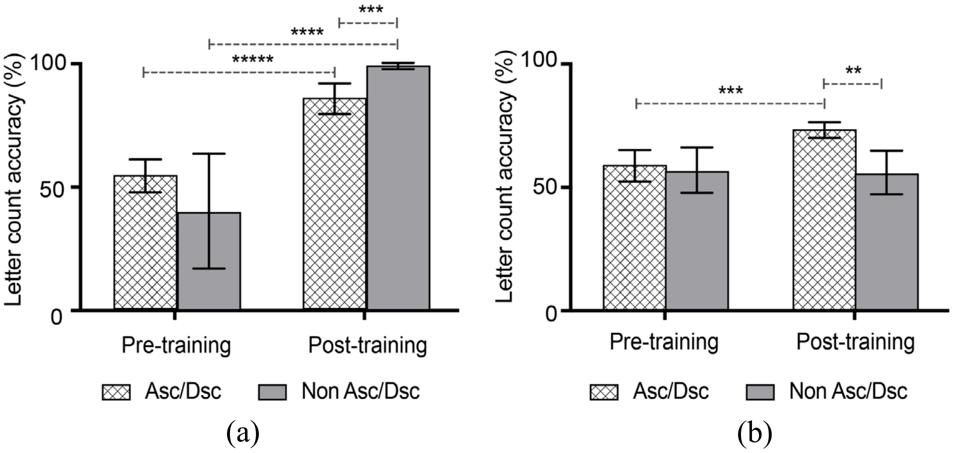


Figure 5. Effect of Asc/Dsc features on letter count accuracy for (a) 10 trained and (b) 50 untrained words. As with the other two measures, subjects were better at letter counting in the trained words. The slight advantage for the non-Asc/Dsc single word was also observed in letter counting. On the other hand, subjects were only better at estimating word lengths of the untrained words when the words did contain Asc/Dsc features. Asterisks represent significance level after Holm–Bonferroni correction (∗∗: $p < 0.01$, ∗∗∗: $p < 0.001$, ∗∗∗∗: $p < 0.0001$, ∗∗∗∗∗: $p < 0.00001$). Error bars represent 95% confidence intervals about the mean.

even those from the same limited category as the trained stimuli. Therefore, one should not regard high performance on a multiple-choice task as proof of a functioning pattern vision system. In fact, the discrimination performance can be hindered when the trained items are presented together with untrained items. This was demonstrated in our results, which showed lower accuracy scores on the 10 words in the post-training session compared to the end of the training session (figure 2(b) inset). There is evidence that increasing the number of training items could increase generalizability, as demonstrated in the tactile recognition of alphanumeric symbols (Arnold and Auvray 2018). Since all English words are constructed from combinations of 26 letters, it is possible that larger stimulus sets may be a practical approach to train reading. With sufficient training, pattern discrimination of all 26 letters may generalize and transfer to word reading. This was demonstrated with an auditory prosthetic used with the Hebrew alphabet (Striem-Amit *et al* 2012b). Interestingly, training on the words, both with the BrainPort (Grant *et al* 2016) and in our study, did not result in letter recognition being the underlying learned skill. Rather, discriminating the shape and features of the whole trained words appear to be the strategy used by subjects. In any case, reading is not the intended use of these devices, as access to text is available and more efficient with other technologies (e.g. text-to-speech; Sorin *et al* 2014). Instead, the intended use of vision prostheses is to allow recognition of many other objects. Training discrimination with letters in clutter-free testing situations, even when successful, is unlikely to generalize and transfer to object recognition in activities of daily living, though it may provide some limited information on the performance of the system.

Similar limitations apply to the study by Grant *et al* (2016), who tested the use of the BrainPort in locating informational signs on the walls of a hallway. The authors referred to the task as an 'orientation and mobility (O&M)' task. The four signs used had symbols marked on them, which the subjects were asked to find. However, the signs themselves had distinct shapes; the Danger (!) sign was triangular in shape, the Stairs sign was a square, and the Men and Women bathroom symbols were round signs. These different shaped signs were dark and were placed on a light wall. Similar to the word 'recognition' tasks, subjects were trained extensively and tested with these same 4 signs five times over the course of a year. Discrimination of the dark sign outlines was, therefore, a viable strategy for performing the task, yet our current study results would suggest that generalization to novel sign shapes or content is unlikely. The relation of such a task to O&M is limited to the ability to locate a single dark sign with a distinct shape on the bright wall, hardly an important aspect of navigation or safe mobility.

Placing emphasis on the performance in forced-choice tasks may inadvertently enable the use of head motor tracing or scanning strategies in vision prostheses. Motor tracing can allow users to extract more information with the low spatial resolution of current vision prosthetic devices, but such results can also lead to incorrect conclusions, such as the interpretation of task success as spatial visual capability. Given the prostheses' limited field of view, users may learn to use head movements to scan along or trace across the contours of objects. Indeed, Caspi and Zivotofsky (2015) found that subjects were able to discriminate low resolution, non-patterned images of the four alternative Landolt C acuity targets using head scanning or tracing. Motor tracing may also be used to produce correct word recognition responses, since the use of large font sizes (da Cruz *et al* 2013a, Grant *et al* 2016) or the use of the prosthesis camera's zoom capability (Nau *et al* 2014a, Grant *et al* 2016) allows motor tracing along the high contrast letter strokes (da Cruz *et al* 2013a, 2013b). While the subjects could not trace the word stimuli in our study, our word stimuli did contain information that was obtainable by motor tracing in the original BrainPort studies (i.e. the word envelope). Thus, our results and those of Caspi and Zivotofsky (2015) show that the low level information available from motor tracing (word length and ascender/descender, and direction of a Landolt C, respectively) could be used to train discrimination in multiple-choice tasks. But since recognition did not transfer to the other 50 untrained word images with similar tracing information available, our results suggest that the utility of motor tracing may be limited to the discrimination of trained stimuli only.

With simple stimuli and clutter-free environments, a prosthetic user could use lateral head movements to scan the prosthesis' narrow field of view across a high contrast line on the ground (da Cruz *et al* 2016; see also use of BrainPort in De Neve (2011); bionic eye prototype in Bionic Vision Australia (2014)). This form of motor scanning may be used in obstacle courses constructed of dark objects placed on a light colored floor (Geruschat *et al* 2012). Caspi *et al* (2018) showed how a patient with the Argus II could trace a curved line in a parking lot despite being trained with straight lines. Though this observation may be considered an indication of learning transfer, the difference between tracing straight and curved lines is really a slight variation on the same task in the narrow field of view. Author EP has dubbed this type of tracing/detecting technique Radar Vision (Jung *et al* 2015). Like radar scanning, it can provide the user the direction of the contrasting line to be traced or of the obstacle to be avoided. Even more limited than radar, however, current devices do not provide information about distance. The direction of the contrasting line or obstacle is not provided by an analysis of spatial image but rather by the synchronization of the signal (radar beep or return like signal) to the motor direction of the scanning head. Similar to radar, this form of motor scanning is effective in non-clutter environments such as the sky or the ocean, but not in urban, cluttered environments. More importantly, this scanning technique does not necessarily involve spatial properties that constitute a functioning pattern vision system. As long as the evaluation task involves simple, clutter-free stimuli and multiple-choice testing, motor tracing can be used successfully, even with a single pixel sensor that provides some type of signal (audio or tactile) when aimed at a contrasting line or obstacle. Single pixel sensors may be a useful aid for blind users, but their utility should not to be confused with the vision restoration expected and desired from retinal or cortical implants.

Multiple-choice testing can be used to determine the resolution limits of visual prostheses. Striem-Amit *et al* (2012a) tested the visual-to-auditory SSD (vOICe) with a 4-alternative forced choice (AFC) tumbling E acuity test and found a high-resolution limit (i.e. 20/360). Labeling the measured resolution limit 'functional resolution' or 'functional acuity', the authors applied this level of sampling to natural scenes and obtained images that were of unreasonably high quality (shown in figure 1(E) of their study). Similarly, Caspi and Zivotofsky (2015) found that head scanning strategies allowed subjects to perform better on a 4-AFC Landolt C acuity test than expected by the geometrical resolution, defined as the angular distance between a pair of electrodes in the retinal implant array. These authors did not present their measured resolution limits with images, but they also argued that the results represented the user's 'recognition acuity', even maintaining that the higher measured acuity demonstrated spatial pattern perception. We would caution against this interpretation. Natural vision is equipped with a much wider dynamic range than current vision prostheses, and it is well known that gray scale dynamic range and resolution may be traded for each other. For example, high resolution is traded for wider dynamic range in halftone printing (Peli 1991), and anti-aliasing of images involves trading dynamic range for higher resolution (Getreuer 2011). If the natural scene images in Striem-Amit *et al* (2012a) were presented at a lower dynamic range (i.e. binary, as may be expected with the Argus II), the estimated perceived quality would be severely reduced (see figure 6). Hence, the higher measured resolution limit cannot serve as evidence of spatial pattern perception with vision prostheses.

To improve the way that we discuss the capabilities of visual prostheses, the language used to describe the change in performance with commonly used evaluation tasks should be reconsidered. Stingl *et al* (2015) and Humayun *et al* (2012) noted significant improvements with the Alpha IMS and statistically better performance with the Argus II, respectively. However, improvements are difficult to quantify or measure, as subjects had little to no light perception without the device. Similarly, Grant *et al* (2016) reported 'improvements' from a baseline of subjects without the use of any assistive device, and the Argus II performance was reported from the baseline of subjects with the device switched off (Humayun *et al* 2012). In addition to the fact that any performance compared to consistent zero performance at baseline will be statistically significant, the use of the term is misleading as it implies existing visual capabilities, be it before or after the use of prosthetic devices. Note that this is just a problem of reporting language and of the interpretations that might follow. Many of these studies properly analyzed the effects by setting threshold goals that subjects had to meet (Grant *et al* 2016), and analyzed the data to determine if the goal was achieved.

We have pointed out the limitations of the multiple-choice testing paradigm in evaluating visual prostheses. While there is nothing wrong with the testing procedures per se, the interpretation of the results as a proof of visual restoration is not supported. What this area of research needs is testing methods that probe the spatial visual capabilities of the prosthesis system directly, demonstrating restoration of visual capabilities. In the process of developing such methods and using them, we will uncover limitations of current systems and hopefully this will guide us towards design changes overcoming these limitations. By considering the physical parameters of the prosthetic system, multiple-choice testing could be useful in determining the thresholds of specific parameters. This was demonstrated in Caspi *et al* (2009), where the resolution limit of a low resolution prosthetic was found to be consistent with the distance between the electrodes. As we have argued repeatedly, task performance alone cannot serve as direct proof for success in restoring vision, evidenced by the higher 'resolution' limit obtained in Caspi and Zivotofsky (2105) through multiple-choice testing. Nevertheless, we remain optimistic that a variant of multiple- choice testing may be developed. One possible approach was demonstrated by Striem-Amit *et al* (2012b), who tested the discrimination of images viewed as soundscape with their auditory prostheses. Instead of just training the discrimination of a few objects, they had the subject discriminate objects taken from seven different categories. Training subjects to recognize the categories and testing them with previously unseen samples from the same categories may provide evidence for generalization and transfer. However, care has to be taken to assure that the categories chosen are not discriminable by incidental low level features such as the word length in our study.

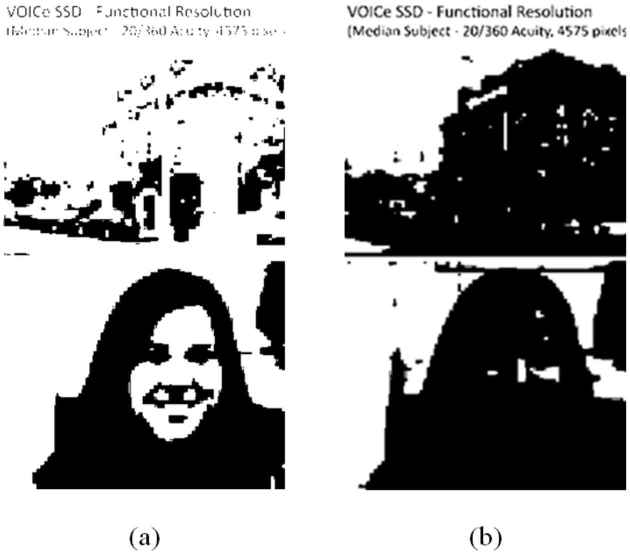

VOICe SSD Functional Resolution
(Median Subject - 20/360 Acuity, 4575 pixels
VOICe SSD - Functional Resolution
(Median Subject - 20/360 Acuity, 4575 pixels
(a)
(b)

Figure 6. Effect of dynamic range on image quality demonstrated with functional acuity derived from 4AFC measure of tumbling E task. The full gray scale wide dynamic range images were taken from figure 1(E) of Striem-Amit *et al* (2012a) and represent the functional resolution of persons using the vOICe visual-to-auditory SSD. Reducing the dynamic range to a 1-bit format as we have done here severely reduces the image quality of both face and street scene images. (a) 1-bit binary with half the pixels black (b) 1-bit binary image with 75% of the pixels set to black. Original image reproduced from Striem-Amit *et al* (2012a), CC BY 4.0.

Acknowledgments

This work was supported in part by the Office of the Assistant Secretary of Defense for Health Affairs, through the Clinical and Rehabilitative Medicine Research Program. Joint Program Committee 8, Neurosensory and Rehabilitative Research Program under Award No. W81XWH-16-1-0033. Opinions, interpretations, conclusions, and recommendations

are those of the author and are not necessarily endorsed by the Department of Defense.

ORCID iDs

Shui'Er Han https://orcid.org/0000-0002-2521-1416
Eli Peli https://orcid.org/0000-0002-1340-9257

References

Ahuja A K and Behrend M R 2013 The Argus™ II retinal prosthesis: factors affecting patient selection for implantation *Prog. Retinal Eye Res.* **36** 1–23

Arnold G and Auvray M 2018 Tactile recognition of visual stimuli: specificity versus generalization of perceptual training *Vis. Res.* **17** 30241–9

Auvray M, Hanneton S and O'Regan J 2007 Learning to perceive with a visuo-auditory substitution system: localisation and object recognition with 'The Voice' *Perception* **36** 416–30

Balota D A, Yap M J, Cortese M J, Hutchison K A, Kessler B, Loftis B and Treiman R 2007 The english lexicon project *Behav. Res. Methods* **39** 445–59

Bionic Vision Australia 2014 *Dianne Ashworth Bionic Eye Prototype Testing, 2014* (Bionic Vision Australia Channel, YouTube, retrieved: 2 February 2018) (https://www.youtube.com/watch?v=6EmleCs0KGY)

Bouma H 1971 Visual recognition of isolated lower-case letters *Vis. Res.* **11** 459–74

Capelle C, Trullemans C, Arno P and Veraart C 1998 A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution *IEEE Trans. Biomed. Eng.* **45** 1279–93

Caspi A 2018 Retinotopic to spatiotopic mapping in blind patients implanted with visual prosthesis *Afeka Conf. (Tel-Aviv)*

Caspi A and Zivotofsky A Z 2015 Assessing the utitlity of visual acuity measures in visual prostheses *Vis. Res.* **108** 77–84

Caspi A, Dorn J D, McClure K, Humayun M S, Greenberg R J and McMahon M J 2009 Feasibility study of a retinal prosthesis: spatial vision with a 16-electrode implant *Arch. Ophthalmol.* **127** 398–401

Chouvardas V G, Miliou A N and Hatalis M K 2008 Tactile displays : overview and recent advances *Displays* **29** 185–94

Cronly-Dillon J, Persaud K and Gregory R P F 1999 The perception of visual images encoded in musical form: a study in cross-modality information transfer *Proc. R. Soc.* B **266** 2427–33

da Cruz L *et al* 2013a The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss *Br. J. Ophthalmol.* **97** 632–6

da Cruz L *et al* 2013b *The Argus II Retinal Implant Allows Letter, Word Reading in Patients with Blindess* (SecondSightEurope Channel, YouTube, retrieved: 2 February 2018) (https://youtu.be/YU1F4TcGRlQ)

da Cruz L *et al* 2016 Five-year safety and performance results from the Argus II retinal prosthesis system clinical trial *Am. Acad. Ophthalmol.* **123** 2248–54

De Neve C 2011 *Neto and His BrainPort* Calgary Herald Channel. YouTube. (Retrieved: 2 February 2018) (https://www.youtube.com/watch?v=_nBs7PnKxzE)

DiCarlo J J, Zoccolan D and Rust N C 2012 How does the brain solve visual object recognition? *Neuron* **73** 415–34

Dorn J D, Ahuja A K, Caspi A, da Cruz L, Dagnelie G, Sahel A, Greenberg R J and McMahon M J 2013 The detection of motion by blind subjects with the epiretinal 60-electrode (Argus II) retinal prosthesis *JAMA Ophthalmol.* **131** 183–9

Edwards T L, Cottriall C L, Xue K, Simunovic M P, Ramsden J D, Zrenner E and Maclaren R E 2018 Assessment of the electronic retinal implant Alpha AMS in restoring vision to blind patients with end-stage retinitis pigmentosa *Am. Acad. Ophthalmol.* **125** 432–43

Fernandes R A B, Diniz B, Ribeiro R and Humayun M 2012 Artificial vision through neuronal stimulation *Neurosci. Lett.* **519** 122–8

Geruschat D, Bittner A and Dagnelie G 2012 Orientation and mobility assessment in retinal prosthetic clinical trials *Optom. Vis. Sci.* **89** 265–75

Getreuer P 2011 Image interpolation with contour stencils *Image Processing On Line* **1** 70–82

Gonzalez R C, Woods R E and Eddins S L 2009 *Digitial Image Processing Using MATLAB* 2nd edn (Knoxville, TN: Gatesmark Publishing)

Grant P *et al* 2016 The functional performance of the BrainPort V100 device in persons who are profoundly blind *J. Vis. Impair. Blind.* **110** 77–88

Hanneton S, Auvray M and Durette B 2010 The vibe: a versatile vision-to-audition sensory substitution device *Appl. Bionics Biomech.* **7** 269–76

Humayun M S *et al* 2012 Interim results from the international trial of second sight's visual prosthesis *Opthalmology* **119** 779–88

Jamieson D G and Morosan D E 1986 Training non-native speech contrasts in adults: acquisition of the english /ð/ and /θ/contrast by francophones *Percept. Psychophys.* **40** 205–15

Jung J H, Aloni D, Yitzhaky Y and Peli E 2015 Active confocal imaging for visual prostheses *Vis. Res.* **111** 182–96

Lozano C A, Kaczmarek K A and Santello M 2009 Electrotactile stimulation on the tongue: intensity perception, discrimination, and cross-modality estimation *Somatosens. Motor Res.* **26** 50–63

Lund K and Burgess C 1996 Producing high-dimensional semantic spaces from lexical co-occurrence *Behav. Res. Methods Instrum. Comput.* **28** 203–8

Margalit E *et al* 2002 Retinal prosthesis for the blind *Surv. Ophthalmol.* **47** 335–56

Meers S and Ward K 2005 A substitute vision system for providing 3D perception and GPS navigation via electro-tactile stimulation *Int. Conf. on Sensing Technology (November)* pp 551–6

Nau A C, Bach M and Fisher C 2013 Clinical tests of ultra-low vision used to evaluate rudimentary visual perceptions enabled by the BrainPort vision device *Transl. Vis. Sci. Technol.* **2** 1

Nau A C, Pintar C, Arnoldussen A and Fisher C 2014a Acquisition of visual perception in blind adults using the BrainPort artificial vision device *Am. J. Occup. Ther.* **69** 6901290010

Nau A C, Pintar C, Fisher C, Jeong J-H and Jeong K 2014b A standardized obstacle course for assessment of visual function in ultra low vision and artificial vision *J. Vis. Exp.* **84** e51205

Ortiz T *et al* 2011 Recruitment of occipital cortex during sensory substitution training linked to subjective experience of seeing in people with blindness *PLoS One* **6** e23264

Peli E 1991 Multi-resolution, error-convergence halftone algorithm *J. Opt. Soc. Am.* **8** 625–36

Rizzo J F and Ayton L N 2014 Psychophysical testing of visual prosthetic devices: a call to establish a multi-national joint task force *J. Neural Eng.* **11** 020301

Rizzo J, Wyatt J, Loewenstein J, Kelly S and Shire D 2003 Perceptual efficacy of electrical stimulation of human retina with a microelectrode array during short-term surgical trials *Investigative Ophthalmol. Vis. Sci.* **44** 5362–9

Shaw J 2016 Bionic vision *Eyenet Magazine* pp 55–60 (San Francisco, CA: American Academy of Ophthalmology) (Retrieved from: https://www.aao.org/eyenet/article/bionic-vision)

Sorin L, Lemarie J, Aussenac-Gilles N, Mojahid M and Oriola B 2014 Communicating text structure to Blind people with text-to-speech *Int. Conf. on Computers for Handicapped Persons* pp 61–68

Stiles N R B and Shimojo S 2015 Auditory sensory substitution is intuitive and automatic with texture stimuli *Sci. Rep.* **5** 15628

Stingl K *et al* 2015 Subretinal visual implant alpha IMS—clinical trial interim report *Vis. Res.* **111** 149–60

Stingl K *et al* 2017 Interim results of a multicenter trial with the new electronic subretinal implant alpha AMS in 15 patients blind from inherited retinal degenerations *Front. Neurosci.* **11** 1–11

Striem-Amit E, Cohen L, Dehaene S and Amedi A 2012b Reading with sounds: sensory substitution selectively activates the visual word form area in the blind *Neuron* **76** 640–52

Striem-Amit E, Guendelman M and Amedi A 2012a 'Visual' acuity of the congenitally blind using visual-to-auditory sensory substitution *PLoS One* **7** e33136

van Meeteren A 1995 Characterization of task performance with viewing instruments *Visual Models for Target Detection and Recogniton* ed E Peli (Singapore: World Scientific) pp 172–91

Wilson H R 1995 Quantitative models for pattern detection and discrimination *Vision Models for Target Detection and Recognition* ed E Peli (Singapore: World Scientific) pp 3–15

Zrenner E *et al* 2011 Subretinal electronic chips allow blind patients to read letters and combine them to words *Proc. R. Soc.* **278** 1489–97

Supplementary materials on next page

Supplementary

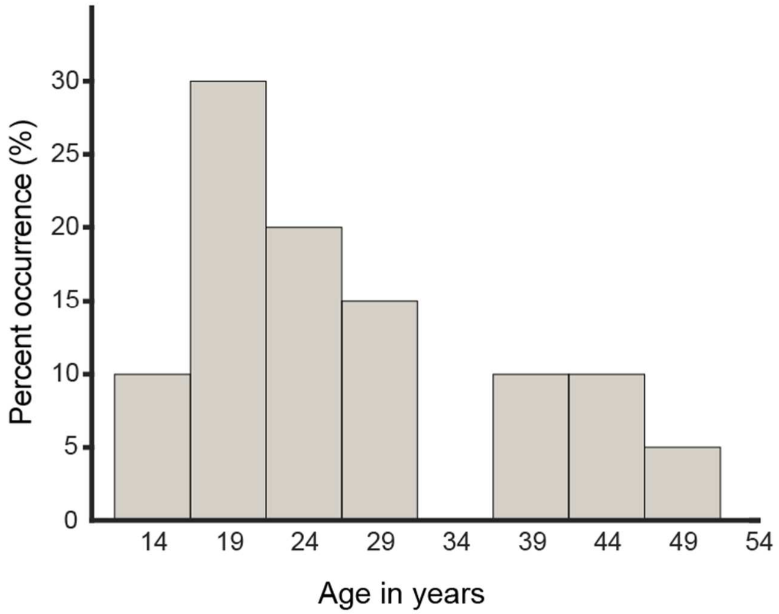


Figure S1. Histogram showing the distribution of age in years for subjects recruited in the study, plotted in terms of percentage occurrence. A bin size of 5 years of age used to compute the histogram. As depicted, 65% of the subjects were young adults, falling within the age range of 19-29 years old. The other 35% was made up of subjects in their early adolescence (10%) and adulthood (25%).

The subjects in our study had a wide age range. Since age has been shown to affect visual perceptual learning, such as picking up task-irrelevant cues (Chang, Shibata, Andersen, Sasaki & Watanabe, 2014), we performed a two-way repeated measures ANCOVA (word group x pre-/post- training, with age as a covariate) on the data. After controlling for subject age, our results revealed significant main effects of training and word group, and a significant interaction between word group and training. There were no significant relationships between subject age and their performance on word group and influence of training. Subject age was also not significantly related to the interaction between word group and training. The statistical results are described in Table S1 below.

Table S1. ANCOVA results examining effect of subject age on training and word group.

Factor	***F***	**(df_1, df_2)**	***p value***
Training	**20.4**	**(1,18)**	**<0.0001**
Word group	**17.2**	**(1,18)**	**< 0.01**
Training x Word group	**13.8**	**(1,18)**	**<0.01**
Training x Age	0.01	(1,18)	0.91
Word group x Age	0.03	(1,18)	0.87
Training x Word group x Age	0.15	(1,18)	0.70

Supplementary Reference

Chang, L.H., Shibata, K., Andersen, G.J., Sasaki, Y., & Watanabe, T. (2014). Age-related declines of stability in visual perceptual learning. *Current biology, 24*(24), 2926-2929. https://doi.org/10.1016/j.cub.2014.10.041.