LETTER **Communicated by Oded Ghitza**

A Self-Organized Artificial Neural Network Architecture for Sensory Integration with Applications to Letter-Phoneme Integration

Tamas Jantvik
Tamas.Jantvik@ltu.se
Lennart Gustafsson
Lennart.Gustafsson@ltu.se
Computer Science and Electrical Engineering, Luleå University of Technology, S-971 87 Luleå, Sweden

Andrew P. Papliński
app@csse.monash.edu.au
Clayton School of Information Technology, Monash University, Melbourne, Victoria 3800, Australia

The multimodal self-organizing network (MMSON), an artificial neural network architecture carrying out sensory integration, is presented here. The architecture is designed using neurophysiological findings and imaging studies that pertain to sensory integration and consists of interconnected lattices of artificial neurons. In this artificial neural architecture, the degree of recognition of stimuli, that is, the perceived reliability of stimuli in the various subnetworks, is included in the computation. The MMSON's behavior is compared to aspects of brain function that deal with sensory integration. According to human behavioral studies, integration of signals from sensory receptors of different modalities enhances perception of objects and events and also reduces time to detection. In neocortex, integration takes place in bimodal and multimodal association areas and result, not only in feedback-mediated enhanced unimodal perception and shortened reaction time, but also in robust bimodal or multimodal percepts. Simulation data from the presented artificial neural network architecture show that it replicates these important psychological and neuroscientific characteristics of sensory integration.

1 Introduction

Many phenomena manifest themselves in more than one sensory modality. In such cases, the integration of several complementary unimodal percepts

Color versions of figures in this letter are presented in the online supplement available at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00149.

into single bi- or multimodal percepts offers several advantages. Time to detection of an event is, for instance, reduced if the event presents itself in more than one modality (Hershenson, 1962). It is also well established that the detection and identification of such an event are more robust against disturbances in one or more modalities (Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001).

Sensory integration has been studied extensively. Calvert, Spence, and Stein (2004) offer a comprehensive review, as do Driver and Noesselt (2008). Functional magnetic resonance imaging (fMRI) has become increasingly important as a tool for noninvasive studies of sensory integration in human subjects (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005). The fMRI instrument has been used to produce massive amounts of data, data that need both modeling and interpretation to aid understanding. A major part of our motivation constitutes the development of functionally mimicking computational architectures that have their foundations in this extensive body of information and literature, thereby exploring selected aspects of brain function.

Speech is an important case of bimodal stimuli; it is primarily perceived by hearing but is also seen in lip, mouth, and other facial movements. It has long been known that visual speech significantly increases auditory speech comprehension under noisy auditory conditions (Sumby & Pollack, 1954), and one study has shown that audiovisual speech yields comprehension improvement even when auditory conditions are good (Remez, 2005).

Another long-known fact is that bimodal sensory integration takes place in association areas, such as the superior temporal polysensory area (STP) (Schroeder & Foxe, 2002; Schroeder et al., 2003). Multimodal convergence also occurs earlier in cortical sensory processing (Ghazanfar & Shroeder, 2006; Kayser & Logothetis, 2007). There is, however, a distinction between the classic multimodal areas, where many neurons have afferents from more than one modality, and the areas previously seen as purely unimodal, where cross-modal influence is modulatory (Kayser & Logothetis, 2007).

Several mechanisms mediate the integration of signals conveying auditory and visual information onto a neural structure. Both feedforward (bottom-up) connections from lower levels to higher levels in the neural hierarchy and feedback (top-down) connections going in the opposite direction, as well as lateral connections, integrate information from different sensory modalities (Calvert & Thesen, 2004; Schroeder & Foxe, 2002; Foxe & Schroeder, 2005).

Feedforward and feedback have both been extensively studied in neural processing. Recent work suggests that a presented stimulus will cause a rapid feedforward sweep of activity with a short delay at each hierarchical level (Lamme & Roelfsema, 2000). This activity is then modulated by feedback.

Feedback plays an important role in the processing of audiovisual speech. According to the review by Price (2000), speech is processed in a network of cortical regions, with early processing taking place in sensory specific cortices (Binder et al., 2000; Calvert et al., 1997; Calvert & Campbell, 2003). Processing for phoneme perception takes place in the left posterior superior temporal sulcus (STSp; Dehaene-Lambertz et al., 2005; Möttönen et al., 2005). Integration of the two modalities of audiovisual speech takes place in the multimodal association area in the superior temporal sulcus (STS) and the superior temporal gyrus (STG; Calvert & Campbell, 2003), located between the sensory-specific auditory and visual areas.

Audiovisual speech exists in two forms: lip reading while hearing, and reading letters while hearing. In both forms, the auditory perception is enhanced compared to purely auditory speech (Frost, Repp, & Katz, 1988; Dijkstra, Frauenfelder, & Schreuder, 2004). The activity in unisensory auditory cortex is increased due to feedback from the bimodal area in the STS to auditory cortex (Calvert, Campbell, & Brammer, 2000; van Atteveldt, Formisano, Goebel, & Blomert, 2004). In this letter, we are interested in the latter variant of audiovisual integration. This integration process has been studied extensively enough to enable the proposition of an artificial neural network architecture whose workings are inspired by these studies.

Letters are processed in unisensory visual cortex in or close to the left fusiform gyrus (Gauthier et al., 2000; Polk & Farah, 1998; Polk et al., 2002). Bimodal integration of phonemes and letters also takes place in the STS, through feedforward processing (Raij, Uutela, & Hari, 2000; van Atteveldt, Formisano, Goebel, & Blomert, 2004). An important series of reports on different aspects of letter-phoneme integration has been presented by the Department of Cognitive Neuroscience at Maastricht University in the Netherlands (van Atteveldt et al., 2004; van Atteveldt, Formisano, Blomert, & Goebel, 2007; Blau, van Atteveldt, Formisano, Goebel, & Blomert, 2008; Froyen, van Atteveldt, Bonte, & Blomert, 2008). An additional finding on the cortical network architecture of letter-phoneme integration indicates effects of congruency of letter and phoneme also in extrastriate visual cortex (Blau et al., 2008).

Earlier work using similar ideas to those in this letter has consisted of modeling the processing of phonemes and letters in sensory-specific and a bimodal association area (Papliński & Gustafsson, 2005, 2006; Gustafsson & Papliński, 2006, Chou, Papliński, & Gustafsson, 2007). The modeling was carried out using an artificial neural network architecture consisting of separate but interconnected self-organizing modules with phonetic and graphic inputs, respectively, and an integrating bimodal module, corresponding to the cortical architecture laid out above. Feedback from the bimodal association area to the auditory cortex was also modeled in the auditory module. The artificial neural

network architecture was called a multimodal self-organizing network (MMSONv1).

Simulations of the MMSONv1 demonstrated that bimodal percepts have an increased robustness against additive noise in phonemes and that this increase of robustness is transferred down the auditory processing stream by feedback. Results from simulations of this artificial neural network architecture thus also demonstrated the increased classification certainty of a noisy phoneme when the corresponding uncorrupted letter is also available.

It has been shown that there is activation in auditory cortex during lip reading (Calvert et al., 1997), even though the sound has been eliminated. The feedback from the bimodal area to the auditory area should thus cause activity there even in the absence of auditory stimuli, provided a visual stimulus is present. In later modeling work (Gustafsson, Jantvik, & Papliński, 2007), an extended artificial neural network architecture with more versatile modules than those used previously (Papliński & Gustafsson, 2005, 2006; Gustafsson & Papliński, 2006; Chou et al., 2007) was shown to replicate this property also, while still replicating the properties of earlier versions of the architecture.

The main contribution of this letter is a new self-supervised and fast algorithm for combining data from self-organized modules that have been trained on different but congruent stimuli and this has yielded an extension of the artificial neural network architecture's repertoire of parallels to cortical function. Using this new architecture, we show how a letter enhances the congruent phoneme when an incongruent phoneme is superposed on the congruent phoneme. We also show that a letter, even in corrupted form, enhances the response to the noisy congruent phoneme. Another new parallel is found in a study of the development over time in this artificial neuron network. With feedback, there is a recurrent process that, after a few loops, converges to a final state. We show that integration of bimodal stimuli yields an initially higher activity that, through feedback, then rises more quickly than that caused by unimodal stimulation. In addition, this letter contains a detailed description of the artificial neural network architecture that also aims to make understanding and potential experimentation by interested third parties easier.

2 Methods

2.1 The Multimodal Self-Organized Network. The focus of this letter is a biologically inspired artificial neural network architecture, referred to as the multimodal self-organized network (MMSON). The architecture comprises interconnected artificial neural network modules, each of which performs a topographic mapping as in Kohonen feature maps (Kohonen, 2001—self-organizing maps), and in addition, for each stimulus, they

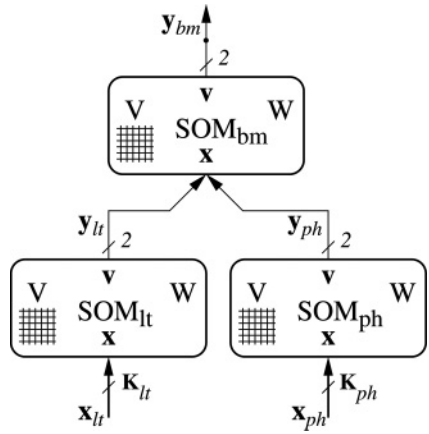


Figure 1: A two-level feedforward-only multimodal self-organized network (MMSON) processing auditory and visual stimuli consisting of self-organizing modules (SOMs). The auditory stimuli are processed in SOM_{ph} and the visual stimuli in SOM_{lt}. Bimodal integration then takes place in SOM_{bm}.

generate an output vector representing the positional coordinates of the artificial neuron with maximum postsynaptic activity. It is important to note that these output vectors are meant to represent semantic labels of the related stimuli. Two types of modules are detailed in sections 2.2 and 2.3, each employing its own learning law.

We start by briefly reviewing the artificial neural network architecture, which we aim to extend: the feedforward-only multimodal self-organizing network (previously abbreviated MuSON; Gustafsson & Papliński, 2006). It is depicted in Figure 1 (see Papliński & Gustafsson, 2005, 2006; Gustafsson & Papliński, 2006; Chou et al., 2007, for details). As can be seen in the figure, it is composed of three modules: SOM_{lt}, SOM_{ph}, and SOM_{bm}. Each module thus generates an output vector, which formally can be described using the general expression

$$\mathbf{y} = F(\mathbf{x}; W, V), \tag{2.1}$$

where $\mathbf{x}$ represents the stimulus for a given module, W is the module's weight matrix, and V describes the structure of the neural grid within the module. The 2D output signal $\mathbf{y}$ gives the 2D position of the winning artificial neuron—the unit with peak postsynaptic output. For simplicity, the 2D positions are encoded as a pair of numbers belonging to the unit disc. With neural grids of rectangular layout with $m \times n$

neurons, such as those used here, these numbers can be trivially calculated as

$$\mathbf{y} = \begin{pmatrix} y_x \\ y_y \end{pmatrix} = \frac{1}{M_R} \begin{pmatrix} c - \dfrac{m-1}{2} \\ r - \dfrac{n-1}{2} \end{pmatrix}, \tag{2.2}$$

$$\text{where } M_R = \sqrt{\left(m - \tfrac{m-1}{2}\right)^2 + \left(n - \tfrac{n-1}{2}\right)^2},$$

and c and r are a neuron's coordinates (column and row) numbered with natural numbers in the straightforward way.

Preprocessed sensory stimuli, $\mathbf{x}_{lt}$ and $\mathbf{x}_{ph}$, form the inputs to their respective unisensory modules, SOM_{lt} and SOM_{ph}. The two-dimensional outputs from these modules, $\mathbf{y}_{lt}$ and $\mathbf{y}_{ph}$, are combined through concatenation, and the results are then formed into inputs to the higher-level bimodal module, SOM_{bm}.

The training of the artificial neural network architecture can be carried out in two ways: in one step by repeatedly presenting the inputs to the sensory modules and the concatenation of the two respective winning coordinates to the bimodal module and then applying the learning rule in all three modules, or in sequential order, with the sensory modules being trained first, whereupon the bimodal module is trained with concatenations of neuron coordinates belonging to congruent classification areas in the two former modules. The semantic of classification areas is described in section 3.1, but simply put, a classification area of a stimulus is the area spanned of the group of artificial neurons that respond most strongly to that stimulus. When the well-known Kohonen learning law is used, as has been done in this and previous work (Papliński & Gustafsson, 2005; Gustafsson & Papliński, 2006), both training techniques yield similar results.

Although the feedforward architecture depicted in Figure 1 will integrate unimodal percepts into bimodal percepts, it is not sufficient for more advanced application purposes due to two reasons: first, feedback connections play an important role in sensory integration, and these are missing from the architecture. Second, there is no quality estimate of the stimuli resulting in the unimodal percepts, and such estimates seem to play a key role when information from different modalities is combined. It has been hypothesized that in the nervous system, different cues are combined in such a way that more reliable cues are given greater weight in the integrated percept; this hypothesis is not new (Taylor, 1962). The hypothesis has of late been experimentally verified in a number of cases (for a recent review, see, for instance, the work of Ernst & Bülthoff, 2004), including the case where the cues are unimodal, such as visual cues for depth estimation (Landy, Maloney, Johnston, & Young, 1995) and cases where the cues

are bimodal, such as audiovisual in the ventriloquist effect (Alais & Burr, 2004) and visual-haptic in estimating the height of an object (Ernst & Banks, 2002).

Consequently, we now extend this architecture according to our source of biological inspiration; the intention is to incorporate feedback, from the bimodal to the auditory processing unit, and to enhance the architecture's ability to exchange information between different processing modules. This latter extension makes use of the activity levels generated by the processing modules. Since the activity level of an artificial neuron of the kind that is used here reflects how well its input agrees with previous training data, this aspect of the response correlates with how well the submitting module recognizes its input. Thus, as a consequence of this enhancement, the modules can communicate graded responses, and these gradings can then be used to perform weighted response fusion. To carry out this fusion while also taking the degree of congruency between the responses into account, formal relations between the signals pending fusion have to exist, and these relations must be exploited using a technique that yields a fused signal that is meaningful.

A frequently occurring method of information exchange between self-organized feature maps entails the use of Hebbian links (Miikkulainen, 1997; Li, Farkas, & MacWhinney, 2004; Li, Zhao, & MacWhinney, 2007; Casey & Pavlou, 2008; Mayor & Plunkett, 2008, 2010). The method used here does not use Hebbian links for this information exhange (others have used different methods as well; Gliozzi, Mayor, Hu, & Plunkett, 2008; Martin, Meredith, & Khurshid, 2009), but we employ our own alternative and novel approach conveying only the bare minimum information: which percept is identified in a map and what confidence level this percept identification has. During fusion of these kinds of condensed labels, they are algorithmically transformed using information present in the self-organized modules and knowledge of stimuli congruencies. This is a fundamental property of our artificial neural network architecture and an important contrast that sets our architecture apart from other artificial neural network architectures that deal with infomation transfer among several self-organized feature maps. Using the algorithmic approach described in the sections that follow for linking self-organized maps yields a reduced computational load in comparison to heuristics such as Hebbian linkage.

2.2 The Extended MMSON with SumSOMs. In the MMSON architecture, initialization is carried out using uncorrupted (i.e., ideal) stimuli only. The presentations of corrupted stimuli, which are formed by adding noise to the uncorrupted stimuli used during initialization, are then unlikely to exactly match any of the neurons' weight vectors, and the resulting output activity hence becomes less than the maximum possible. In this way, the peak magnitude of the output activity can be taken as a

measure of the reliability of a stimulus. We therefore also need to introduce processing of this magnitude into the complete artificial neural network architecture.

To this end, we extend the artificial neural network architecture to use modules whose output signals do not merely consist of the position $\mathbf{v}$ of a winner neuron, resolved via the structural description V as before, but also the activity level a of this neuron. To fuse two outputs of this kind in the desired way, we employ coordinate transformation and postsynaptic activity field combination instead of mere concatenation of coordinates. More specifically, we introduce a neural network configuration SumSOM (summing self-organized module), which combines the outputs coming from a pair of modules (SOMs or SumSOMs, or both) and classifies this combined signal. The coordinate transformation is applied to at least one of the SumSOMs' incoming signals, prior to combination, to permit the combination step to be straight-forward.

In this letter we use the phrase *postsynaptic activity field,* or simply *field,* to denote the postsynaptic activities laid out in a matrix of the same two-dimensional configuration as the neural lattice yielding the activities. This field is formed using a function

$$\Phi = \Theta(\boldsymbol{d}; V) \tag{2.3}$$

that, given a vector $\boldsymbol{d}$ and structural description V, transforms the postsynaptic activities from vector format to the desired matrix format specified by V. An alternative way of describing this is to say that we have two complementary ways of indexing. One is to index according to the structural description V, which in this letter amounts to two-dimensional indexing (i, j), say. Given these indices and the maximal number $j_{\max}$ that, for example, j can assume, the other way can be described as sequential indexing, with index k, say, so that

$$k = j_{\max} \times (i - 1) + j.$$

The SumSOM produces an output of the same kind that the simpler self-organized module does: a classification coded as a 2D position of maximum activity and the activity level at that location. Therefore, the output from a SumSOM can be interpreted in the same way as the output of the standard SOMs used in previous work. Our intention is for these SumSOMs to be seen as extensions of the SOMs, allowing the processing of more than one input signal while replicating the SOMs' behavior during the application phase.

Two configurations of the SumSOM are described here; one in which the output of one module is transformed in order to enable modulation of the postsynaptic activity field of another module and one in which the

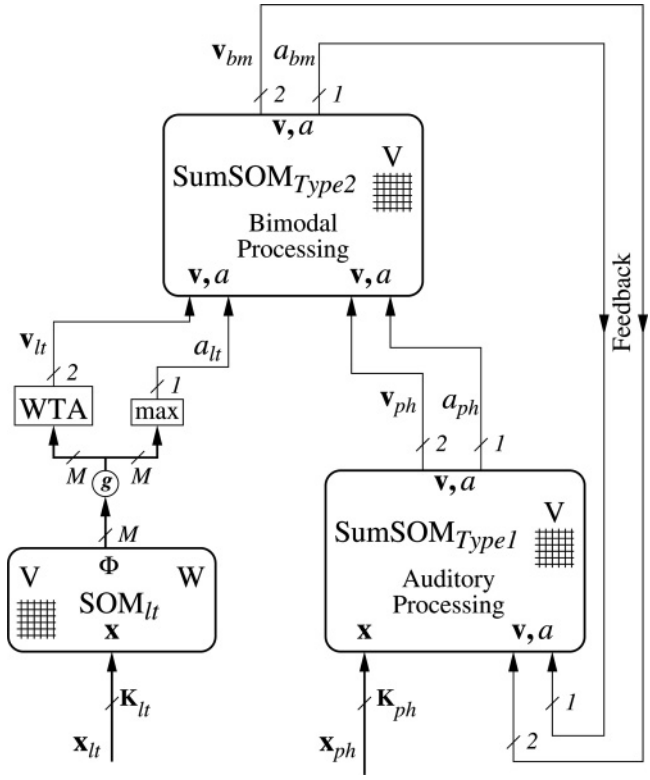


Figure 2: A two-level MMSON with feedback processing auditory and visual stimuli. The MMSON consists of self-organized modules, SumSOMs and additional circuitry. SumSOMs combine signals that may come from SOMs or SumSOMs, or both. The number of neurons in the letter module is denoted by *M*.

outputs of two modules are both transformed so that the fused response lies in a space that differs from both of the modules' output spaces. We call the former configuration a SumSOM of type 1 and the latter a SumSOM of type 2.

An outline of the extended artificial neural network architecture, in which both SOMs and SumSOMs are used as building blocks, is depicted in Figure 2. Inspiration for the architecture's coarse design, with its blocks and their connections, has largely stemmed from the cortical network proposed by van Atteveldt et al. (2004) and also from the findings of Schroeder and Foxe (2002): with feedforward convergence to multisensory cortices and auditory feedforward and visual feedback convergence to auditory association cortex.

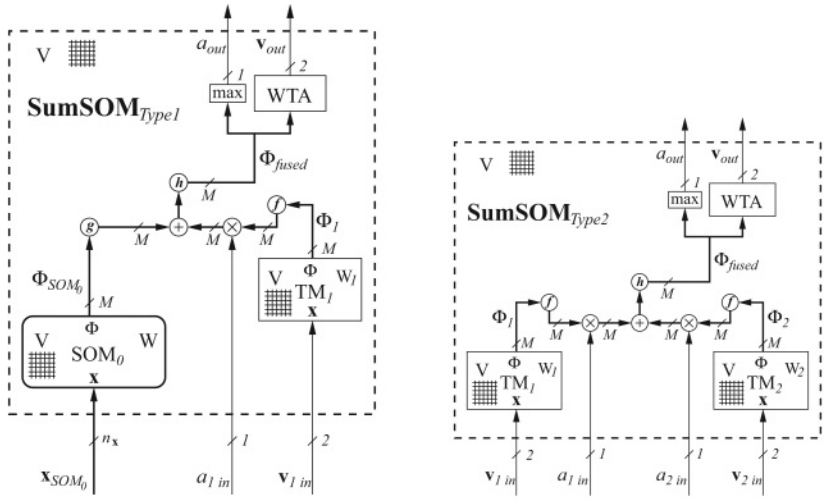


Figure 3: (a) $SumSOM_{Type1}$ enables modulation of SOM_0's response field using the input, $(\mathbf{v}_{1\,in}, a_{1\,in})$, coming from another module. M is the number of neurons in SOM_0. (b) $SumSOM_{Type2}$ fuses together outputs, $(\mathbf{v}_{1\,in}, a_{1\,in})$ and $(\mathbf{v}_{2\,in}, a_{2\,in})$, of two modules (SOMs or SumSOMs). Here, M is the number of neurons in the neural lattice template. See the text for further details.

The SumSOM described in this letter is similar to the module with the same name previously presented by Gustafsson et al. (2007). It is based on the same ideas and also serves the same purpose of enabling the MMSON to fuse signals from different modules while taking their adherent activity levels (i.e., intensities) into account. The difference in the description of it here, and of that in previous work, lies in the way the module is initialized.

2.3 Circuitry of the SumSOM Modules. This section describes what the SumSOM modules contain and ideas for their use. The specific details of how we initialize the different parts of these modules are described in section 2.5.

Figure 3a shows a schematic of a type 1 SumSOM, which enables the output of another module (i.e., SOM or SumSOM), coming in on $\mathbf{v}_{1\,in}$ and $a_{1\,in}$ (as a 2D coordinate of a neuron and its activity, respectively), to modulate the postsynaptic activity field of another SOM, here labeled SOM_0.

During initialization of this module, it is assumed that SOM_0 has finished its organization phase and that there is a relation that defines a one-to-one correspondence between the coordinates of peak activity generated by

SOM_0's training data set (i.e., $\{\mathbf{x}_{SOM_0}\}$) and the coordinates coming in on $\mathbf{v}_{1\,in}$. Initialization consists of letting the layout of the neural lattice TM_1 be identical to that of SOM_0 and arranging the weights, W_1, in TM_1 in such a way that the location of the peak activity in its postsynaptic field coincides with that of SOM_0 for corresponding inputs. TM_1 is thus to be seen a transformation network.

Upon completion of the respective training and weight assignment phases of SOM_0 and TM_1, $SumSOM_{Type1}$ essentially fuses its input pairs $(\mathbf{x}_{SOM_0}; (\mathbf{v}_{1\,in}, a_{1\,in}))$ together by transforming and superimposing the induced activity fields of its two neural networks, treating the result as an integrated response field and forming the fused output as the location and intensity of the maximum activity in this combined field. In detail, the postsynaptic activity fields in SOM_0 and TM_1, caused by their respective inputs on $\mathbf{x}_{SOM_0}$ and $\mathbf{v}_{1\,in}$, are forwarded as

$$\Phi_{SOM_0} = \Theta\,(W \cdot \mathbf{x}_{SOM_0};\, V)$$

and

$$\Phi_1 = \Theta\,(W_1 \cdot \mathbf{v}_{1\,in};\, V)\,.$$

Each element $\Phi^{(i,j)}_{SOM_0}$ (superscripts denote element indices) of the former field is fed through the function g, which makes changes in the postsynaptic activity field more prominent by transforming the activities into a field of linear and normalized measure of angular deviation. The function is defined as

$$g(\Phi^{(i,j)}, k_0) := \frac{\pi/2 - \arccos(\Phi^{(i,j)})}{k_0(\pi/2)}. \tag{2.4}$$

In the latter field, the function f is applied to each element $\Phi^{(i,j)}_1$, where

$$f(\Phi^{(i,j)}, \Phi, k_1) := \frac{\Phi^{(i,j)}}{k_1 \max(\Phi)}, \tag{2.5}$$

$\max(\Phi)$ returns the largest value in Φ, and naturally,

$$\Phi_1 = \Theta\,(W_1 \cdot \mathbf{v}_{1\,in};\, V)\,.$$

The resulting field is then element-wise multiplied with $a_{1\,in}$. After these multiplications, we have an activity field in which the location of peak activity has been reallocated from $\mathbf{v}_{1\,in}$ to $F\,(\mathbf{v}_{1\,in}; W_{TM_1}, V_{TM_1})$ (F from equation 2.1) while the magnitude of this activity remains equal to $a_{1\,in}$, that is, equal to the magnitude of the peak activity in the module providing $\mathbf{v}_{1\,in}$.

k_0 and k_1 are constants that regulate how much weight should be assigned to each of the two fields. Setting

$$k_0 = k_1 = c \in \mathbf{R}^+$$

results in weightings that yield fair combinations.

The two resulting fields are then superimposed by adding corresponding elements to each other, after which a function $\boldsymbol{h}$ is applied to each element. $\boldsymbol{h}$ is defined as

$$\boldsymbol{h}(\Phi^{(i,j)}, \Phi) := \frac{\Phi^{(i,j)}}{\max(\Phi, 1)}. \tag{2.6}$$

$\boldsymbol{h}$ thus acts as a saturating function that makes sure that the maximum activity does not exceed unity by rescaling the field if needed. Here, $\max(\Phi, 1)$ returns the largest value in the set formed by the union between the values in Φ and unity. The end result is the field Φ_{fused}, where

$$\Phi^{(i,j)}_{fused} = \boldsymbol{h}\big(\boldsymbol{g}\big(\Phi^{(i,j)}_{SOM_0}\big) + a_{1\,in}\boldsymbol{f}\big(\Phi^{(i,j)}_{1}\big)\big),$$

which is functionally comparable to those induced within the simpler self-organized modules containing only a two-dimensional neural lattice. The same methods of determining the maximum activity's position and the activity intensity can be used, thereby forming the unit's output $(\mathbf{v}_{out}, a_{out})$.

The other SumSOM type performs weighted signal fusion of outputs coming from two modules while giving responses according to a predefined template. The idea behind this SumSOM type is that it allows us to phenomenologically mimic the convergence of signals stemming from different cortical areas. A schematic is shown in Figure 3b.

While initializing this module, we assume a one-to-one correspondence between pairs of coordinates that are being input on $\mathbf{v}_{1\,in}$ and $\mathbf{v}_{2\,in}$. Additionally, the setup phase of a type 2 SumSOM needs an external template lattice that dictates the resolution of the SumSOM's response field and supplies a predetermined winner position for each corresponding pair of coordinates. Initialization begins with letting the neuron lattices in both TM_1 and TM_2 be of the exact same type as the template lattice (i.e., be of the same two-dimensional configuration). The weights of these networks are then assigned so that corresponding training signals on $\mathbf{v}_{1\,in}$ and $\mathbf{v}_{2\,in}$ yield the same positions of winner neurons in both networks and that these positions agree with the positions determined by the template.

The operation of a $SumSOM_{Type2}$, after it has been initialized, is similar to that of a $SumSOM_{Type1}$. Input signals received via $(\mathbf{v}_{1\,in};\ \mathbf{v}_{2\,in})$ induce activity fields in the two transformation maps TM_1 and TM_2. Before these activity fields are superimposed, both are transformed through the application of the function equation $\boldsymbol{f}$ as defined in equation 2.5 which is given

the arguments $(\Phi_1^{(i,j)}, \Phi_1, k_1)$ and $(\Phi_2^{(i,j)}, \Phi_2, k_2)$ in the two respective cases (where all variable names are local to this module). These altered fields are then element-wise multiplied with $a_{1\,in}$ and $a_{2\,in}$, respectively, after which they are superimposed. Before the location and magnitude of the peak activity are determined, the function $\boldsymbol{h}$, defined in equation 2.6, is applied for the same reason as before. The resulting Φ_{fused}, which here consequently is calculated as

$$\Phi_{fused}^{(i,j)} = \boldsymbol{h}\big(a_{1\,in}\boldsymbol{f}\big(\Phi_1^{(i,j)}\big) + a_{2\,in}\boldsymbol{f}\big(\Phi_2^{(i,j)}\big)\big),$$

is thus the unit's response field, and the position of the peak activity is output on $\mathbf{v}_{out}$ while its magnitude is output on a_{out}.

2.4 Stimuli. Before describing how we initialize this presented artificial neural network architecture and demonstrate our applications of it, we describe in more detail the stimuli that we use and how they are preprocessed.

The inputs $\mathbf{x}_{lt}$ to the letter processing module (see Figure 2) consist of 22-element vectors. The visual stimuli yielding these vectors are obtained from the representation of letters in the 12-point Times New Roman font. Preprocessing is carried out as follows. Each black and left-aligned font symbol is placed on white background and then transformed into a 21×25 pixel image. Afterward the collection of images is scanned vertically so that each letter is represented by a 525-element double-precision stimulus vector. In order to reduce computational time, we reduce the dimensionality of the letter stimuli. We employ principal component analysis (PCA) and discard the principal component scores of the components spanning the two dimensions in the dimensionally reduced subspace of least variance. Because the total number of letter stimuli is being equal to 23, the dimensionality of the transformed letter stimuli is thus reduced to 21. In the final preprocessing steps, we find the longest vector in the set of vectors containing the principal component scores and rescale all of these vectors with the inverse of its length, so that all vectors have a length shorter than or equal to unity. Then we project these rescaled vectors up on the unit hypersphere of dimensionality $N + 1$, where N is the dimension of the original vectors, by applying the vector-valued function

$$S(\mathbf{x}) := \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(N)} \\ \sqrt{1 - \sum_i \left(x^{(i)}\right)^2} \end{pmatrix} \tag{2.7}$$

to each vector containing our rescaled principal component scores. The end result is thus a set of $\mathbf{N}_{lt} = 23$ vectors, each containing $\mathbf{K}_{lt} = 22$ elements.

We have performed alternative dimensionality-reducing processing of the letters such as measuring their coincidence with vertical, horizontal, and diagonal lines. The subsequent self-organization of the SOM has the same characteristics as when principal components are used (not shown).

The preprocessed stimuli inputs $\mathbf{x}_{ph}$ to the auditory processing module consist of $\mathbf{K}_{ph} = 37$-element vectors. In our study, we use the phonemes as pronounced by a female Swedish speaker with particularly clear diction. She read a number of Swedish words from which we parsed the initial phoneme—$\mathbf{N}_{ph} = 23$ phonemes in all. A set of 36 mel-cepstral coefficients was determined for each phoneme. A discussion of the mel-cepstrum and its use in representation of speech is given by Gold and Morgan (2000). The mel-cepstrum, however, has one well-known disadvantage: the mel-cepstral coefficients are very sensitive to the addition of white noise to the speech sound. This is a crucial observation from our point of view because we are interested in studying the effects that the adding of noise to the stimuli has on integration, and this high sensitivity may blow these effects out of proportion. There are many ways of mitigating this problem, and we have chosen to use one of the many different desensitizing methods—that proposed by Tyagi & Wellekens (2005). As with the letter stimuli, we find the longest vector, rescale all vectors with its inverse length, and apply the vector-valued function S (see equation 2.7) to the results. This yields our $\mathbf{x}_{ph}$.

2.5 Architecture Initialization. Given that there exist different algorithms of self-organization and different ways of achieving the functional goals put forth in section 2.3, the initialization of the proposed artificial neural network architecture can be carried out in several ways. This section lays out one approach.

In essence, our approach consists of organizing the three maps discussed in section 2.1 and arranging the additional neuron weights in the SumSOMs, using data from the three self-organized modules, in such a way that congruent inputs produce correctly fused outputs.

Thus, as an initial step of initialization, the two sensory SOMs, SOM_{ph} and SOM_{lt}, are trained with their respective preprocessed stimuli. The former SOM is trained with the vectors derived from phonemes, while the latter is trained with the vectors derived from pictures of letter symbols (see section 2.4 for the details of these stimuli).

The training algorithm used is a variation of Kohonen's learning law. In this letter, we work with normalized stimuli located on hyperspheres. Therefore we can use the simple dot-product learning law (Kohonen, 2001) for training, in which case the update of a weight vector $\mathbf{w}_j$ for the jth neuron is described by

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta \cdot \Lambda \cdot (\mathbf{x}^T - \mathbf{w}_j(n)), \tag{2.8}$$

where η is the learning rate, Λ is a neighborhood function centered at the position of the winning neuron (see the appendix for details on model parameters such as the exact definition of Λ), and $\mathbf{w}_j(n)$ and $\mathbf{w}_j(n+1)$ are the current and the next weight vectors of the jth neuron, respectively. A normalization step can be carried out on demand,

$$\mathbf{w}_j(n+1) = \frac{\mathbf{w}_j(n+1)}{\|\mathbf{w}_j(n+1)\|},$$

in which case $\mathbf{w}_j(n+1)$ is thus overwritten.

After training, classification areas are determined using the procedure that follows. We present each training vector to the organized SOM, tag the neuron of maximal activity, and assign it to the training vector's known class. We call these neurons ideal neurons, and there thus exists one for each stimulus. We then go through all untagged neurons and determine to which tagged neuron's weights its own weights are closest and assign the untagged neuron to the same class. When the procedure is finished, every neuron has been assigned to a class, and all neurons belonging to the same class make up a classification area. This procedure is carried out for both SOMs (SOM_{ph} and SOM_{lt}).

SOM_{bm}, the third and the top-most module depicted in Figure 1, is then trained with a transformed version of concatenations of neuron coordinates from the two sensory modules belonging to congruent classification areas. The transformation consists of remapping the concatenations onto a unit sphere (of dimension 5) and is carried out as follows. All vectors are first rescaled in equal proportion so that the magnitude of the longest vector that can possibly arise is unity. This rescaling factor is thus

$$\frac{1}{\sqrt{4 \times \left(\frac{1}{\sqrt{2}}\right)^2}} = \frac{1}{\sqrt{2}},$$

as all coordinates are encoded as described by equation 2.2. Seeing that all vectors must now be on the unit disc of dimension 4, function S defined in equation 2.7 is applied to them. The desired transformation has thus been carried out.

The training of this module consists of two phases where it is trained only with concatenations of coordinates belonging to congruent ideal neurons initially. When the coarse organization has formed, the training set is augmented to cover all possible concatenations of coordinates belonging to congruent classification areas. Following the training, ideal neurons are determined by presenting the transformed concatenations of the congruent ideal neurons from the sensory modules and noting the winner coordinates in SOM_{bm}. Thereafter, classification areas are determined as described in the paragraph above. The described training approach decreases the risk of

obtaining a mapping that yields incorrect classifications of congruent coordinates while increasing the chance of having ideal neurons of this same SOM centered in classification areas.

Our next step is to build the full MMSON architecture, which we start by letting the bimodal association area be modeled by a type 2 SumSOM and the auditory processing area by a SumSOM of type 1. We place the trained SOM_{ph} inside the auditory module, let the layout of the neural lattice TM_1 in this module be identical to that of SOM_{ph}, and assign each neuron in TM_1 to the same classification class as its corresponding neuron in SOM_{ph}. SOM_{bm} serves as the template for the topology of the bimodal association area and is used as such when weights are assigned. Therefore, the bimodal processing module's transformation maps, TM_1 and TM_2, are set to have identical layout with SOM_{bm}, and the neurons in both maps are assigned to the same classification class as their corresponding neuron in SOM_{bm}.

This procedure of first organizing the feedforward network and then completing the artificial neural network architecture with feedback connections is in agreement with results from developmental studies of cortex. According to the work of Gogtay and colleagues (2004), primary sensory cortices mature first, while the multimodal integrative association areas in superior temporal cortex develop later. It is reasonable to assume that feedback connections from the latter to the former develop on the development of the latter. In visual cortex, it has been shown that feedforward connections develop earlier than feedback connections, "likely making their establishment dependent on sensory experience" (Kral & Eggermont, 2007).

We now turn to the assignment of weights in the transformation maps within the different modules. First, we assign the weights of TM_1 in the bimodal module. We collect the coordinates $\mathbf{y}_{lt_k}$ of all ideal neurons in SOM_{lt}, where $\mathbf{y}_{lt_k}$ contains the coordinates for the ideal neuron in response to stimulus k, and set the weights of all neurons of the kth classification area in TM_1 to $S(\mathbf{y}_{lt_k})$. This simple algorithm will make the classification area containing the peak activity in the bimodal module's TM_1 the same as that in SOM_{lt}. A corollary of this procedure is that all neurons belonging to a classification area in TM_1 will have equal activity, and if the classification area containing the peak activity in the two transformation maps is the same, then TM_1 will not yield a change of winner neuron coordinates.

The weights of the bimodal module's other transformation map, TM_2, are assigned in the following way. For every neuron of SOM_{ph}, we determine which classification area k it belongs to and concatenate its coordinates $\mathbf{y}_{ph}$ with the congruent ideal neuron's coordinates from SOM_{lt} (i.e., with $\mathbf{y}_{lt_k}$). We then apply the function S (defined in equation 2.7) to this resulting vector, present the result to SOM_{bm}, and note the coordinates $\mathbf{y}_{bm}$ of the winning neuron in this module. If the weights of the TM_2-neuron with coordinates $\mathbf{y}_{bm}$ are unassigned, or if $\mathbf{y}_{ph}$ contain the coordinates of an ideal neuron, we set these weights to $S(\mathbf{y}_{ph})$. If $\mathbf{y}_{ph}$ does not contain the

coordinates of an ideal neuron and the TM_2-neuron's weights are already assigned to, say, $\mathbf{w}$, we overwrite these weights with $S(\mathbf{y}_{ph})$ if

$$\|\mathbf{y}_{ph_k} - \mathbf{y}_{ph}\| > \|\mathbf{y}_{ph_k} - S^{-1}(\mathbf{w})\|,$$

where $\mathbf{y}_{ph_k}$ are the coordinates belonging to the ideal neuron for stimulus k and S^{-1} is the inverse of S. An earlier assignment of weights of a neuron in classification area k is thus overwritten if the distance from the area's ideal neuron $\mathbf{y}_{ph_k}$ to the coordinates $\mathbf{y}_{ph}$ is larger than the distance between $\mathbf{y}_{ph_k}$ and the previously assigned coordinates $S^{-1}(\mathbf{w})$.

Moving on to the last transformation map, TM_1 of the auditory module, we repeat the previous steps of presenting the S-transformed concatenation of each neuron's coordinates of SOM_{ph}, $\mathbf{y}_{ph}$ belonging to classification area k, with its congruent ideal neuron's coordinates from SOM_{lt} while noting the coordinate of the winner neuron $\mathbf{y}_{bm}$ in SOM_{bm}. Interwoven with this routine, we set the weights of the neuron located at coordinates $\mathbf{y}_{ph}$ in TM_1 to $S(\mathbf{y}_{bm})$. As a last step, when all neurons have been assigned weights, we go through the transformation map, visiting each classification area and checking for neurons with identical weights. We then nullify the weights of all neurons that have a nonunique weight vector, except for the one closest to the ideal neuron of the classification area. This last step is an attempt to improve stability and aesthetics; when possible, it will recursively move the location of maximum activity closer to the ideal neuron of the active classification area.

3 Results and Simulations

3.1 The Self-Organized Maps for Letters, Phonemes, and Bimodal Stimuli. A typical example of resulting mappings in the respective sensory modules representing letter and phoneme processing after self-organization is depicted in Figures 4a and 4b. When a training vector is presented to its corresponding SOM, the SOM's response field is highest from a population of neurons in a connected area, and thus these neuronal populations constitute the detectors of the respective stimuli. We call these areas classification areas, or simply patches, and the figure shows each of these areas corresponding to their respective stimuli (precise details of how these areas are determined are given in section 2.5).

The resulting map for the 23 letters in our material is shown in Figure 4a. We notice that visually similar letters such as i, l, and t are grouped together, a property that is a known characteristic of networks trained with Kohonen's self-organizing algorithm. Likewise, in the map for the corresponding phonemes, shown in Figure 4b, such phonemes that are similar, such as the plosives p, k, and t, are grouped together. Kohonen (1988, 2001) offers further discussions of self-organizing maps such as these, and the related "phonetic typewriter" from 1988.

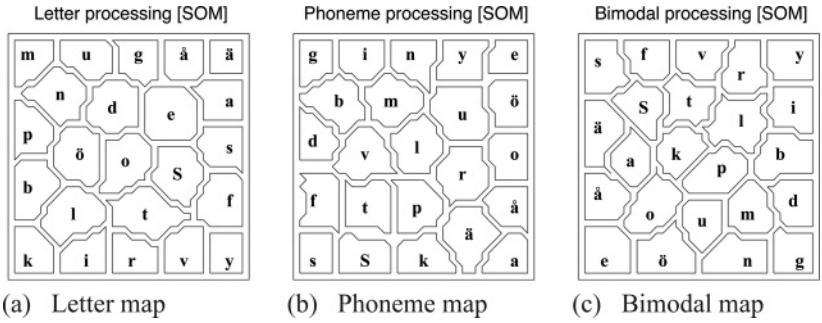


Figure 4: Areas of classification for labeled letters, phonemes, and letter-phoneme combinations after self-organization. The labels share their positions with the ideal neurons. In all three modules, the response field consists of the output signals of 36 × 36 neurons.

An example of the arrangement within the bimodal module, which is self-organized on concatenations of coordinates of congruent classification areas supplied by the sensory modules is depicted in Figure 4c. The similarity characteristics of this map are derived from the placement of the patches in the letter and phoneme maps and thus only indirectly reflect the features of the letters and phonemes. After the completed initialization procedure (detailed in section 2.5), the bimodal processing module thus integrates letter and phoneme stimuli.

3.2 Application Dynamics. During the application phase (after training and the additional initializations), the sensory stimuli are input to the artificial neural network architecture's sensory processing units via $\mathbf{x}_{lt}$ and $\mathbf{x}_{ph}$ (the stimuli are detailed in section 2.4). The two sensory units output the positions of their respective maximum activities, $\mathbf{v}_{lt}$ and $\mathbf{v}_{ph}$, along with the magnitude of these activities, a_{lt} and a_{ph}. These signals are then fused and classified in the bimodal module, and this classification, contained in $\mathbf{v}_{bm}$, is fed back to the auditory module, together with the peak activity level a_{bm} in the bimodal module's fused response field. The processing in the auditory module is modulated by this feedback, and this results in new output. This process of information flow, from the auditory module to the bimodal module and then back to the auditory module, will be referred to as a loop count. The first loop is, however, defined as the state after the first feedforward sweep, that is, the state reached just before the feedback modulates the auditory processing module. The state reached just before the feedback modulates the auditory module the second time will thus be referred to as the second loop, and so on.

The resulting recurrent process continues until it either converges or reaches a maximum number of allowed loops. At the outset, the feedback

signals are undefined, and thus one must assign some initial values in the auditory processing module. Our way of doing this is described in the appendix.

3.3 Application 1: Robustness of Auditory Recognition Against Sensory Noise. Because our inspiration for developing this artificial neural network architecture primarily originated from studies of cerebral cortex, we are naturally interested in closing the loop by exploring the behavior of our architecture that parallels cortical function. In this and the next three sections, we thus relate the dynamics of the artificial neural network architecture to psychophysical and brain imaging studies.

Our first application is identification of noisy phonemes and comparing performance between cases when the congruent letter stimuli are also available and when they are not. Human performance on identifying noisy phonemes in the presence of task-irrelevant letters has been extensively studied and recently reported on by Blau et al. (2008). In this study, the authors determined the effect of the letter in the identification of phonemes by presenting both congruent and incongruent letters, with and without visual noise, in conjunction with somewhat degraded phonemes to human subjects. The behavioral results revealed that "the speed and accuracy of speech sound identification critically depended on the quality of the visual stimulus (congruency-by-noise interaction) where facilitation/inhibition effects were reduced as a function of letter quality." fMRI analyses showed enhancement with congruent letters in both the multimodal STS area and the auditory association cortex.

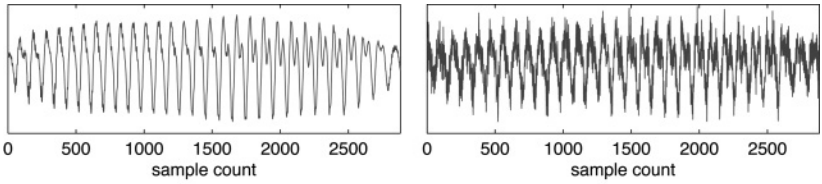
The same behavioral results have been obtained for continuous speech. Zekveld, Kramer, Kessens, Vlaming, & Houtgast (2008) studied speech comprehension when text from an automatic speech recognizer was displayed simultaneously or with small delays. The authors concluded that "people are apparently able to listen to speech in noise while simultaneously reading partly incorrect text to enhance speech comprehension."

Here we present an analysis of robustness of comprehension, based on simulations with our artificial neural network architecture. The analysis is carried out by forming new phoneme stimuli of the form

$$\mathbf{x}_{ph} = \beta \cdot \textbf{phoneme} + \alpha \cdot \textbf{white noise},$$

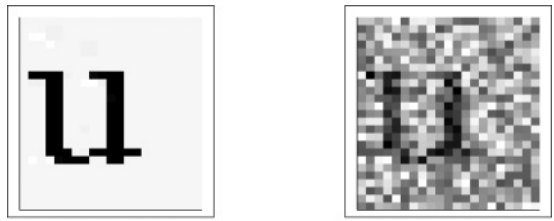
with $\alpha + \beta = 1$. α ranges from null to unity in steps of 0.05 (5%), and the noise is uniformly distributed. In Figures 5a and 5b, we show the waveform of the phoneme u unperturbed and with 35% noise (i.e., $\alpha = 0.35$, $\beta = 0.65$), respectively.

The collection of visual stimuli is augmented as well, with a noisy variant of each letter. The noisy variants are formed by using the ideal stimuli as starting points and then changing the intensity of each gray scale pixel



(a) Ideal phoneme 'u' (b) Perturbed phoneme 'u'

Figure 5: Ideal (left) and perturbed (right) auditory stimuli (phonemes). The method of perturbation consists of adding white and uniformly distributed noise to the ideal (i.e., originally recorded) waveform. See the text for further details.



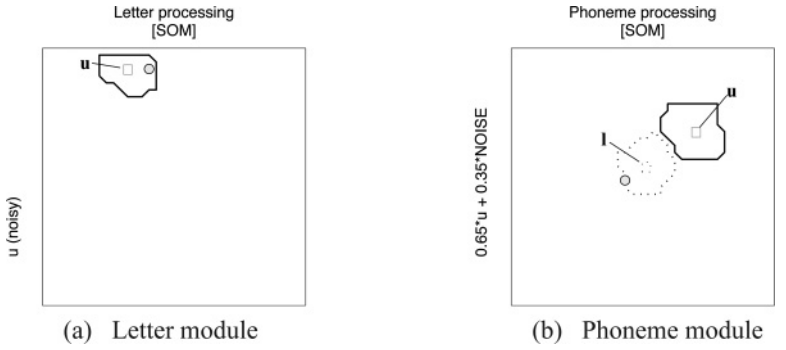
(a) Ideal letter 'u' (b) Perturbed letter 'u'

Figure 6: Ideal (left) and perturbed (right) visual stimuli (letters). The perturbed visual stimuli are constructed by taking the ideal stimuli and changing the contrast of each pixel. The amount of contrast change is dictated by a white and uniformly distributed random process. See the text for further details.

as dictated by a uniformly distributed random variable. The variable's range is between minus 60% and plus 60% of the pixels' intensity range (if the pixel value becomes invalid, we reflect the value through its nearest intensity range boundary point). In Figures 6a and 6b, one can see the ideal and perturbed versions of the letter u, respectively. We keep the same preprocessing procedure as described in section 2.4, which means that the preprocessed letter vectors lie in the Euclidean space of dimension 45.

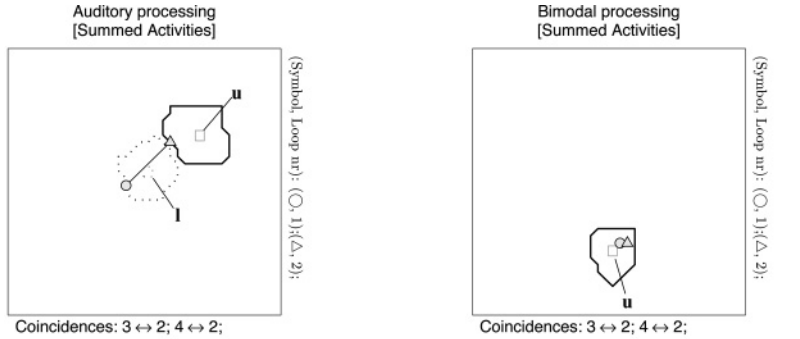
When preprocessed noisy stimuli are presented to the trained unimodal SOMs, the peak activity of the induced response fields is lower than when presenting ideal (learned) stimuli. As these peak activities are propagated to the sites of summation, the activity field derived from the noisiest source will be lowest and therefore have the least influence on the fused activity. This is in accord with the findings for other cases of sensory integration (Alais & Burr, 2004; Helbig & Ernst, 2007).

In what follows, an example of the artificial neural network architecture's developing dynamics after stimuli presentation will be shown. Figures 7a and 7b show the outcomes of presenting the noisy phoneme u (with 35% noise) and the congruent noisy letter to the sensory modules (i.e., the preprocessed versions of the stimuli shown in Figures 5b and 6b). As can be



(a) Letter module (b) Phoneme module

Figure 7: Sensory modules' responses to the noisy phoneme u (right) and its noisy letter counterpart (left). Patches drawn using continuous curves indicate the areas of classification for the stimuli classes the input to the network belong to, while the dotted patch indicates how the SOM actually classified the inputs. Rectangular symbols indicate the neurons responding to ideal stimuli, and the gray-filled circles denote the positions of peak activities during the current simulation.



(a) Auditory processing module (b) Bimodal processing module

Figure 8: Responses at the two sites of summation: in the auditory and bimodal processing modules. The gray-filled circle and triangle indicate the location of peak activity after the first and second loop counts, respectively. After loop count 2, the locations of peak activities do not move. Patch indications bear the same meanings as in Figure 7.

seen, the phoneme stimulus is so noisy that it is misclassified as an l. Because there are no feedback effects in these unimodal modules (see section 2.3 for the details of the auditory processing module), the initial activity fields remain the same throughout the simulation.

Figures 8a and 8b show the development at the sites of summation. The recurrent activity is brief, and this is typically the case. The locations of

initial maximum activities in the different modules are shown with circle symbols and the location of final peak activity with a triangle. In the summed activity of the bimodal module, the difference between the final and initial activities is small. The winning neuron changes within the same patch, but the classification thus remains the same. This, however, is not always so; the bimodal module as a whole is part of a recurrent loop, and the development can hence be more dynamic. The notable result of the recurrent activity here is that the peak of the combined activity of the auditory module moves from the l-patch to the u-patch. The noisy letter has thus corrected the classification in the auditory processing in our model.

Kislyuk, Möttönen, and Sams (2008) state, "Our findings show that visual stream can qualitatively change the auditory percept at the auditory cortex level, profoundly influencing the auditory cortex mechanisms underlying early sound discrimination." Thus there is a clear correspondence between our simulation result here and psychophysical experimental findings. The signal paths of our model that give rise to these results agree with those reported from the Department of Cognitive Neuroscience at Maastricht

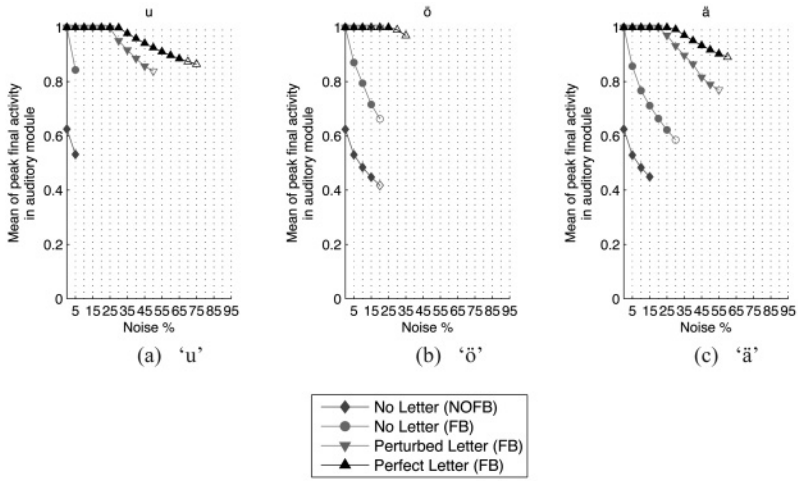


Figure 9: Limits of correcting efficiency of ideal and noisy letters to congruent noisy phonemes. The phonemes were perturbed by 16 different noise realization. Opaque markers denote cases when the congruent letter stimuli could correct all realizations, unfilled markers denote cases when at least one noise realization could not be corrected by the congruent letter stimuli, and curves end when no noise realization could be corrected by congruent letter stimuli. The variation of the peak activity due to different noise realizations was low (standard deviation < 0.015). FB and NOFB signify feedback and no feedback, respectively. See the text for details.

University (van Atteveldt et al., 2004, 2007; Blau et al., 2008; Froyen et al., 2008), as stated above, and they also agree with the anatomical examinations of the corresponding areas in macaque monkeys (Schroeder & Foxe, 2002; Schroeder et al., 2003).

In Figures 9a to 9c, we show the limits of the correcting efficiency of ideal and noisy versions of the letters u, ö, and ä to the congruent noisy phonemes and the mean peak neural activity level when the stimulus is correctly identified. We also show the contrasts in outcome when feedback exists versus when it has been removed. The curves with diamond markers show the performance of the raw input to the phoneme map, that is, how much noise can to be added before the phoneme input is perturbed enough to be misclassified by its sensory map.

Results agree well indeed with expectations: the effect of the feedback on the activity levels is an increase of the activity in the auditory module, while its effect on classifications is a bit more complex: congruent and uncorrupted letter stimuli have larger correcting capacity, but the very noisy letter stimuli (formed as explained previously in this section) also improve comprehension.

The results for the other letter-phoneme pairs are not identical, but the qualitative results in the plots shown are representative. The general result is that the feedback modulates the system by its ability to improve both the location of the winner neuron and its activity; in all cases, feedback increases the activity level in the auditory module, and both an uncorrupted and a very noisy letter aid comprehension of the corresponding phoneme.

3.4 Application 2: Feedback and Faster Neuronal Activity Gains. Recall from section 1 that in addition to gains in detection rate, the focus of the previous section, an important, highly related, and well-studied effect of multisensory integration is that of gains in reaction time (Hershenson, 1962; Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004; Molholm, Ritter, Javitt, & Foxe, 2004; Hecht, Reiner, & Karni, 2008). It is therefore of interest to examine what properties the artificial neural network architecture can exhibit from this point of view.

To this end, we examine the results shown in Figures 10a to 10c. These figures depict the effect of the feedback on the bimodal module at three different levels of noise; the three plots show the median (marker) and the range (bars) of the peak activity in the module as a function of loop count. The two measures contain data from only those runs in which the MM-SON architecture stabilizes with a correct classification of the multisensory stimulus. For reference, the light and dark dashed lines show the median activity levels for the scenarios with perfect letter stimuli and no letter stimuli, respectively, when the feedback is removed. Our results clearly indicate the detrimental effects of noise on the activity level and the ability of the feedback to have the system iteratively recover from the losses due to it, where the degree of improvement shows a clear differential dependence

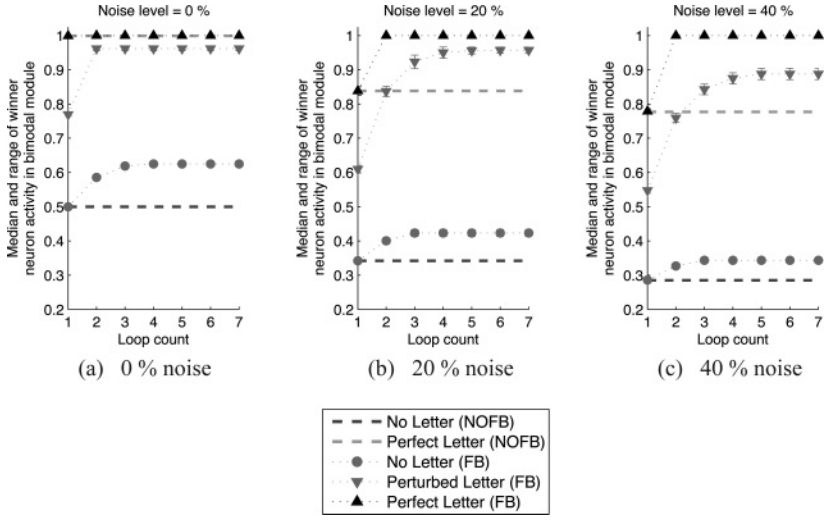


Figure 10: Gains in neural activity in the bimodal module due to integration at three different levels of noise. In each graph, the three curves with markers show how the peak activity in the auditory module develops as a function of the loop count. The median of the peak activities is indicated by the markers, and the bars show the ranges. FB and NOFB signify feedback and no feedback, respectively. See the text for details.

on the complementary letter stimuli. The architecture's explanation for this result is that increased noise causes lower activity in the auditory module at the outset, and this results in a delay in reaching higher activity in the bimodal module. Another result that can be discerned from these plots is that there is not only a delay of reaching the maximum activity but the asymptotic activity reached decreases when noise has contaminated both the auditory and the visual stimuli.

These simulation findings may be put into our context by imagining how our architecture's behavior exerts its surroundings. A neural network, located downstream from our architecture, responsible for response initiation, which may reasonably be modeled as an integrate-and-fire network that thus accumulates incoming signals over time until a threshold is reached (Ratcliff & Smith, 2004), would be affected in ways that are in agreement with experimental findings. In the noise-free scenarios, the multisensory cases would still yield faster accumulation of activity than the unisensory cases, and hence would reach a firing threshold faster and with a shorter response time. This is a well-known psychophysical result.

Barutchu et al. (2010) examined the effect of audiovisual integration in noise on reaction times. This recent study contains three interesting findings that can also be put in close relation to our artificial neural

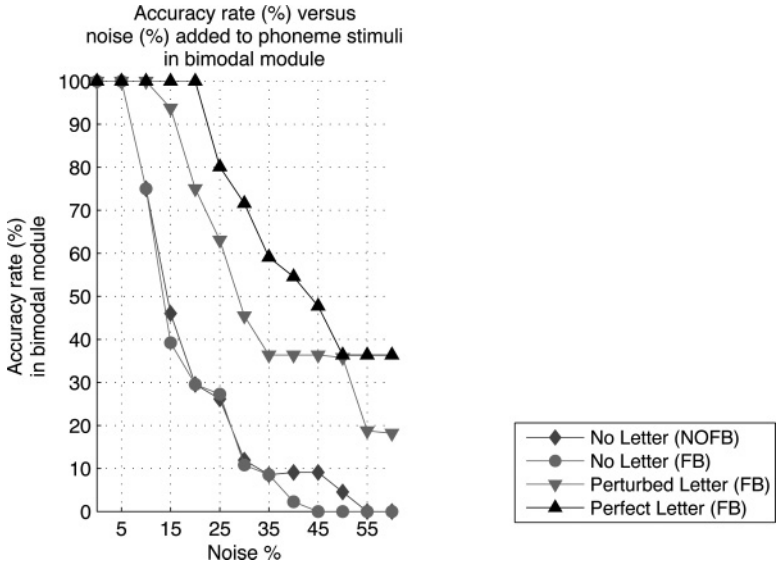


Figure 11: Overall classification accuracy in percentage of the bimodal module as a function of noise level.

network architecture's operation; reaction time length seems to be an increasing function of the level of noise, reaction times are shortened with multisensory stimuli even when they are noisy, and at moderate levels of noise, when multimodal information is available, the main effect is this lengthening on reaction time while accuracy remains high. Inspection of Figures 10a to 10c indicates that our architecture's behavior parallels the first two findings well; the downstream neural network receives less stimulation when noise levels rise, and it will take longer for its units to reach their threshold, but the stimulation is increased in the cases with multisensory stimuli. That the artificial neural network architecture also exhibits the third finding can be understood by inspecting Figure 11. The plot in this figure shows the overall classification accuracy rate, in percentage units, of the bimodal module as a function of the level of noise added to the phoneme stimuli. What is thus shown is the fraction of simulation runs when the MMSON stabilized with a correct classification. When combining the information in this plot with the information in Figure 10, one sees also the qualitative similarity between the last finding of Barutchu et al. (2010) and the results from our architecture; the classification rate of the bimodal module remains high even though the outgoing activity of the model decreases as noise is moderately increased (up to 20% to 30%, say).

3.5 Application 3: Simultaneously Presented Phonemes. One common situation where sensory integration is invoked arises when one is

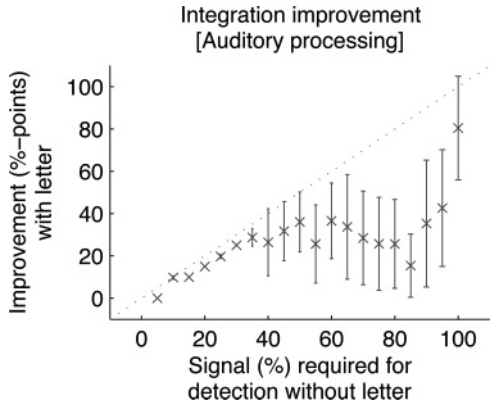


Figure 12: Results from the simultaneously presented phonemes scenario. The plot shows the average improvement (in percentage units) in detection threshold with the congruent letter stimuli as a function of the detection threshold level without the letter stimuli. The bars show the magnitude of one standard deviation. Data are collected from 20 different organizations of the artificial neural network architecture. See the text for details.

watching one speaker, while more than one speaker is heard (Sumby & Pollack, 1954). Here, we test the parallel scenario when two phonemes are spoken by the same speaker, recorded and played simultaneously, and at the same time, the letter congruent with one of the phonemes is shown. The question is: What is the smallest required fraction of the congruent phoneme for this phoneme to appear as the winner in the auditory processing module?

To answer this question, we present phoneme stimuli of the form

$$\mathbf{x}_{ph} = \alpha \cdot \mathbf{phoneme}_1 + \beta \cdot \mathbf{phoneme}_2,$$

with $\mathbf{phoneme}_1$ being the test phoneme, $\alpha + \beta = 1$, and a step size of 0.05 (5%). We are also interested in monitoring the difference that the presentation of the congruent letter yields. Consequently, we carry out this test with two visual conditions: with the letter stimuli congruent to $\mathbf{phoneme}_1$ and without any letter stimuli. From a purely technical point of view, the result from this experiment gives a perspective of how the perfect letter stimuli influence the auditory module through the feedback path. More specifically, it shows the corrective power of the perfect letter stimuli. The results are shown in Figure 12.

The plot shows the average improvement (in percentage units) in detection threshold with the congruent letter stimuli as a function of the detection threshold level without the letter stimuli. These data were collected from 20 different initializations of the artificial neural network architecture. Thus,

the uninitialized MMSON was replicated 20 times, and each was initialized and tested independently of one another. The variations of improvement due to the different initializations are indicated by the bars; each bar spans one standard deviation.

It should be mentioned that the task of identifying two simultaneously spoken phonemes without letters seems to be quite difficult for human subjects. We conducted a test with 12 male subjects, aged 25 to 40, all without hearing impairment. The subjects had been informed about the purpose of the test and given their consent. The test was carried out as follows.

Of our 23 phoneme stimuli, we selected a subset of 11 stimuli that had approximately the same duration, and we formed all the possible 50/50 combinations of them (55 in total) by superposing their waveforms. At the beginning of the test, each subject was presented with each of the 11 phoneme stimuli while being shown the congruent letter. Each stimulus was played four times interleaved with a 1 second pause. Then the subject was presented with half of the stimulus combinations (28/27 split). Each subject was assigned a sequence number (1–12). If the subject's sequence number was odd, the combinations presented were selected at random, while those with an even sequence number were presented with the combinations not presented to the subject with the immediately preceding sequence number. The combinations were played once, in randomized order, and the subject was given 10 seconds to write down what he had heard. All stimuli were presented binaurally using a pair of headphones. The average rate of correctly detecting both phonemes was 41% among these subjects.

For comparison, we extracted the exact parallel data from our simulations. To this end, we examined the response fields in the auditory processing module after presenting stimuli with equal mixtures of two phonemes. If the two classification areas with the highest activity coincided with the two stimuli in the mixture, we counted the result as a success and as a failure otherwise. The average success rate from the same 20 organizations as above was 29%. This discrepancy may depend on our way of preprocessing the stimuli. In particular, significant information may be lost due to the use of the mel-cepstrum transform, which comprises only magnitude, but no phase information.

3.6 Application 4: Activation of the Auditory Module by Visual Stimuli Alone. Studies on human cerebral cortex have shown activation of parts of auditory cortex during silent lip reading (Calvert et al., 1997; Pekkola et al., 2005). In this section, we demonstrate that such an effect of sensory integration is also manifest in our artificial neural network architecture, when visual speech consists of letter reading. Silence in our experiment is modeled by letting the response of each neuron in the phoneme SOM be random between 0 and 0.1 with a uniform distribution. In the experiment accounted for here, the realization of such a response field yielded peak activity in the patch classifying a. The outcome in the auditory and bimodal

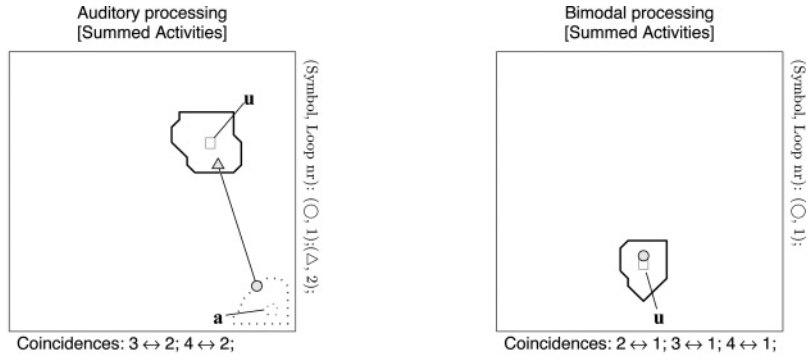


Figure 13: Induced response plots as a result of the MMSON receiving phonetic silence and the letter u. In the bimodal processing module, the silent stimulus has no significant effect. The feedback from the bimodal module becomes the dominant influence on the response field in the auditory processing module after loop 1. The peak activity in the auditory processing module is, however, significantly lower than it would be if the phoneme u were presented. Figure semantics are the same as in Figures 7 and 8.

processing modules induced by the presentation of the stimulus vector representing letter character u, and this particular response field realization is depicted in Figure 13. In the bimodal processing module, the patch representing u instantly becomes the most active one as the influence from the auditory module is too low to make any contribution of consequence. With the help of feedback, the maximum activity in the auditory processing module becomes located within the u patch after only one feedback loop. The peak activity level in the auditory processing is approximately 0.66, significantly lower than the level of 1 attained when the MMSON is presented with the uncorrupted phoneme u instead of silence.

Our interpretation of this experiment is that the artificial neural network architecture is able to exhibit yet another result gathered from studies of brain function: the auditory subnetwork becomes activated by visual stimulation, but the activation is lower than that achieved when auditory stimuli are used.

4 Discussion

The multimodal self-organizing network presented in this letter has been described as an artificial neural network architecture inspired by cortical brain studies related to sensory processing and sensory integration in general and to the processing of audiovisual speech stimuli in particular. The architecture was applied to artificial audiovisual speech

stimuli. Its performance and behavior were tested in phoneme recognition settings and the results compared to certain aspects of cortical letter-phoneme integration. We showed that the MMSON architecture exhibits important characteristics that are indeed similar to those obtained from studies on the brain within psychology and neuroscience.

4.1 Related Work. Existing approaches to modeling or simulating sensory integration are diverse indeed. In this section we review work that has been influential or is closely related to the work presented in this letter.

The model of the superficial and deep layers of superior colliculus by Casey and Pavlou (2008) employs self-organizing maps associated by Hebbian links that are trained using the activity product rule with renormalization (Haykin, 1998; Brown, Kairiss, & Keenan, 1990). The established associations forms an approximate translation from one of the map's postsynaptic coordinate spaces into the other's thereby facilitating the combination of postsynaptic activities, which forms a bimodal activity field. Whereas the main focus of the model is to demonstrate one way of organizing an association between the two maps, the authors' experimental goal is to simulate behavioral studies of sensory integration that show an enhancement of activity at the bimodal level when congruent multisensory inputs are presented. The model is purely unidirectional, with convergence existing at the bimodal level only.

Magosso, Cuppini, Serino, Pellegrino, and Ursino (2008) presented a different model of sensory integration within the superior colliculus. This model also consists of three sheets of equal size, each containing simplified artificial neurons. The sheets model visual, auditory, and multimodal areas, respectively. Here, though, the units are equipped with lateral connections whose weights are distributed in the Mexican hat manner. Model stimuli are represented as spatiotemporal activity distributions. Units in the unimodal areas have gaussian receptive fields, but they project only to one designated neuron in the multimodal area, respectively. Projections from the multimodal sheet to the unimodal ones are implemented in the same manner, where each multimodal unit projects back to only the units it receives input from. The units' activation rates are computed using a first-order transfer function and a static sigmoidal relationship. There is no learning in this model, as all links have predetermined weights and the artificial neural network architecture is thus hard-wired. Weights are, however, varied manually by the authors in order to model different integration phenomena, and the authors speculate that the reason for the range of characteristics exhibited by neurons in real neural networks is randomness in synaptic strengths.

The basic module of the model described by Rolls (2004) is a pattern association network that can be made to associate two stimuli using an associative (Hebb-like) learning rule. When three of these modules are interconnected, a recurrent artificial neural network architecture for audiovisual

processing is built, where two modules process sensory stimuli and multimodal convergence is carried out in the third. The multimodal module feeds its signals back to both sensory processing modules. Although the method of modeling is quite different from that described here, the modeling goals are rather similar. Three points are made: the individual modules exhibit the effects of facilitation, where responses are increased for congruent inputs, and suppression, where responses are suppressed for incongruent inputs. The model can also display both the behavior where one stimulus dominates over the other and a McGurk-type effect, where the stable state corresponds to neither of the input stimuli.

Martin et al. (2009) presented a self-organizing map model of multisensory integration. The multimodal input is modeled with purely artificial stimuli of low dimension. Inputs are made up of m component short-range integer vectors, where m is the number of modality types, while the integers model the degree of activation of a modality type. All tests are run on artificial data. The SOM self-organizes in such a fashion that unimodal stimuli activate peripheral regions of the SOM, and bimodal and trimodal stimuli activate regions with peaks between the unimodal activations. Since the model uses only one SOM, there is no top-down or lateral feedback in the model, and the output is determined in a static way. The authors give particular care to make the multimodal enhancement show inverse effectiveness. This is accomplished by adding a layer of sigmoid functions operating on the inner products delivered by the SOM.

Some work models integration by linking together pairs of self-organized maps using Hebbian synapses. An early example is the DISLEX model by Miikkulainen (1997), where phonological and orthographic maps are, respectively, linked to a semantic map as dictated by the co-occurrence of their respective input stimuli. The links are unidirectional, meaning that no recurrent dynamics exists. This mental lexicon model is shown to be able to mimic language dysfunction, semantic slips, category-specific aphasic impairments, and dyslexic behavior through selective lesion procedures. DevLex (Li, Farkas, & MacWhinney, 2004) and DevLex-II (Li, Zhao, & MacWhinney, 2007) are both based on DISLEX and model early lexical development. The former model focuses on capturing how the linguistic environment changes dynamically during language learning, while also offering psychological explanations for a handful of aspects of acquisition of language. It consists of two growing SOMs (Li et al., 2004), which respectively model phonological and semantic areas in a child's cerebral cortex, that are linked by bidirectional and bi-weighted Hebbian synapses. The latter artificial neural network architecture (Li et al., 2007) consists of three networks: an SOM modeling word form, an SOM modeling a semantic map, and a sequence map modeling the output (i.e., articulation) of words. The maps are linked together with unidirectional Hebbian synapses to model word comprehension and word production, meaning that there exist only two sets of intermap links. This model is also employed for explaining and

predicting various phenomena of early language learning, but here the main focus is on simulating the challenge toddlers face when they need to articulate phonemic sequences of words. Apart from the former architecture containing growing SOMs instead of Kohonen SOMs, both models' workings are very similar to that of Miikkulainen (1997); they use the same rules and strategies for determining both afferent and lateral synapses and the same functions for cross-map activation.

Mayor and Plunkett (2008) model early word learning using two self-organized maps that are linked by bidirectional, single-weight, Hebbian synapses. This yields a recurrent artificial neural network architecture that enables the modeling of cross-modal effects in yet another way. By varying different parameters, such as the onset time of the synaptic weight change of the Hebbian links relative to the two maps' self-organization progress, the count of epochs the Hebbian links are trained on, and the connectivity density, they offer explanations for different psychological findings of word learning in toddlers and young children (Mayor & Plunkett, 2010). The main modeling goal is to link words and objects together in a taxonomic way, so that the presentation of an object should coactivate the correct corresponding word, while the presentation of a word should coactivate the correct set of corresponding objects. Because the primary focus is on offering psychological explanations, few technical details are described. Nevertheless, as the authors use a different but comparable approach to model sensory integration, it would be interesting to carry out architectural comparisons between our model and theirs.

Mayor and Plunkett coauthor a paper where the impact of auditory labels on infants' visual categorization is modeled (Gliozzi et al., 2008). The basic building block of the modeling architecture is again the SOM, but here the authors suggest a different approach for modeling the integration of the two sensory modalities. After two SOMs have been trained on their respective unimodal stimuli, integration is modeled by training a bimodal SOM (called "global") on the unimodal SOM's activity patterns in response to congruent stimuli (activity is defined in a similar way as in this letter). This integration approach may be compared to that used in our feedforward artificial neural network architecture (section 2.1), where we train our bimodal SOM on concatenated coordinates belonging to the units in the unimodal SOMs that respond most strongly to respective congruent input stimuli.

A quite early work drawing ideas from neural network models and sensory integration for improving the classification of auditory input was described by Yuhas, Goldstein, and Sejnowski (1989); Yuhas, Goldstein, Sejnowski, and Jenkins (1990). The research topic of our letter has thus been of interest for quite some time. Studying integration processes of auditory and visual signals from the artificial neural network perspective is interesting not only for replicating aspects of brain function but also for advancing the technical field of automatic speech recognition. The system

presented by Yuhas et al. (1989, 1990) is essentially a feedforward network trained in a completely supervised fashion. In this system, which serves as a vowel recognizer, a multilayer perceptron (MLP; Haykin, 1998) is used to classify vowels using short-term spectral amplitude envelopes (STSAE) of the phonemes representing the vowels. To improve the STSAEs of noise-contaminated versions of these vowel phonemes, the system uses estimates of the envelopes, which are derived from a complementary information source: the visual images of the mouth uttering the same phonemes. This is done by training another MLP to predict STSAEs using the visual images. The estimates stemming from the two complementary information sources are integrated by forming weighted averages of them, where the weights are dependent on the noise level.

Finally, we mention the work of Sit and Miikkulainen (2006, 2009), who have developed a hierarchical artificial neural network architecture based on self-organizing neural lattices that aims to model the neural organization in V1 and V2. Despite some resemblances of that work and ours, it differs from the work we have presented here in several ways. Our work's inspiration is focused around the functional aspects of multisensory processing in cortical areas higher in the cortical hierarchy, while the work of Sit & Miikkulainen focuses on examining the organization of synaptic weights in lower levels in the cortical hierarchy, which process more elementary unimodal stimuli. The units in their model receive afferent and lateral inputs gated by synaptic weights that have been self-organized using Hebbian learning with divisive postsynaptic normalization. The earlier model (Sit and Miikkulainen, 2006) also contains feedback connections but these have been removed in the more elaborate description (Sit and Miikkulainen, 2009). Each unit receives afferent input from a subset of units in the neural lattice located at the immediately lower hierarchical level, while model input consists of low-level sensory data: pictures coded as intensity maps. The main contributions in that work, apart from the presented artificial neural network architecture, are a correspondence between the organized orientation maps of the model and of those found in cortices of monkeys, and predictions about connection properties (synaptic weight distributions) of neurons in V1 and V2 and what kind of stimuli V2-neurons might prefer. Thus, in spite of analogies between the work presented here and that of Sit and Miikkulainen (2006, 2009), both the work methods and the aims of the developed artificial neural network architectures are quite dissimilar.

4.2 Opportunities for Future Work. In this letter, we have studied the processing of noisy but congruent stimuli. It is obvious that the inputs in the simulations may as well be incongruent and the occurrence of a McGurk (McGurk & MacDonald, 1976) type effect is then to be expected. This is currently being studied.

In section 2.1 we observed that complementary cues are combined in such a way that more reliable cues are given greater weight in the integrated percept. Psychophysical experimenters have presented more precise information on this matter. Investigations of sensory integration with incongruent visual and haptic information where the visual stimulus is blurred to varying degrees have shown a smooth transition between which of the two information sources dominate the subject's percept (Ernst & Banks, 2002; Helbig & Ernst, 2007). Similar experiments have been performed in the audiovisual setting where the task was to localize a sound source, and the same effects were observed: either vision dominated the percept, or audition, or a weighted average of both information sources was formed (Alais & Burr, 2004). These results and other evidence (Ernst & Bülthoff, 2004; Cheng, Shettleworth, Huttenlocher, & Rieser, 2007) have formed the consensus of opinion stating that the brain estimates the quality of the incoming sensory data and integrates them in a statistically optimal (in the likelihood sense; see the tutorial by Myung, 2003) way (Ernst & Banks, 2002). In short the maximum likelihood estimation (MLE), which can be derived from Bayes' theorem (1763), is a specification of how to combine multiple sources of information in a way that minimizes the variance of the combined result, under certain assumptions (see, e.g., the appendix of the book chapter by Yuille & Bülthoff, 1996, and Jacobs, 2002, for more information). Seeing that the presented artificial neural network architecture exhibits several major characteristics of sensory integration, a natural next step is to examine how the properties of optimality can be incorporated as well.

The above studies pertain to sensory integration; the multimodal self-organized network is not restricted to such studies, however. It may be applied to any number and combination of stimulus modalities and to any depth of the processing hierarchy. It can also accommodate any number of feedback connections. As was shown in the case of letter reading combined with silence, the artificial neural network architecture gracefully accepts a zero input in one or more modalities. In a network with several hierarchical levels, we may then simulate the situation where there are no sensory inputs but the network is stimulated at its top level (presumably through some thought process). In such a process, called imagery (Mechelli, Price, Friston, & Alumit, 2004), it is known from brain activity imaging studies that activity spreads to levels lower in the hierarchy, perhaps even to the primary sensory cortices (Slotnick, Thompson, & Kosslyn, 2005). In the multimodal self-organized network, this is made possible through the feedback paths provided in the architecture.

Appendix: Model Parameters

When the three modules using the learning law stated in equation (2.8) are being trained, the learning rate η is set to decrease exponentially as a

function of number of epochs passed ε,

$$\eta(\varepsilon) = e^{-\frac{9}{2}\left(\frac{\varepsilon-1}{\varepsilon_{total}-1}\right)^2}, \tag{A1}$$

where ε_{total} is the total number of epochs the modules are trained for.

The neighborhood function Λ also shrinks as the number of epochs increase. If $\boldsymbol{\zeta}$ is the coordinate of the winning neuron, then the neighborhood function's value at coordinates $\mathbf{y}$ is

$$\Lambda(\varepsilon, \boldsymbol{\zeta}, \mathbf{y}) = e^{-\lambda(\varepsilon)\sum_i \left(y^{(i)}-\zeta^{(i)}\right)^2}, \tag{A2}$$

where $\lambda(\varepsilon)$ is the function describing the shrinkage,

$$\lambda(\varepsilon) = \frac{1}{2(\lambda_i - (\lambda_i - \lambda_f)\left(\frac{1}{2}\left(\frac{\varepsilon-1}{\varepsilon_{total}-1}\right)\right)^{\delta})^2}, \tag{A3}$$

with $\lambda_i = 1$, $\lambda_f = 0$ and $\delta = \frac{1}{8}$.

The maps within the two sensory modules are trained for 8000 epochs each. As explained in section 2.5 the bimodal map is trained in two phases. The map is trained for 50,000 epochs. During epochs 1 to 20,000, it is trained using concatenations of congruent ideal neuron coordinates only, and the training set is then extended to encompass all concatenations of coordinates belonging to congruent stimuli areas.

The weights k_0 (used in function $\mathbf{g}$, equation 2.4) and k_1 (used in function $\mathbf{f}$, equation 2.5) in the auditory processing module are both set to 1.6. The same weight is used in the $\mathbf{f}$ functions of the bimodal module. When $\mathbf{g}$ is applied to the letter module's response field, k_0 is set to unity.

As mentioned in section 2.2, feedback from the bimodal module is necessarily undefined when the stimuli are first presented to the MMSON. We have chosen to remedy this by defining the initial response field of TM_1 in the auditory processing module to be all null. We also rescale the initial activity field in the bimodal module, which is thus induced during the first feedforward sweep, by the factor

$$\frac{max\left(\Phi_{SOM_0}\right) + max\left(\Phi_{SOM_{lt}}\right)}{2max\left(\Phi_{fused_{BP}}\right)}. \tag{A4}$$

Here, $max\left(\Phi_{SOM_0}\right)$, $max\left(\Phi_{SOM_{lt}}\right)$, and $max\left(\Phi_{fused_{BP}}\right)$ are the maximal (i.e., peak) activities in the auditory module's SOM_0 feature map, the feature map organized on letters, and the original response field of the bimodal module, respectively. This last step, which thus sets the peak activity of the bimodal module's initial response to the mean of the peak activities in the two unisensory maps, prevents diminishing the influence of the feedback

signal on the initial fused response of the auditory processing module. Without this measure, the peak activity of the bimodal module's initial response, and hence of the the auditory module's transformation map TM_1, would in most cases be only a fraction of the peak activity in the auditory module's feature map SOM_0.

Acknowledgments

We acknowledge the encouraging support from Jerker Delsing at Luleå University of Technology and the financial support of the Ph.D. Polis collaboration program between Luleå University of Technology and Monash University in Melbourne.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262.

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Exp. Brain Res., 166*, 559–571.

Barutchu, A., Danaher, J., Crewther, S. G., Innes-Brown, H., Shivdasani, M. N., & Paolini, A. G. (2010). Audiovisual integration in noise by children and adults. *Journal of Experimental Child Psychology, 105*, 38–50.

Bayes, T. (1763). Essays towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society (1683–1775), 53*, 370–418.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P.S.F., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10*, 512–528.

Blau, V., van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2008). Taskirrelevant visual letters interact with the processing of speech sounds in heteromodal and unimodal cortex. *European Journal of Neuroscience, 28*, 550–509.

Brown, T. H., Kairiss, E. W., & Keenan, C. L. (1990). Hebbian synapses: Biophysical mechanisms and algorithms. *Annual Review of Neuroscience, 13*, 475–511.

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. *Cognitive Brain Research, 10*, 349–353.

Calvert, E. G., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S.C.R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*(5312), 593–596.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visual speech. *J. Cognitive Neuroscience, 15*(1), 57–70.

Calvert, G., Campbell, R., & Brammer, M. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649–657.

Calvert, G. A., & Thesen, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *J. Physiology Paris, 98*, 191–205.

Casey, M., & Pavlou, A. (2008). A behavioral model of sensory alignment in the superficial and deep layers of the superior colliculus. In H. He & X. Xu (Eds.), *IEEE International Joint Conference on Neural Networks (IJCNN), 2008 (IEEE World Congress on Computational Intelligence)* (pp. 2750–2755), Piscataway, NJ: IEEE.

Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological Bulletin, 133*, 625–637.

Chou, S., Papliński, A. P., & Gustafsson, L. (2007). Speaker-dependent bimodal integration of Chinese phonemes and letters using multimodal self-organizing networks. In *Proc. 20th Int. Joint Conf. Neural Networks*. Piscataway, NJ: IEEE.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage, 24*, 21–33.

Dijkstra, T., Frauenfelder, U. H., & Schreuder, R. (2004). Bidirectional grapheme-phoneme activation in a bimodal detection task. *J. Physiology Paris, 98*, 191–205.

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on "sensory-specific" brain regions, neural responses, and judgments. *Neuron, 57*, 11–23.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8* 162–169.

Foxe, J. J., & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport, 16*(5), 419–423.

Frost, R., Repp, B., & Katz, L. (1988). Can speech perception be influenced by simultaneous presentation of print? *J. Mem. Lang., 27*, 741–755.

Froyen, D., van Atteveldt, N., Bonte, M., & Blomert, L. (2008). Cross-modal enhancement of the mmn to speech-sounds indicates early and automatic integration of letters and speech-sounds. *Neuroscience Letters, 430*, 23–28.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *J. Cognitive Neuroscience, 12*, 495–504.

Ghazanfar, A. A., & Shroeder, C. E. (2006). Is neocortex essentially multisensory? *TRENDS Cog. Sci., 10*, 278–285.

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2008). The impact of labels on visual categorisation: A neural network model. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 397–402). Austin, TX: Cognitive Science Society.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *PNAS, 101*, 8174–8179.

Gold, B., & Morgan, N. (2000). *Speech and audio signal processing*. Hoboken, NJ: Wiley.

Gustafsson, L., Jantvik, T., & Papliński, A. P. (2007). A multimodal network for sensory integrationof letters and phonemes. In *Proceedings of the 11th International*

Conference on Artificial Intelligence and Soft Computing (pp. 25–31). Calgary, AB: Acta Press.

Gustafsson, L., & Papliński, A. P. (2006). Bimodal integration of phonemes and letters: An application of multimodal self-organizing networks. In *Proc. Int. Joint Conf. Neural Networks* (pp. 704–710), Piscataway, NJ: IEEE.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall PTR.

Hecht, D., Reiner, M., & Karni, A. (2008). Multisensory enhancement: Gains in choice and in simple response times. *Experimental Brain Research, 189*, 133–143.

Helbig, H., & Ernst, M. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research, 179*(4), 595–606.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology, 63*(3), 289–293.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences, 6*(8), 345–350.

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate crossmodal information? *Brain Struct. Funct., 212*, 121–132.

Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience, 20*, 2175–2184.

Kohonen, T. (1988). The "neural" phonetic typewriter. *Computer, 21*, 11–22.

Kohonen, T. (2001). *Self-organising maps* (3rd ed.). Berlin: Springer-Verlag.

Kral, A., & Eggermont, J. (2007). What's to lose and what's to learn: Development under auditory deprivation, cochlear implants and limits of cortical plasticity. *Brain Resarch Reviews, 56*, 259–269.

Lamme, V., & Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neuroscience, 23*, 571–579.

Landy, M., Maloney, L., Johnston, E., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Res., 35*, 389–412.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158*(4), 405–414.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks, 17*, 1345–1362.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science, 31*, 581–612.

Magosso, E., Cuppini, C., Serino, A., Pellegrino, G. D., & Ursino, M. (2008). A theoretical study of multisensory integration in the superior colliculus by a neural network model. *Neural Networks, 21*(6), 817–829.

Martin, J. G., Meredith, A. M., & Khurshid, A. (2009). Modeling multisensory enhancement with self-organizing maps. *Frontiers in Computational Neuroscience, 3*, 8.

Mayor, J., & Plunkett, K. (2008). Learning to associate object categories and label categories: A self-organising model. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 697–702). Austin, TX: Cognitive Science Society.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review, 117*(1), 1–31.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mechelli, A., Price, C., Friston, K., & Alumit, I. (2004). Where bottom-up meets top-down: Neuronal interactions during perception and imagery. *Cerebral Cortex, 14*, 1256–1265.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language, 59*(2), 334–366.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14*(4), 452–465.

Möttönen, R., Calvert, G. A., Jääskeläinen, I., Matthews, P. M., Thesen, T., Tuominen, J., et al. (2005). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *NeuroImage, 30*, 563–569.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*, 90–100.

Papliński, A. P., & Gustafsson, L. (2005). Multimodal feedforward self-organizing maps. In *Lecture Notes in Compute Science* (pp. 81–88). Berlin: Springer.

Papliński, A. P., & Gustafsson, L. (2006). Feedback in multimodal self-organizing networks enhances perception of corrupted stimuli. In *Lecture Notes in Artificial Intelligence* (pp. 19–28). Berlin: Springer.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Möttönen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: An FMRI study at 3t. *NeuroReport, 16*, 125–128.

Polk, T. A., & Farah, M. J. (1998). The neural development and organization of letter recognition: Evidence from functional neuroimaging, computational modeling, and behavioral studies. *PNAS, 98*, 847–852.

Polk, T. A., Stallcup, M., Aguire, G. K., Alsop, D. C., D'Esposito, M., Detre, J. A., et al. (2002). Neural specialization for letter recognition. *J. Cognitive Neuroscience, 14*(2), 145–159.

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. *J. Anat., 197*, 335–359.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron, 28*, 617–625.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333–367.

Remez, R. (2005). Three puzzles of multimodal speech perception. In E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), *Audiovisual speech* (pp. 12–19). Cambridge, MA: MIT Press.

Rolls, E. T. (2004). Multisensory neuronal convergence of taste, somatosensory, visual, olfactory, and auditory inputs. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 311–331). Cambridge, MA: MIT Press.

Schroeder, C. E., & Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research, 14*(1), 187–198.

Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N., & Hackett, T. A. (2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology, 50*, 5–17.

Sit, Y. F., & Miikkulainen, R. (2006). Self-organization of hierarchical visual maps with feedback connections. *Neurocomputing, 69*, 1309–1312.

Sit, Y. F., & Miikkulainen, R. (2009). Computational predictions on the receptive fields and organization of V2 for shape processing. *Neural Computation, 21*, 762–785.

Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2005). Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral Cortex, 15*, 1570–1583.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Taylor, M. (1962). Figural after-effects: A psychophysical theory of the displacement effect. *Canadian Journal of Psychology, 16*, 247–277.

Tyagi, V., & Wellekens, C. (2005). On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (pp. 529–532). Piscataway, NJ: IEEE Press.

van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex, 17*, 962–974.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43*, 271–282.

Yuhas, B., Goldstein, M. H., & Sejnowski, T. (1989). Integration of acoustic and visual speech signals using neural networks. *Communications Magazine, IEEE, 27*, 65–71.

Yuhas, B., Goldstein, M. H., Sejnowski, T., & Jenkins, R. (1990). Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE, 78*, 1658–1668.

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–161). Cambridge: Cambridge University Press.

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M.S.M.G., and Houtgast, T. (2008). The benefit obtained from visually displayed text from an automatic speech recognizer during listening to speech presented in noise. *Ear and Hearing, 29*, 838–852.

Received July 2, 2010; accepted December 24, 2010.