Why verbalization of facial features increases false positive responses on visually-similar distractors: A computational exploration of verbal overshadowing

Aya Hatano (hatano.aya@c.mbox.nagoya-u.ac.jp), Taiji Ueno (taijiueno7@gmail.com),
Shinji Kitagami (kitagami@cc.nagoya-u.ac.jp), and Jun Kawaguchi (kawaguchijun@nagoya-u.jp)
Department of Psychology, Graduate School of Environmental Studies, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya City, Aichi 4648601, JAPAN

Abstract

Verbal overshadowing refers to a phenomenon whereby verbalization of a non-verbal stimulus (e.g., he had *slant eyes*) impairs subsequent non-verbal recognition accuracy. In order to understand the mechanism by which this phenomenon occurs, we constructed a computational model that was trained to generate an individual-face-specific representation upon input of a noise-filtered retinotopic face (i.e., face recognition). When the model verbalized the facial features before receiving the retinotopic input, the model incorrectly recognized a new face input as one of the different, yet visually-similar, trained items (that is, a false-alarm occurred). In contrast, this recognition error did not occur without prior verbalization. Close inspection of the model revealed that verbalization changed the internal representation such that it lacked the fine-grained information necessary to discriminate visually-similar faces. This supports the view that verbalization causes unavailability/degradation of fine-grained non-verbal representations, thus impairing recognition accuracy.

Keywords: verbal overshadowing; face recognition; computational modeling; verbalization

Introduction

Language is the principal medium for carrying out daily communications. This is still true when communicating our non-verbal experiences, such as recounting a crime scene we have witnessed, or describing the physical appearance of a criminal. Particularly, if we do not have a record of the event such as a picture or video, then conveying an eyewitness memory relies on language. A crucial question in cognitive science, therefore, is the influence of verbalization on non-verbal memory. Many studies have revealed that language has extra-communicative functions, in that it affects such cognitive functions as perception, learning, and memory. For example, in a seminal study by Schooler and Engstler-Schooler (1990), participants watched a video of a bank robbery for 30 seconds and following which half of the participants described the appearance of the bank robber. Subsequently, all of the participants were shown a line-up that consisted of the bank robber's photo and seven distractors. Results revealed that participants who had verbalized the bank robber's appearance were worse at recognizing the target individual than those who had not, a phenomenon known as *verbal overshadowing*. The procedure of these experiments can be experienced beyond an experimental setting. For example, during criminal investigations, an eyewitness may provide a statement describing the appearance of a criminal and subsequently identify them from a line-up. In such situations, it is crucial to prevent a false accusation and to examine the credibility of the eyewitness's testimony. Therefore, it is both theoretically and practically important to clarify the mechanism by which verbal overshadowing occurs. For this purpose, we constructed a parallel-distributed processing (PDP) model to simulate the effect of verbalization on subsequent visual recognition.

A closer review of the literature allows us to gain further insight into this phenomenon and therefore to establish a more specific aim for our model. First, although not all of the past studies have split the recognition scores into positive and negative trials, false alarm is sometimes more susceptible to verbalization before recognition than hit rates; that is, participants often inaccurately identify distractors as a target rather than miss a correct target (Meissner, Brigham, & Kelley, 2001). Furthermore, recognition accuracy in this study was positively correlated with accuracy of the verbal description prior to recognition. Based on these observations, Meissner et al. proposed a *recording interference account* that assumed verbalization rendered the representations less accurate (compared to visual representations), thus impairing subsequent visual recognition.

Second, Kitagami, Sato, & Yoshikawa (2002) revealed that verbal overshadowing is also sensitive to the degree of similarity between targets and distractors (manipulated with a morphing technique). Specifically, verbalization impaired subsequent visual recognition only when distractors were highly similar to the target (using a 9-alternative choice task with a "not present" response choice), but the impairment disappeared when similarity was low. It is also worth noting that this manipulation involved a change in the distractors, but not in the target picture itself. We revisited the original data and revealed that accuracy was impaired due to the more frequent choice of a distractor (a false alarm) rather than an incorrect choice of "not present" (a miss). Schooler (2002) explained this result with the *transfer inappropriate processing shift hypothesis*. This hypothesis assumes that verbalization induces a processing shift from visual to verbal, and that a shift to verbal processing makes fine-grained non-verbal information about faces unavailable. This non-verbal information is crucial for discriminating the target from others (see also, Maurer, LeGrand, & Mondroch, 2002), especially in a high-similarity condition (Kitagami et al., 2002). Although Schooler's hypothesis does not necessarily assume a correlation between recognition accuracy and verbal description accuracy (see also Kitagami

et al., 2002; Fallshore & Schooler, 1995), this hypothesis and the recording interference hypothesis by Meissner et al. (2001) share two ideas: First, both assume that fine-grained non-verbal information is necessary for face recognition. Second, both expect that a verbal representation which is generated during verbalization lacks such fine-grained information, thus impairing visual recognition.

More recently, Clare and Lewandowsky (2004) introduced an alternative hypothesis, arguing that verbalization shifts the criterion threshold such that participants say "The target is not present in the display" more frequently when in fact the target is present. Although this account can explain a range of existing data, two issues deserve consideration. First, even when a "not present" response was disallowed (that is, responses were forced choice), verbal overshadowing was observed in some studies, especially when elaborative verbalization was encouraged (Fallshore & Schooler, 1995). Second, the shifting criterion hypothesis cannot explain the fact that false alarm is more susceptible to verbalization than hit rate (Kitagami et al., 2002; Meissner et al., 2001). Thus, as Clare and Lewandowsky also speculated, there may be two mechanisms by which verbalization impairs subsequent visual recognition: One is shifting-criterion (Clare & Lewandowsky, 2004), and the other is degradation (Meissner et al., 2001) or unavailability (Schooler, 2002) of fine-grained non-verbal representations crucial for face recognition, especially when a distractor is visually confusing. This study focused on the latter possibility, and investigated how the nature of representations changes upon verbalization, and how this affects subsequent visual recognition. Computational modeling is an effective approach for this purpose. An explicitly implemented computational model allows a modeler to directly look at the nature of computations/representations that are underpinning a simulated behavior. The PDP model here was trained for three facial processing tasks: One was to represent the retinotopic input of a face in a non-verbal format (visual encoding/recognition); a second was to activate the correct units for verbal labels upon the same retinotopic input of a face (verbal encoding); the third was to represent a face in a non-verbal format upon verbal inputs (the mental imagery of a face upon verbal cues). After being trained for these tasks, the model was forced to activate some verbal units (i.e., verbalization), and we investigated how this forced activation changed the nature of the computed representation in the model, and how it affected subsequent visual encoding of a retinotopic input.

Method

Model Architecture

Figure 1 shows the architecture of the PDP model, built with LENS software (http://tedlab.mit.edu/~dr/Lens/). Three peripheral layers were connected bi-directionally with a single hidden layer. In order to reduce the computational
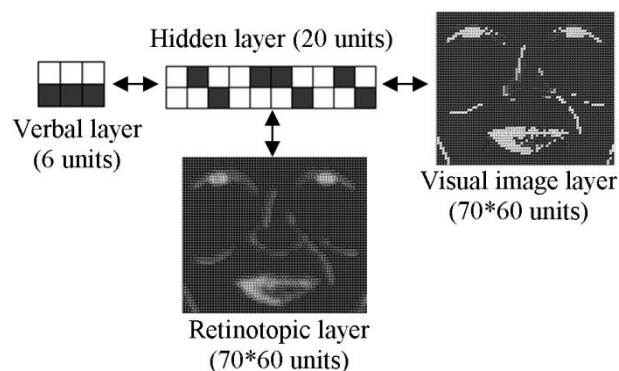

Figure 1: Architecture of the model (Hinton diagram).

demand in this large model, units between layers were connected sparsely, such that a unit was not connected with others if the external input/target value of that unit was always zero (e.g., a unit in the top-left corner). The bottom layer was named the *retinotopic layer*, and its activation patterns represented a filtered (Gaussian noise) face stimulus. The left layer was named the *verbal layer*, and each unit in this layer represented a verbal label for facial features in a localist manner. The right layer was named the *visual image layer*, and its activation pattern represented a non-filtered (without Gaussian noise) face stimulus. With this architecture and the representations in each layer, we trained the model for the three tasks described below.

Tasks

Visual encoding of a face from a retinotopic input (visual recognition). In this task, retinotopic face pattern was hard-clamped onto the retinotopic layer. Then, the network was trained to activate the non-filtered, *unique* visual face information of the same person in the visual image layer (*individuation or visual recognition* – see later).

Verbal encoding of facial features from a retinotopic input (verbalization). In this task, the input was the same as the previous visual recognition task, but the network was trained to activate the correct verbal units for each presented face. For example, if the face had *slanted eyes*, then the model had to turn on the unit for *"slant eyes"*, and had to turn off the unit for *"drooping eyes"*.

Mental imagery upon verbal cues. In this task, the verbal labels of facial features were presented onto the verbal layer, and the network was trained to activate the visual face information in the visual image layer. As we will explain later, the accuracy for this task never reached 100% because different faces sometimes shared the same verbal labels (i.e., different targets from the same input pattern).

Recognition. A standard experimental task on human recognition memory employs a N-alternative forced choice task to probe recognition process, particularly when examining verbal overshadowing. The model, however, was not trained for making an explicit N-alternative forced

"choice". Therefore, we should adopt a proper measure to probe the model's recognition. It is one of the most debatable issues in cognitive psychology regarding what process/mechanism is underpinning recognition. Following previous studies (e.g., Plaut & Behrmann, 2011), we examined whether the model could represent item-specific information (i.e., unique face) as an approximation of recognition process. If the model computes item-specific information of an "old" face in the visual layer from a "new" retinotopic (noisy) input, then it can be considered the model identifies this input as old face by mistake (especially after a verbal label for the old face was activated). In this way, we can at least measure false alarm safely, which is the target of the current study with this procedure.
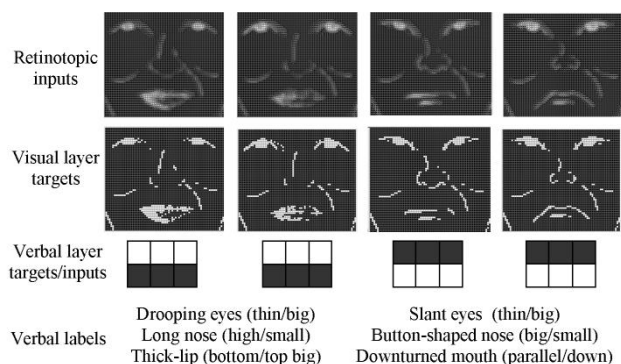

Figure 2: Four examples of the training patterns. Note that two examples within each half share the same verbal labels, and thus the same pattern activations in the verbal layer. (However, they are different faces with different specific features as shown in the parentheses).

Representations (face stimuli)

Figure 2 shows examples of the face pictures that we created using montage software (http://www1.mahoroba.ne.jp/~matumoto/nitaroS.html). Sixty-four face pictures were created by combining four types of eyes, four types of nose, and four types of mouth (see the bottom row of Fig.2 for the possible features) in the following steps. First, we selected two verbal labels for each part of the face – slant eyes, drooping eyes, long nose, button-shaped nose, downturned mouth, and thick lips. Next, we selected two specific types for each verbal label (e.g., slant -eyes [thin] and slant eyes [big] for the label slant eyes, as shown in the right two faces in Fig. 2). In this way, we created four types of eyes, nose, and mouth, resulting in 64 different faces by combining 4 by 4 by 4. In order to make the model trainable, we did not include other features such as hair. Finally, the size of each picture was 70*60 pixels, and the color information in each pixel was binarized (i.e., black pixel → 1; white pixel → 0). The resultant 4200-bit vector pattern was used as the target pattern of the visual image layer in the visual recognition tasks (see second row of Fig. 2).

The original bit patterns were transformed into the retinotopic input pattern by smoothing with Gaussian convolution (SD = 0.5) (Plaut & Behrmann, 2011). The original bit patterns were smoothed by Gaussian convolution (SD = 0.5). In summary, the model had to map a noise-filtered retinotopic input (top row of Fig. 2) into a clearer visual representation (second row of Fig. 2), which is necessary for visual recognition.

The pattern activations in the verbal layer represented the verbal labels in a localist manner (third row of Fig. 2). For example, when presented with a retinotopic pattern of the drooping-eyes, long-nose, and thick-lip face, then the model had to activate the first, second, and third units of the verbal layer (the left two cases of Fig. 2 show these examples). In the mental imagery upon verbal cues task, the same units in the verbal layer were turned on, and the network was trained to activate the visual images in the internal image layer. The accuracy in this task can never be 100% because sometimes a different target should be generated from the same input pattern (i.e., the same verbal labels). For example, slant eyes (thin) and slant eyes (big) shared the same verbal label, slant eyes. Therefore, the same unit (slant eyes) was turned on for these two cases, but different output patterns (thin or big eyes) should be generated in the visual image layer. This is true to humans: We can imagine various kinds of faces but cannot specify a unique face by simply hearing "slant eyes, long nose, and thick lips". A small amount of Gaussian noise (SD = 0.2) was added to the input for the hidden layer to encourage this layer to adopt more polarized outputs.

Training

Among the 64 face patterns, only 55 patterns were presented during training, and the remaining nine untrained patterns were used to evaluate the network's generalization performance. Furthermore, this allowed us to investigate how differently the model behaved with the trained faces ('old' items) and untrained faces ('new' items), as it was crucial for us to investigate the effect of verbalization before the recognition phase (described later in more detail).

In each trial, units in the corresponding layer (retinotopic or verbal layer) were hard-clamped to their input values, and the network was allowed to cycle 10 times. In each time step, the activation spread to the next layer, gradually being scaled by the values of the interconnecting weights, and then the network would settle into the steady state (an attractor). After 10 cycles of updates, the discrepancy between the output activation pattern generated by the network and the correct target pattern was calculated, and the connection strength was adjusted to reduce the discrepancy. The model was trained with a learning rate of 0.05, and with a decay parameter set to 0.0000001. When we evaluated the final performance, we used a strict criterion such that the output was scored correct if the discrepancy was within 0.5 in every unit of the target layer (i.e., the activation was less/more than 0.5 if the target was zero/one for each unit respectively).

Given that young infants recognize their parents easily, it would be natural to assume that visual recognition skills are acquired earlier than an ability to verbalize facial features, or to imagine a face upon verbal cues. Thus, all 55 of the face stimuli were first trained for the visual recognition task. After learning to generate a steady state for more than 50% of the training items in this task, the other two language-related tasks were included in the training schedule.

Results

Training tasks

Five independent simulations were run with different random seeds, and we confirmed consistent results across five cases. The training finished after 2837 epochs of training (in each epoch an item appeared once for each task in a random order), at which point the network's performance reached 100% in both the visual recognition and the verbalization tasks from a retinotopic input for both trained and untrained items (i.e., generalization). The accuracy in the visualization task from verbal labels was 0% (see above for the reason).

Visual recognition with/without verbalization

In order to investigate the visual recognition process of a retinotopic input after/without verbalization, we recorded the activation patterns in the visual image layer (right column of Fig. 3) when the network settled on 10 cycles after the retinotopic input presentation (left column of Fig. 3). The upper two rows of Fig. 3 show the pattern activations for a trained ('old') face (*drooping eyes [thin], long nose [high], and thick lip [bottom big]*) and for a visually-similar, yet untrained ('new') face (*drooping eyes [big], long nose [high], and thick lip [bottom big]*), respectively. Both retinotopic inputs were correctly mapped onto every unit of the visual image layer. This means that two visually-similar faces were successfully discriminated (see the bigger eyes represented in the second row), unless they were preceded by the verbalization process (100% accuracy in computation of the individual-specific face information for all the nine untrained items). Next, the middle two rows of Fig. 3 show the activations for the same two items as the upper rows but after verbalization. Specifically, we simulated the following situation: Imagine that the network had encountered the 'old' face shown in top row of Fig. 3 (*drooping eyes [thin], long nose [high], and thick lip [bottom big]*), and the network had verbalized the correct labels (*drooping eyes, long nose, and thick lip*). To simulate this situation, the three verbal units for these labels were manually turned 'on' (generating the outputs of 1.0) and the network was allowed to cycle 10 times, during which the activations spread into the other layers (it updated its internal status 10 times). After 10 cycles, a retinotopic input for the trained face ('target') and that for the visually-similar, yet untrained face ('new') were presented respectively, and the network was allowed to update its status 10 times until each input pattern was mapped onto a
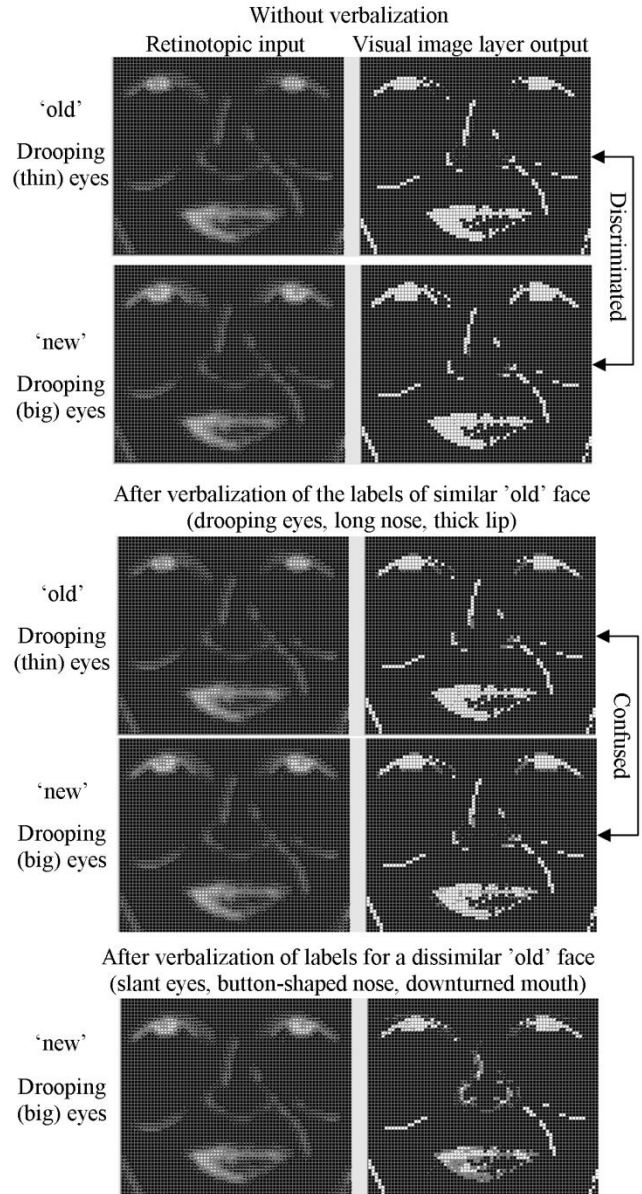

Figure 3: Activation patterns in the visual image layer (right) upon retinotopic inputs (left) for trained 'old' face and for untrained 'new' face in the visual recognition task. Upper two rows: without verbalization. Middle two rows: after verbalization of a similar 'old' face. Bottom row: after verbalization of a dissimilar 'old' face.

steady pattern in the visual image layer (right column). A visual inspection reveals the retinotopic input for the visually-similar 'new' face (*drooping eyes [big]*) was mapped onto the pattern for the 'old' face (*slant eyes [thin]*) in the visual image layer (false alarm). Euclid distance of the output pattern from the target "new" face was larger than that from the lure "old" face (i.e., similar to the "old" face pattern). This means that the model actually computed the item-specific information of the "old" face (false alarm). The same analysis was conducted for all the nine "new" items (against its visually-similar "old" face, respectively),
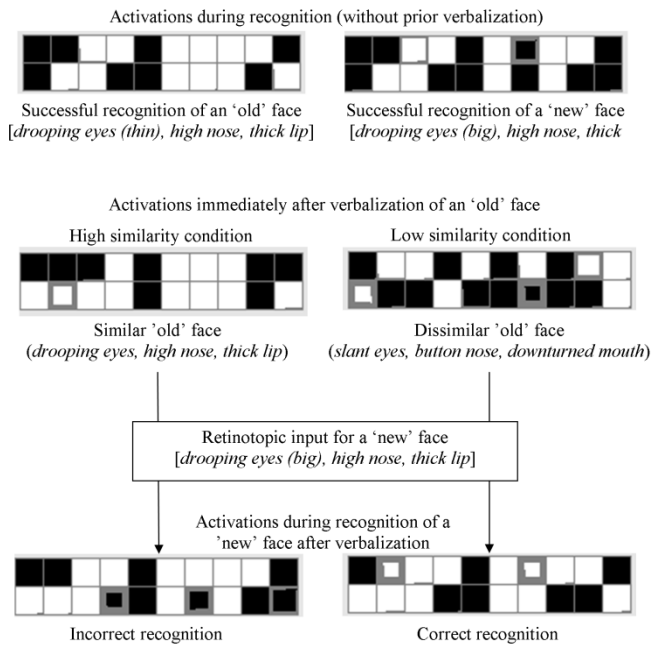
Figure 4: Internal activation patterns in the hidden layer (Hinton diagram: more white units denote higher activation values than black units) at the various kinds of time point (see main text).

and averaged across the five individual simulations. The resultant recognition accuracy was 60% (40% false alarm), $SD = 14.9\%$, which was significantly lower than 100% (t (4) = 5.99, p = .003). This confirms that the example result in Figure 3 is generalized across other patterns. In contrast, the retinotopic input for the 'old' item was mapped onto the correct pattern (*drooping eyes [thin]*) in the visual image layer, though less weakly than when presented without verbalization (top row). Taken together, these results confirm that false alarm was more susceptible to verbalization than hit rate. Finally, the bottom row of Fig. 3 shows the simulated result in the condition where the distractor was dissimilar to the target. Specifically, the activation patterns were taken from the same untrained 'new' face (*drooping eyes [big], long nose [high], and thick lip [bottom big]*), but after activations of the irrelevant verbal units (*slant eyes, button-shaped nose, and downturned mouth*). Thus, the model had encountered a dissimilar person, and verbalized the dissimilar labels before visual recognition of a 'new' item. As a result, the network did not settle into the pattern of the dissimilar target face (*slant eyes and downturned mouth*), but the represented pattern was more similar to the correct pattern (*drooping eyes and thick lip*). In other words, the model did not confuse the presented retinotopic input (the 'new' face) with the previously encountered 'old', yet dissimilar face, thus avoiding false alarm in this low-similarity condition (Kitagami et al., 2002).

Finally, in order to understand the mechanism of verbal overshadowing, the pattern activation in the hidden layer was measured at the various kinds of time point (Figure 4). First, the Hinton diagrams at the top row of Figure 4 show

the internal activation patterns when the network successfully discriminated two visually similar 'old' and 'new' faces, respectively (shown in the upper panels of Fig. 3). A visual inspection reveals these two representations are very similar. This concurs with the idea that fine-grained representations are crucial in face recognition (Maurer et al., 2002), without which one would be easily mapped to the other, incorrect, face pattern in the visual image layer (i.e., incorrect recognition).

Next, the left diagram of the middle row of Figure 4 shows the activation pattern immediately after verbalization of 'drooping eyes, high nose, and thick lips'. As a result, this internal representation immediately after verbalization was neither identical with that for visual recognition of the 'old' face (top left) nor that for visual recognition of the 'new' face (top right), concurring with the idea that verbalization generates the representation that lacks fine-grained information crucial for face recognition (Maurer et al., 2002). Though lacking such detailed information, it was nonetheless closer to the representation for the 'old' face (top left) than that for the 'new' face (top right). In other words, the model's internal status had already moved towards the pattern for the 'old' face. We will explain later why this representation increased the false alarm of the model when the distractor was similar to the target.

Discussion

The present computer simulation examined how internal representations changed upon verbalization and how this affected subsequent visual recognition. Without verbalization, the model represented the correct and unique pattern activation for each old face and for a visually-similar new face, respectively, in the visual image layer. This confirms that the model did not confuse two visually-similar retinotopic inputs. On the other hand, the model failed to represent the correct pattern for a new face following the forced activation of verbal units for an 'old', visually-similar face (i.e., verbalization). Instead, the represented pattern in the visual layer assimilated to that for the visually-similar 'old' face, suggesting that the model could not differentially recognize the 'new' face from the 'old' face (a false alarm). Importantly, this assimilation was weakened when the preceding verbalization included the features of an 'old', yet dissimilar face. Therefore, these results mirrored Kitagami et al. (2002), who found that participants' false alarm increased upon verbalization when the distractors were similar to the target.

Explicit implementation of a computer model allowed us to directly look at the internal representations to understand why the model behaved in this way. In a normal situation (without verbalization), the model computed very similar, yet unique, internal representations for retinotopic inputs of visually-similar faces. This fact is consistent with the idea that fine-grained representation is necessary for visual recognition of faces (Maurer et al., 2002), especially when discriminating a target from similar distractors. When the model verbalized the facial features, this internal

representation changed such that it was neither identical to that of an 'old' face, nor that of a ''new' face, supporting the argument that verbalization either degrades the fine-grained representation (Meissner et al., 2001), or renders it unavailable (Schooler, 2002). Nonetheless, it was closer to the representation for the verbalized 'old' face than to that of a 'new' face. In order to understand why this representation induced a false alarm for a visually-similar face, it is useful to describe the general activity of PDP models here. During training, a PDP network finds a unique *attractor state* (a unique abstract pattern in the hidden layer) associated with each input pattern. Therefore, generating a correct output is sometimes described as if the internal activity of the hidden layer *falls into its unique attractor basin*. Though they are unique, similar inputs are associated with similar/neighboring attractor basins (as shown in the top two panels of Fig. 4). Consequently, if the internal representation of the model is degraded for some reason, a similar input can incorrectly drift and fall into the wrong attractor basin, generating an incorrect output. In the current model, verbalization generated the internal representation that lacked fine-grained information crucial for visual recognition (middle-left row of Fig. 4). Though it lacked such information, it was nonetheless closer to the representation of the 'old' face (the top and middle left rows are similar in Fig. 4). In other words, the model's internal status had moved towards the attractor basin for the 'old' face by verbalization. In such a situation, a similar retinotopic input, which would have settled into a unique, yet similar/neighboring attractor basin without verbalization (top-right of Fig. 4), was easily captured by the attractor for the 'old' face. The resultant (captured) internal activation pattern is shown in the left bottom row of Fig. 4, which was more similar to the top-left pattern than the top-right pattern (i.e., incorrect recognition). This is the mechanism by which verbalization impairs subsequent recognition, especially when the distractor is similar to the 'old' face. In contrast, when the dissimilar (inconsistent) labels were verbalized before visual recognition, the hidden layer activation pattern was very different to that for a subsequent 'new' face (i.e., the top and middle right rows are not similar). In this case, the network is not captured by this dissimilar attractor basin, as is shown in the bottom right diagram of Fig. 4.

In summary, as the present study has demonstrated, a computer simulation is a useful tool for investigation of verbal overshadowing. It is difficult to examine verbal overshadowing empirically, given that the standard paradigm involves a single-trial measurement. Therefore, many participants are necessary for detecting a reliable effect, and it is difficult to systematically manipulate a variable as a within-subject factor. In such a situation, it is worthy to implement a computational model in order to understand the mechanism and to provide a theory-driven question that can be empirically testable in human experiments. Of course, any computational modeling should be concerned whether the model's representation/process is the same as the human's, but previous studies have demonstrated that investigating the internal representation of a model is a useful approach to advance the cognitive theory (e.g., Plaut & Behrmann 2011). Furthermore, the current model can be extended to other types of perceptual stimuli (not just face). Thus, we expect that the present study would be an important step to clarify the relationship between language and perception in general. Finally, one issue deserves consideration: The current model simulated the increase in false alarm upon verbalization (Kitagami et al., 2002; Meissner et al., 2001), rather than a missed response. Although some studies failed to detect a significant difference between these two measures (Schooler & Engstler-Schooler, 1990, a recent experimental and computational study (Clare & Lewandowsky, 2004) suggested that it was actually the increase in "not present" responses, a response type that was not implemented in the current model. A future modeling target would be to understand the mechanism by which verbalization increases both "not present" responses and false alarms, of which the latter particularly occurs when the distractors and targets are similar.

References

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology-Learning Memory and Cognition, 30*(4), 739-755

Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 1608-1623.

Kitagami, S., Sato, W., & Yoshikawa, S. (2002). The influence of test-set similarity in verbal overshadowing. *Applied Cognitive Psychology*, 16, 963-972.

Maurer, D., LeGrand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Science*. 6, 255-260

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition*, 29, 176-186.

Plaut, D. C. & Behrmann, M. (2011). Complementary neural representations for faces and words: A computational exploration. *Cognitive Neuropsychology*, 28, 251-275.

Schooler, J.W. (2002), Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, 22, 989-997.

Schooler, J. W., & Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, 22, 36-71.