In R.L. Goldstone, P.G. Schyns & D.L. Medin (Eds.) (1998).
Psychology of Learning and Motivation, Vol. 36.
San Diego, CA: Academic Press.

Perceptual Learning from Cross-Modal Feedback

Virginia R. de Sa and Dana H. Ballard

August 2, 1998

1 Introduction

Ultimately we must understand how humans and animals are able to learn the complicated tasks they do. An important component of that learning process is the learning of how to form useful categories from sensory data. Thus the focus of this chapter is that of learning to classify—learning to recognize that particular patterns belong to the same class which is different from the set of classes that represent other patterns. That such learning can be difficult is illustrated by a commonly used two-dimensional vowel dataset taken from Peterson and Barney (1952)[1], shown in Figure 1. The data represent different utterances of the common vowels of english. As you can easily see, the distributions from different classes overlap making error-free classification impossible and simple clustering non-optimal.

Learning algorithms for classification have been the subject of study in the field of pattern recognition since the 1950s. Such algorithms attempt to find decision boundaries that best divide the distributions of exemplars from different classes. More recently such learning algorithms have been cast in the form of networks. These algorithms are termed "neural networks" since their primitives are models for biological neurons. One advantage of casting the algorithms in this form is that they can be more directly related to biological neural processing and development.

Such algorithms usually work by adjusting the connection strengths between model neurons (or units) and for that reason are termed *connectionist algorithms*. There have been many examples of applications of connectionist algorithms that are able to provide insight into the computational processes in the brain. The technique used in (Lehky & Sejnowski, 1988; Zipser & Andersen, 1988) is to train a back-propagation (Werbos, 1974; Rumelhart, Hinton, & Williams,1986) network with many examples of hypothesized inputs and desired outputs and then observe the resulting receptive field structure (as determined by the pattern of connectivity from the input neurons). The logic behind this work is that if a network constructed to do a task using abstract neuronal elements results in receptive fields similar to those observed in the brain, it is possible that the brain is performing a similar computation. While this argument has led to many computational insights, the advances in understanding the brain are somewhat diminished in the use of a biologically implausible learning algorithm that cannot address the question of how the particular computational network might develop in the brain. The two major drawbacks with back-propagation as a model of neural learning are first, the necessity of the target output with each input pattern, and second, the requirement of the "back-propagation" of the error measures to neurons in the hidden layers (those that are neither inputs or outputs). The first problem of requiring desired outputs to learn cognitive tasks is not unique to back-propagation or connectionist modeling in general. Many cognitive models assume that the answer is somehow magically available to the learner during

[1]The original dataset contains more dimensions (formant frequencies) but the two-dimensional version is commonly used for benchmarking algorithms.

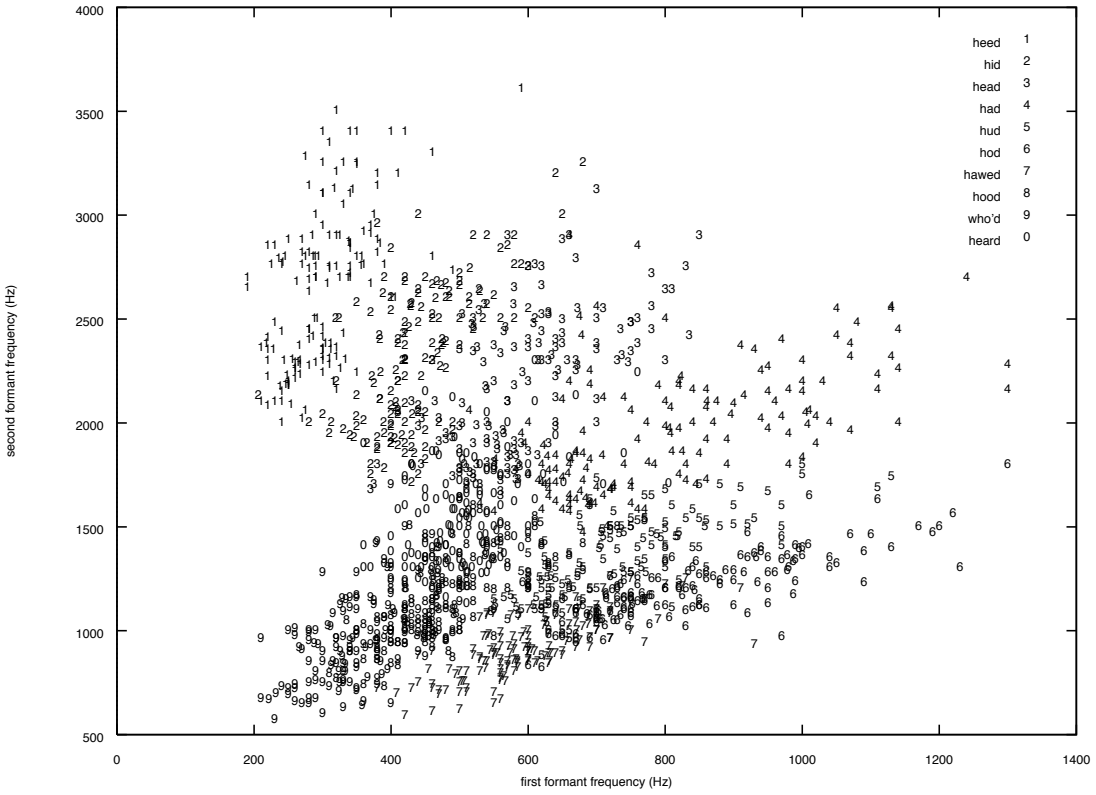

Figure 1: **The Peterson-Barney vowel dataset.** Each number (0-9) represents a different vowel class. Each plotted number corresponds to the formants measured from one instance of the uttered vowel plotted in terms of the first two formant frequencies (F1 and F2).

learning. These algorithms are called *supervised* algorithms and are tantamount to assuming that the desired mapping to be learned has already been learned by some entity whose output is available to the output of the training model!

Many models of the development of the early areas of the visual system (for example Linsker, 1986c; 1986a; 1986b; Miller, Keller, & Stryker, 1989; Obermayer, 1990; Obermayer, Ritter, & Schulten, 1992) use more biologically plausible learning rules. These models provide a mechanism for the development of the orientation columns, ocular dominance columns and even the particular patterns of interaction observed. The learning algorithm in these models depends only on the correlations between input patterns, and maps very well to neurophysiological ideas about the rules governing neural plasticity. The correspondence between the biologically motivated network models of the first visual cortical area (V1) and the known anatomy and physiology of monkey and cat V1 are extremely striking. However, they have been limited to modeling only this and earlier stages of processing and do not address the particular cognitive task in which an area may be participating. It seems that in order to learn high-level cognitive tasks, some indication of the desired outputs are needed during training. In other words,

> ...purely unsupervised learning based strictly on statistics, does not lead to conceptualization. This is due to the implicit assumption that every distinguishable input state potentially carries a different message. In conceptualizing, on the other hand, different input states which carry the same useful message are grouped together. This grouping

> requires some further knowledge that distinguishes signal from noise, or provides a measure of closeness on the signal space (Kohonen, 1984), or provides active supervision as in perceptual learning. (Redlich, 1993, p. 302)

This raises the question — Is there some way to model high level cognitive tasks using more biologically plausible learning algorithms? This chapter attempts to address this question for the particular task of object recognition or classification. The major idea is that, while it is implausible that the neuronal output label is available, it is true that the environment is providing extra information not currently considered in most unsupervised algorithms (algorithms that do not provide the target signal). We argue that the world provides much more information at a global level than is available to any one sensory modality and that this information can be used to learn classifications within each modality. We use this to address the problem of required supervision in developing the *Minimizing-Disagreement* algorithm. The model is based on gross cortical anatomy and achieves powerful classification results using biologically plausible computations without requiring the correct output signal.

We begin by clarifying the inherent weakness of unsupervised algorithms for classification and discuss various versions of adding supervision to the basic competitive algorithm.

1.1 Unsupervised Learning

A general way of representing sensory inputs is in terms of n-dimensional points, or vectors, groups of which can be summarized with prototypes or *codebook vectors* (because their purpose is to recode the input). A simple classification border between two such codebook vectors representing different classes is the $(n-1)$-dimensional hyperplane midway between them. With several codebook vectors per class and several classes, non-linear boundaries may be devised by taking the border from the *Voronoi* tessellation of these prototype points. Each codebook vector is assigned a class label and patterns are classified as belonging to the class of the closest codebook vector. These classifiers are often termed piecewise-linear classifiers for the shapes of their classification borders.

In learning algorithms, classification borders are moved indirectly by moving the codebook vectors. Competitive learning (Grossberg, 1976a, 1976b; Kohonen, 1982; Rumelhart & Zipser, 1986) is an unsupervised, biologically plausible (Coultrip, Granger, & Lynch, 1992; Miikkulainen, 1991) way of achieving this for easily separable data clusters. In Competitive learning, different neurons become responsive to different input patterns. If there are more patterns than output neurons within a competing group, then similar patterns (where similarity is defined in terms of closeness in the input space) will map to the same competitive neuron. In this way competitive learning algorithms learn to group similar inputs without a supervisory signal.

1.2 Supervised Learning

The natural grouping produced by competitive learning is a disadvantage in many classification tasks that require the mapping of dissimilar inputs to similar outputs. In order to achieve classification based on a task-dependent semantic similarity as opposed to similarity of the inputs, it is necessary to provide more information. A common form of additional information is the label of the correct class for each input pattern, which is provided during a training phase.

One simple use of the labels is to use them to determine the cluster to which each datapoint belongs. In Figure 2 we term this approach Supervised Competitive learning. A more powerful use of the labels is to monitor the number of misclassified patterns and minimize that number. Kohonen's (1990) LVQ2.1 algorithm is designed to find optimal borders in this way (though see de Sa & Ballard,1992; Diamantini & Spalvieri, 1995). The idea is to move the codebook vectors so that

the border between them moves toward the crossing point of the distributions. The distributions are sampled near the currently estimated border in order to decide which way to move the border.

The LVQ2.1 learning rule moves weight vectors only when exactly one of the two closest weight vectors is from the correct class. Then, as long as the pattern falls within a "window" around the current border, the weight vector from the correct class is moved toward the pattern and the one from the other class is moved away.

It can be described informally as:

> If the pattern is near a current border, move the codebook vector of the same class towards the pattern and that of the other class away.

or more rigorously,

$$\vec{w_i}(t+1) = \vec{w_i}(t) - \epsilon(t)(\vec{\xi}(t) - \vec{w_i}(t))$$

$$\vec{w_j}(t+1) = \vec{w_j}(t) + \epsilon(t)(\vec{\xi}(t) - \vec{w_j}(t))$$

where $\vec{w_i}$ and $\vec{w_j}$ are the two closest codebook vectors and $\vec{\xi}$ belongs to C_j but not C_i and $\vec{\xi}$ falls in the "window" between $\vec{w_i}$ and $\vec{w_j}$. The resulting border movement increases the chances of an incorrectly classified pattern being correctly classified on a subsequent trial.

1.3 Summary

To summarize, the standard unsupervised algorithm classifies only based on input similarity. It is oblivious to the particular labels associated with each different task. For the three pattern distributions given at the top of Figure 2, each classification task has the same input patterns (just with different labels) and so it classifies them all the same way. This only results in a correct separation for the first problem. By adding supervision to the standard competitive learning algorithm we can provide it with task-specific information to allow it to classify the problems in Figure 2 differently. For appropriately separated classes, the simple algorithm of competitive learning in the augmented input space (with a dimension for each class) can provide appropriate boundaries. The algorithm is still limited, however, as it is not designed for optimal placement of the boundaries. For optimal classification with arbitrary class separation (and overlapping classes) the LVQ2.1 algorithm is best as it actually takes into account the current performance in order to improve it. These kinds of algorithms show that information about misclassifications can greatly increase the classification abilities of a piecewise-linear classifier. The power of the supervised approach lies in the fact that it directly minimizes its final error measure (on the training set). The positions of the codebook vectors are placed not to approximate the probability distributions but to decrease the number of misclassifications.

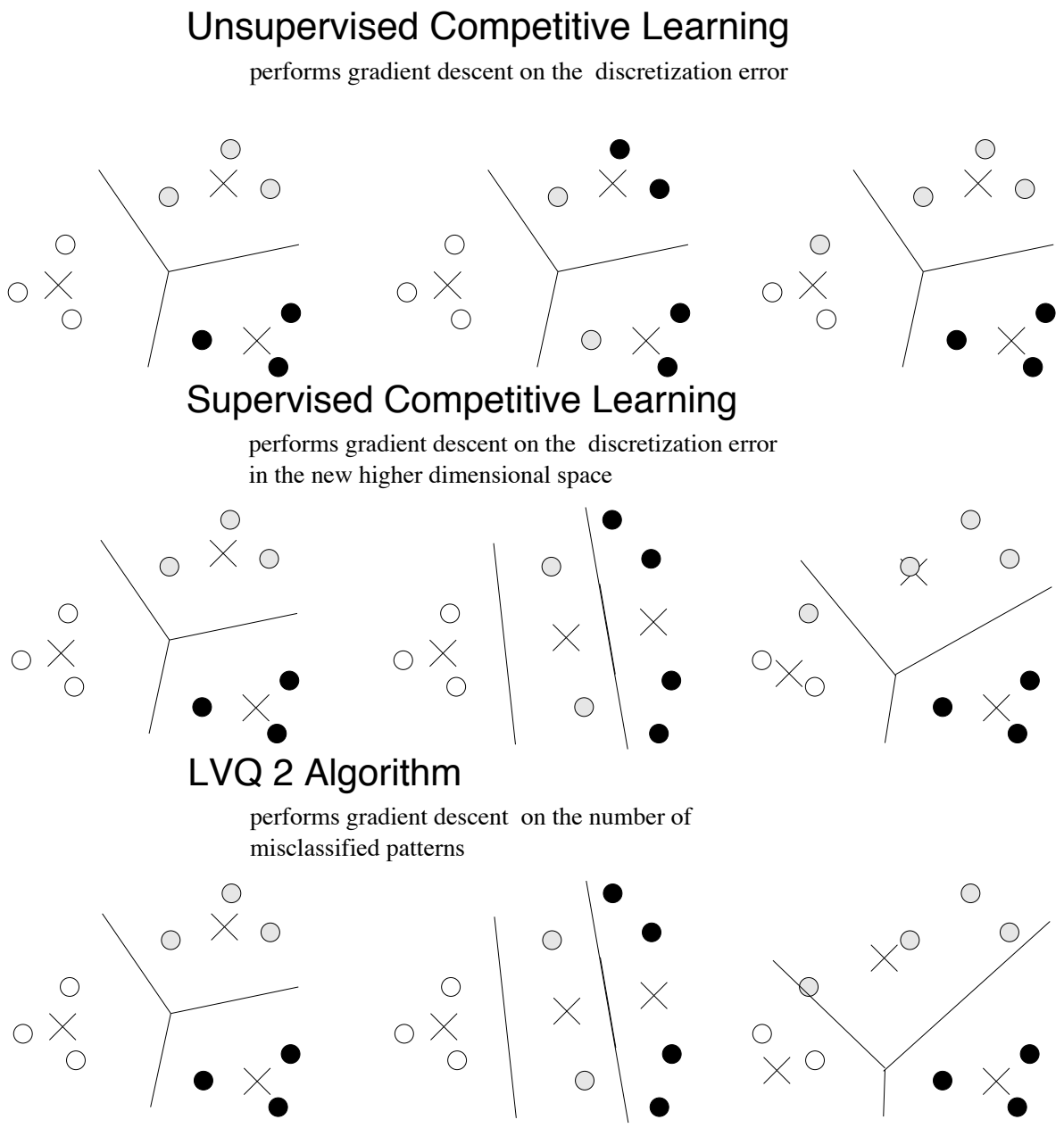

Figure 2: **Performance of the different algorithms on three different classification problems.** The three different colored circles (white, gray, and black) represent three different classes of patterns. Only the LVQ2 algorithm is able to separate the three classes appropriately. See the text for more description.

2 Cross-Modal Information

The power of supervised algorithms in handling difficult classification tasks makes them an attraction, but unfortunately, when modeling human category learning, one must deal with the fact that there is no omniscient homunculus (or pre-trained network) correctly labeling all the incoming sensory patterns. The algorithm employed by the brain does not have the correct answer always available, cannot measure its classification errors while learning, and hence can not directly minimize them.

One solution is to use the *cross-modal structure* between signals from two or more modalities to assist in the development of a piecewise-linear classifier within each modality. This takes advantage of the structure available in natural environments that results in sensations to different sensory modalities (and sub-modalities) that are correlated. For example, hearing "mooing" and seeing cows tend to occur together. So, although the sight of a cow does not come with an internal homuncular "cow" label it often co-occurs with an instance of a "moo."

These correlations have been noted before and are often used for justification in supervised cognitive models. Thus providing a 'cow' label with cow images (in a model learning to recognize animals) may be justified with the statement that "an infant is told 'cow' when shown a cow." The major point of disagreement of this model from the supervised cognitive models stems from the observation that the spoken word 'cow' is not a useful teaching signal until the auditory system has started to correctly parse and group speech signals. This is immediately apparent to those who have tried to build a machine speech recognition system or even observed sound spectrograms of spoken words. Similarly, as computer vision researchers are well aware, the cow picture is not a useful teaching signal for the 'cow' acoustic signal until the visual system has started to appropriately group visual images. Thus although the world provides extra information in cross-modal structure it is *not* in the form of the correct neuronal target output.

If it were possible to use a multimodal feature space that combined visual and auditory signals, the class distributions would be more distant and more readily separable. Thus if X is the feature space for vision and Y is the feature space for audition, Then $X \times Y$ should be more separable. This suggests that unsupervised clustering, or density estimation in this joint visual-auditory space would be helpful. However, simple clustering in the combined feature space, means that full-featured patterns would be needed to access the correct category (in other words, a full auditory-visual feature vector would always be needed). More suitable density estimation techniques that are capable of handling missing features rapidly become infeasible in high dimensions due to the large number of parameters. Thus the desire is to use the greater structure in the multimodal feature space while still processing in the lower dimensional spaces.

The key is to process the "moo" sound to obtain a self-supervised label for the network processing the visual image of the cow and vice-versa. This idea is schematized in Figure 3. In lieu of correct labels each network receives the output of the other network. Again, this is fundamentally different than running two separate supervised learning systems—the networks actually develop together. Initially they do not provide very good labels for each other but the algorithm allows the networks to develop into better classifiers together. Because the networks help each other we refer to this kind of algorithm as *self-supervised*.

This general strategy was introduced in the IMAX algorithm (Becker & Hinton,1992). The Minimizing-Disagreement (M-D) algorithm differs in that it was specifically designed to perform multi-class categorization with all communication occurring through natural and neurophysiologically plausible one-way connections. Phillips, Kay, and Smyth (1995) have recently developed a plausible implementation of IMAX for the binary classification case. Our algorithm uses reentrant feedback to allow multi-modality information to influence the development of uni-modal systems

as in (Reeke, Sporns, & Edelman, 1990; Edelman, Reeke, & Gall, 1992), however, we concentrate solely on classification, but attack real classification tasks with overlapping input patterns. The fact that both networks are learning makes this approach significantly harder than approaches where one modality trains another (Munro, 1988; Carpenter, Grossberg, & Reynolds, 1991; Tan 1995) or others that combine two already trained networks (Yuhas, Goldstein, & Sejnowski, 1988; Stork, Wolff, & Levine, 1992).

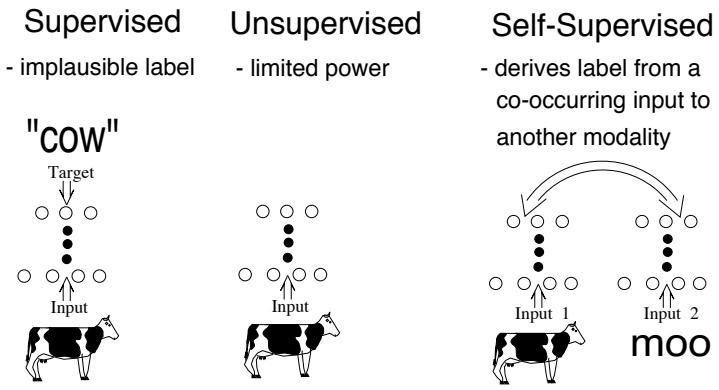

Figure 3: **The idea behind the self-supervised algorithm.**

To start we give some of the biological motivation behind the idea of using the cross-modality structure.

2.1 Psychophysics

The coincidence between auditory and visual events is pervasive throughout our experience. From just after birth, babies are able to turn crudely toward sounds. Over time we develop a rather precise auditory-visual spatial map. This enables us to be able to pick up correlations between auditory and visual sensations. For example, with sound localization we are able to notice that mooing comes from cows.

Another example of information picked up in this way is the ability to read lips. Anyone who has conversed in a noisy environment, is aware of the improved speech recognition achieved when the speaker's face (particularly the lips) is visible. (This has also been demonstrated in more controlled experimental conditions (Sumby, & Pollack, 1954).) The visual signal from the motion of the lips, jaw, and tongue help the auditory system to understand the speech.

This ability to recognize relationships between lip movements and emitted sound develops early. By four and a half months of age infants are able to recognize that particular lip motions go with particular sounds. Kuhl and Meltzoff (1984) showed that infants looked significantly longer at the matching face when presented with the sound /a/ or /i/. Their preference was specific to the actual speech information as they did not show this effect when the speech signals were replaced with tones that followed the duration, amplitude envelope and onset/offset timing of the original speech sounds (Kuhl & Meltzoff, 1984).

The effect of lip movement on speech recognition is even more prominent when the stimuli are experimentally manipulated so that the visual and acoustic signals are discordant. In the

experiments of (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978) subjects are presented with acoustic stimuli of various consonant-vowel pairs and simultaneously shown images of faces speaking a different consonant with the vowel. Thus for example when presented acoustically with a /ba/ syllable and visually with a face speaking /ga/, 98% of adult subjects hear /da/ (McGurk & MacDonald, 1976). The result is very striking and not subject to conscious control. It shows that visual and auditory stimuli are able to interact to produce a unified percept, different from the stimuli actually given to either modality. Furthermore it seems that the ability of visual signals to influence acoustic classification is at least partially learned. Pre-school and school children show significantly less of the effect than do adults (McGurk & MacDonald, 1976).

Another example of experienced auditory-visual correlations affecting an auditory perception is given as an anecdote in Howells (1944). Howells reports that a driver who was familiar with an intersection that made a concurrent whistle sound as the lights changed for stop and go, failed to notice when the whistle was later disabled.

> On the next trip to the crossing, in spite of counter suggestions recently received, the driver reported a distinct hallucination of the whistle (Howells, 1944, p.89).

In another anecdotal example where the frequencies of the whistles were different for lights changing to the stop or go signal. Howells reports

> Few drivers noticed that there was a difference in the tones, or even that a whistle sounded at all, until the wires controlling the whistles were accidentally crossed by a repairman, so the usual combinations of color and tone were reversed. The result was general confusion and a collision at the crossing (Howells, 1944 p. 90).

Under more controlled conditions, Howells (1944) trained subjects with tones followed by colored screens. On 95% of the trials, the color-tone pairing was consistent and on the other 5% of trials it was reversed. After an initial period, half the presentations were at full saturation and the other half much paler. Subjects showed increasing errors (on the pale screens preceded by the inconsistent tone) with training. Subjects tested with white stimuli after conditioning reported always the color associated with the co-occurring sound and subjects instructed to produce a white color in the presence of the high or low tone were offset in the direction opposite to that imposed during training.

A similar example of auditory events influencing visual perception is demonstrated in a cross-modal experiment in (Durgin, 1995; Durgin & Proffitt, 1996). The experiment involved repeated brief presentations of random dot patterns in two rectangular areas of a screen. On each presentation, one of the two areas received 25 dots/deg^2 and the other 2 dots/deg^2. The visual presentations were paired with auditory tone stimuli such that the pitch of the tone was perfectly correlated with the side of the denser dot pattern. After 180 flashed presentations, a staircase procedure was used to determine the perceived density equivalence (for test patterns with dot densities between the two trained densities) between the two areas when presented with each of the two tones. The experiment showed that there was a significant effect of the tone on the perceived density relationship between the patterns in the two areas. The simultaneous presentation of the tone associated with a denser texture in one area during training, lead to an impression of greater dot density in that area during testing. To match a constant density, the difference between the density required in the presence of the high pitch and that with a low pitch was 10% (Durgin, 1995).

Hefferline and Perera (1963) have shown that correlated proprioceptive (an invisibly small thumb twitch detected electromyographically) and auditory events (tone) can lead to a subject later reporting that he "still heard it (the tone)" to subsequent proprioceptive events in the absence of the tone.

Zellner and Kautz (1990) have also shown that color can affect odor perception. In their experiments, colored solutions were perceived as having a more intense odor.

> Even after being told the solutions were of equal concentrations, they [subjects] insisted that the solutions were not the same intensity (Zellner and Kautz, 1990, p. 396)

It is clear that the co-occurrence of multi-sensory signals can assist or interfere with processing. There is also evidence that after experience with cross-modal correlations, a uni-modal discrimination can be affected by a stimulus to the other modality. We would like to go one step farther and hypothesize that this multi-modality integration is important not only for improved recognition but is useful for the development of recognition features in both individual modalities. Along these lines, there is some evidence that exposure to auditory-visual co-occurrences is important for normal attentional development. Quittner and colleagues (Quittner, Smith, Osberger, & Mitchell, 1994) report that deaf children show reduced *visual* attention (in a non-auditory task) than hearing children. The authors conclude that though auditory information is not used in their tested task the development of focused visual attention is helped by auditory experience (presumably coincident with visual experience).

2.2 Neurobiology

The previous section examined results showing that information from different sensory modalities is combined in determining our perception. Often, the combination is not subject to conscious control. It is as if the results are not simply being combined at a high-level output stage but are able to influence each other in the individual processing stages. This is corroborated by neurophysiological studies which have found responses of cortical cells in primary sensory areas that respond to features from other sensory modalities. For example Spinelli, Starr and Barrett (1968) found sound frequency specificity in cells in primary *visual* cortex of the cat and Fishman and Michael (1973) found that these bimodal cells tend to be clustered together. As support for the unified percept observed in psychophysical studies, the stimuli are able to affect the same cell. In fact acoustic responses in a single cell could be inhibited by inhibitory visual stimuli (Morrell, 1972). More recently Maunsell and colleagues (Haenny, Maunsell, & Schiller, **?**; Maunsell, Sclar, Nealey, & DePriest,1991) have shown responses in visual neurons in Area V4 to oriented tactile stimuli that the animal has been trained to match to subsequently presented oriented visual gratings.

Sams and colleagues, (Sams, Aulanko, Hämääinen, Hari, Lounasmaa, Lu,& Simola, 1991) have also shown effects of visual input on auditory processing in humans. Using magnetoencephalographic (MEG) recordings, they showed that although a visual signal by itself did not result in a response over the auditory cortical area, different visual signals changed the response to the auditory signal. They again used the McGurk effect stimuli. Subjects were trained with a higher percentage of either agreeing or disagreeing stimuli. Significantly different neuromagnetic measurements were made to the frequent and infrequent stimuli. As no similar difference occurred when two different light stimuli occurred with the sounds, they argue that this shows that the visual information from the face is reaching the auditory cortex.

The neurophysiological and psychophysical evidence must be reconciled with the fact that anatomically the information from the different sensory modalities goes to spatially separate, segregated cortical areas. Retinal input goes to occipital cortex whereas auditory input goes to auditory cortex in the temporal lobe. Even within the auditory and visual cortex there are many different areas which seem to be specialized to processing different parts of the signal. For instance color processing seems to be mostly separate from motion processing (Merigan & Maunsell, 1993; Desimone & Ungerleider, 1989). There is a significant restriction on the amount of cross-modality

interaction that can occur. This is thought to be due to restrictions on connectivity; it is not physically possible to have all neurons connected to all other neurons (or even any significant fraction). Therefore input from each modality and submodality must be processed separately, at least in the early stages, with little cross-talk.

As there are no direct afferent (feed-forward) connections from one input modality to another, the information from other modalities could either be coming bottom-up from shared subcortical structures such as the superior colliculus or alternatively top-down from the multi-sensory integration areas such as entorhinal cortex and other limbic polymodal areas. This idea has been suggested before (for example Rolls, 1989) and seems to be supported by the evidence from visual cortex. As stated by Spinelli et al. (1968)

> non-visual stimuli affect the activity of ganglion cells only minutely (Spinelli, Pribram, & Weingarten, 1965; Spinelli & Weingarten, 1966; Spinelli, 1967; they affect that of the geniculate cells to a greater extent (Meulders, Colle, & Biosacq-Schepens, 1965) and very markedly affect cortical cells (Murata, Cramer, & Bach-y-Rita, 1965). Even more interaction appears to be present in prestriate cortex (Buser & Borenstein, 1959).(p. 82)

Thus we know from psychophysical studies that information from different modalities is combined and that information from one modality can assist or interfere with classification in another. The physiological evidence supports this finding in showing that input to other modalities can influence processing in another sensory pathway. This combined with the anatomical evidence that shows no direct input from one modality's transducers to another pathway, suggests that this information is coming top-down through feedback pathways from multi-sensory areas. Furthermore we suggest that this integration may not just affect the properties of developed systems but play an important role in the learning process itself. Just as lip-reading is a learned classification ability, correlations between inputs to different sensory modalities may affect other classification learning in the individual modalities. In this section we will investigate the power that this kind of integration might provide to learning classifiers in the individual modalities.

2.3 Using Cross-Modality Information for Self-Supervised Learning

Following the anatomical evidence presented earlier and acting under the assumption that it is infeasible to have neurons receiving input from the sensory transducers of all the senses, we propose an architecture such as that schematized in Figure 3 in which each modality has its own processing stream (or classification network) but access to each other's output at a high level. This information can reach the lower levels within each processing stream through feedback within the stream.

One way to make use of the cross-modality structure in a network like this is to cluster the codebook vectors in their individual spaces, but use the joint structure to learn to label co-occurring codebook vectors with the same label. After clustering in the input space, the activity patterns in the resulting hidden unit (codebook) activation space (the space of dimensionality equal to the sum of the number of codebook vectors in each space) can be clustered. For example using Competitive learning on the second layer of weights in a network such as that in Figure 4, will tend to cluster the codebook vectors. Each codebook vector is given the label of the output neuron in whose cluster it belongs (the neuron to which it projects most strongly); thus these weights can be considered implicit labeling weights.

The above use of cross-modality structure is useful and we will use it to initialize our algorithm (calling it the initial labeling algorithm), however a more powerful use of the extra information is for better placement of the codebook vectors themselves. The insight in this algorithm is that we

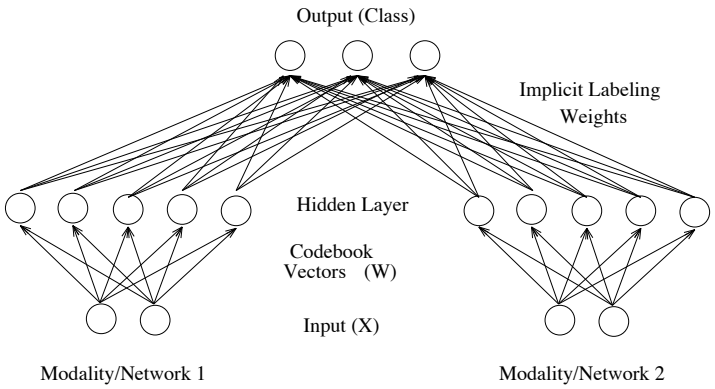

Figure 4: **Network for learning the labels of the codebook vectors.** The weight vectors of the hidden layer neurons represent the codebook vectors while the weight vectors of the connections from the hidden layer neurons to the output neurons represent the output class that each codebook vector currently represents. In this example there are 3 output classes and two modalities each of which has 2-D input patterns and 5 codebook vectors.

.

can make use of the joint cross-modal information without directly connecting all codebook vectors to each other and without requiring implausible communication.

The true classification goal of minimizing the number of misclassified patterns is explicitly supervised in that in order to monitor the number of misclassified patterns, one must be aware of the real class labels. For an unsupervised error function, we propose the *Disagreement Error*—the number of patterns classified differently by the two networks. The motivation behind this goal is that, in the absence of an external label, the output of the other network can take the role of the label. Each network will provide a "label" for the other network to aim for. The two modalities, representing different but co-occurring information from the same source, teach each other by finding a local minimum in their output disagreement.

We will explain the computational motivation for the learning rules with a simple 1-Dimensional example. Consider the 2-class example in Figure 5. Each class has a particular probability distribution for the sensation received by each modality. If modality 1 experiences a sensation from its pattern A distribution, modality 2 experiences a sensation from its own pattern A distribution. That is, the world presents patterns from the 2-D joint distribution shown in Figure 6 but each modality can only sample its 1-D marginal distribution (the sum of the curves in the subfigures in Figure 5, shown along the sides in Figure 6). If b_1 is the midpoint between codebook vectors of different classes in modality 1 and b_2 is the same for modality 2, we can write the disagreement error, $E(b_1, b_2)$, in terms of these class borders (or thresholds). If the codebook vectors are such that the codebook vector for Class A is to the left of that for Class B in both modalities, the disagreement error is written as:

$$E(b_1, b_2) = Pr\{x_1 < b_1 \ \& \ x_2 > b_1\} + Pr\{x_1 > b_1 \ \& \ x_2 < b_2\} \quad (1)$$

$$= \int_{-\infty}^{b_1} \int_{b_2}^{\infty} f(x_1, x_2) dx_1 dx_2 + \int_{b_1}^{\infty} \int_{-\infty}^{b_2} f(x_1, x_2) dx_1 dx_2 \quad (2)$$

where $f(x_1, x_2)$ is the joint probability density for the two modalities. We stress again that this energy function does not depend on the class information but does make use of the joint density

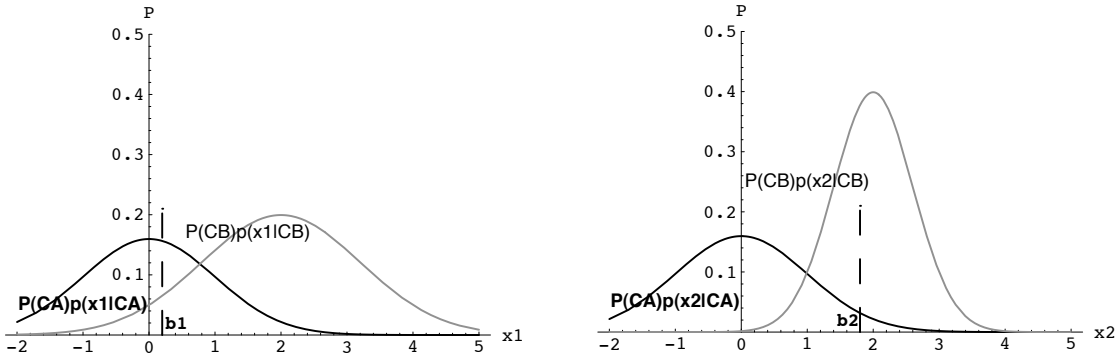

Figure 5: **An example world as sensed by two different modalities.** If modality 1 receives a pattern from its Class A distribution, modality 2 receives a pattern from its own class A distribution (and the same for Class B). Without receiving information about which class the patterns came from, they must try to determine appropriate placement of the boundaries b_1 and b_2. $P(C_i)$ is the prior probability of Class i and $p(x_j|C_i)$ is the conditional density of Class i for modality j

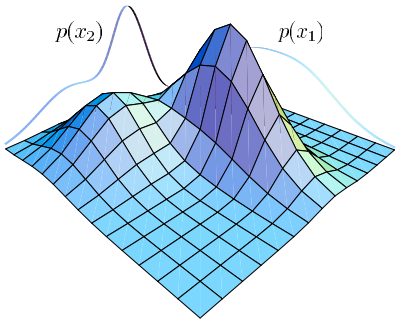

Figure 6: **The joint distribution for the Example in Figure 5.** The higher dimensional joint distribution has greater structure and is used to guide the placement of borders in the individual modalities. The figure shows an example where the two variables are conditionally independent but this is not required (although as mentioned later it helps significantly).

of all inputs and thus contains more information than that in the distributions of the individual modalities.

Applying stochastic gradient descent (Robbins & Monro, 1951; Wassel & Sklansky, 1972; Sklansky & Wassel, 1981; de Sa, 1994a; de Sa, 1994b; Diamantini & Spalvieri, 1995) to this error function gives a simple "on-line" rule for updating the position of both networks' classification borders

> If the pattern received by a modality is close to a current border, move the codebook vector of the class *that is output by the other modality* towards the pattern, and that of the other class away.

This rule moves the borders to increase the local proportion assigned to the class output by the other modality.

Similarly one can derive a generalization of this rule for multi-dimensional space that moves borders through moving codebook vectors. The Minimizing-Disagreement (M-D) algorithm applies this rule after each presentation of multimodal stimuli; it is summarized in Figure 7[2]. We will often refer to step 2 of the algorithm as the initial labeling stage and step 3 as the Minimizing-Disagreement stage. Weights are only updated if the current pattern falls near the middle between two codebook vectors of different classes (The specification of this "window" and the decrease in learning step size are as in Kohonen's 1990 supervised algorithm.)

To minimize disagreement with respect to the labels (last point of step 3 in Figure 7), the implicit labeling weights of the winning codebook vector unit in one modality are moved towards the label of the co-occurring codebook vector in the other modality. In order to avoid the undesirable solution where both networks always output the same class, we use weight normalization of the "implicit labeling weights" from the hidden units to the output units in Figure 4. In a biologically plausible implementation of the algorithm, the codebook vectors can be informed of the output class of the other network by using feedback weights from the output to codebook vector units. (These weights can be appropriately learned using Hebbian (Hebb, 1949) learning (de Sa & Ballard, in preparation)). The close codebook vectors with augmented activation are moved towards the input pattern and close ones that are not augmented with feedback activity are moved away. This type of learning rule is compatible with experimentally observed neural plasticity in visual cortical cells (Artola, Bröcher, & Singer, 1990).

Intuitively we can understand that this update rule minimizes the disagreement as follows. If Network 2 outputs Class A then Network 1 increases its future probability of saying Class A by moving its Class A codebook vector toward the pattern and the one from the other class away. If Network 2 outputs Class B then Network 1 increases its future probability of saying Class B similarly.

One might note that the global minima of the Minimizing Disagreement Energy are all codebook positions for which all patterns are classified as the same output. However, with small enough overlap in the joint space, a local minimum exists between the two class centers. An initial border determined by most simple clustering algorithms would start within the basin of attraction of this minimum. We can show (de Sa & Ballard, in preparation) that an appropriate local minimum exists beyond the case where clusters could be separated given the individual modalities alone, but just short of what could be achieved if one could look for clusters in the joint space. The algorithm is able to extract most of the greater structure in the higher dimensional joint distribution without requiring the extra parameters for modeling in this large space. We could add additional terms

[2]The rule shown was derived from generalizing the 1-Dimensional rule and is slightly simpler than the rule derived from differentiating in the multi-dimensional space (as in Diamantini & Spalvieri, 1995). As it is simpler and in our hands has performed as well as the multi-dimensionally derived rule, it was used in the simulations.

to ensure this (analogous to the individual entropy terms $H(Y_1), H(Y_2)$ in the IMAX (Becker & Hinton, 1992) algorithm). However for the datasets we have encountered this is not necessary and in fact with the vowel dataset in the next section, yielded slightly worse performance (probably because the addition of the terms changes the position of the local minimum—it is no longer minimizing the disagreement). However it is possible that for problems with more overlap between classes (where an appropriate local minimum might not exist), terms like this might help if the current algorithm does not perform well.

1. Initialize codebook vectors (randomly from data vectors or by unsupervised clustering)
2. Initialize labels of codebook vectors using the labeling algorithm
3. Repeat for each presentation of input patterns $X_1(n)$ and $X_2(n)$ to their respective modalities
   - Find the two nearest codebook vectors in each modality to their respective input patterns
   - Find the hypothesized output class in each modality (as given by the label of the closest codebook vector)
   - For each modality update the weights according to the following rules (Only the rules for modality 1 are given)

     Updates are performed only if the current pattern $X_1(n)$ falls within c(n) of the border between two codebook vectors of different classes (one of them agreeing with the output from the other modality). In this case

     $$\vec{w}_{1_{i^*}}(n) = \vec{w}_{1_{i^*}}(n-1) + \epsilon(n)\frac{(X_1(n) - \vec{w}_{1_{i^*}}(n-1))}{||X_1(n) - \vec{w}_{1_{i^*}}(n-1)||}$$

     $$\vec{w}_{1_{j^*}}(n) = \vec{w}_{1_{j^*}}(n-1) - \epsilon(n)\frac{(X_1(n) - \vec{w}_{1_{j^*}}(n-1))}{||X_1(n) - \vec{w}_{1_{j^*}}(n-1)||}$$

     where $\vec{w}_{1_{i^*}}$ is the codebook vector with the same label, and $\vec{w}_{1_{j^*}}$ is the codebook vector with another label.
   - Update the labeling weights

Figure 7: **The Minimizing-Disagreement (M-D) algorithm—a self-supervised piecewise-linear classifier.**

In summary, instead of minimizing the misclassifications (which cannot be done without a measure of the misclassifications provided by the labels), the Minimizing-Disagreement algorithm

minimizes the disagreement between the outputs of the two networks. This makes the algorithm fundamentally different from supervised algorithms. Before both networks have developed, they are providing bad label estimates — they must develop together. In the next section we review some experimental results that show that the networks are able to assist each other and significantly improve their classification abilities over their initial configuration. *We test the performance of each modality separately. The purpose of the algorithm is not to develop a method of combining sources of information for discrimination, but to develop a method of combining information for better development of the individual modules themselves.*

3 Artificial Modality Experiments

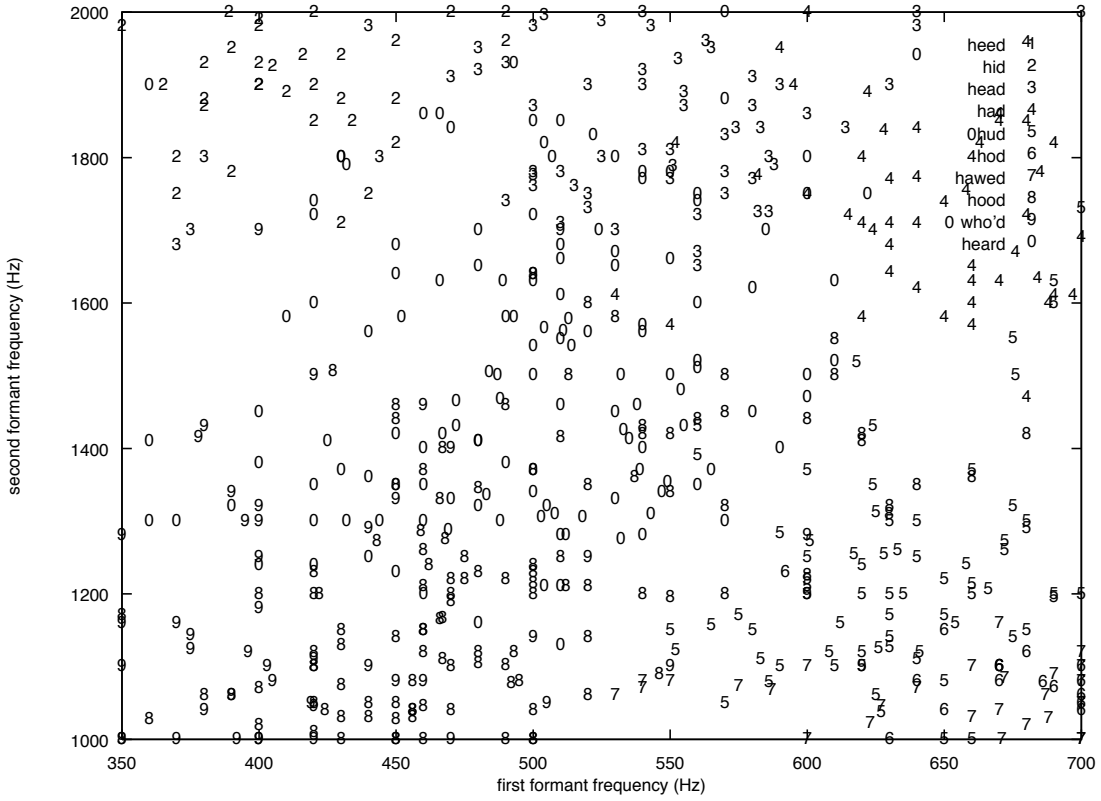

Figure 8: **A Close-up of part of the Peterson-Barney vowel dataset.** Note the large amount of overlap between the pattern classes in this central area.

In order to compare performance between the M-D algorithm and uni-modal supervised and unsupervised algorithms it is important that the difficulty in both modalities of the M-D problem be the same as that of the uni-modal problem. This is because the M-D algorithm's performance within a modality depends not only on the difficulty of the pattern classification within that modality but also on that of the other "labeling" modality; it is this modality that is providing a better or worse teaching signal [3] . For this reason, we ran a set of experiments that applied the M-D algorithm to two networks both simultaneously learning to classify the same dataset. Thus in this case *the two "modalities" are artificial and represent two independent sources drawing from the same input distributions.* Each source is the input to one modality. Accuracy was measured for each "modality" separately (on the training set). It is important to note that we are not combining the outputs of the two modalities as in methods that reduce variance by averaging over different networks (see for example Breiman 1996). The purpose of combining the two instances of the dataset is to learn appropriate classifiers for an individual dataset. The results described below are also tabulated in Table 1 and displayed in Figure 14.

The dataset chosen was the Peterson and Barney (1952) vowel formant dataset which consists of the formant frequencies[4] of 10 vowels spoken in a /hVd/ context (e.g. /had/, /hid/,...). We employed a version, used by and obtained from Steven Nowlan, consisting of the frequencies of

[3]This also addresses the criticism that it is just the "easier" modality supervising the other modality

[4]Formant frequencies are peak frequencies in the acoustic signal and reflect resonances of the vocal tract. These

the first and second formants of vowels spoken by 75 speakers (32 males, 28 females, 15 children) [5]. (See Figures 1 and 8.) We stress again, each modality receives an independently chosen two-dimensional vowel from the chosen class. The purpose in these experiments was not to study vowel recognition but to better understand and examine the M-D algorithm in an easily visualizable (two-dimensional) dataset.

In the first experiment, the classes were paired so that the modalities received patterns from the same vowel class. If modality 1 received a pattern from the /a/ vowel class, modality 2 received an independently chosen pattern from the same (/a/) vowel class and likewise for all the vowel classes (i.e. $p(x_1|C_j) = p(x_2|C_j)$ for all j).

3.1 Better Border Placement

After the initial labeling algorithm stage (the second step in Figure 7), the accuracy was $60 \pm 5\%$ reflecting the fact that the initial random placement of the codebook vectors does not induce a good classifier. After application of the Minimizing-Disagreement stage (the third step in Figure 7) the accuracy was $75 \pm 4\%$. At this point the codebook vectors are much better suited to defining appropriate classification boundaries.

The improvement in border placement can be seen by observing the arrangement of misclassified patterns after each stage. Figures 9 and 10 show the misclassified patterns after the initial labeling and Minimizing-Disagreement stage of the algorithm. These results are averaged over 20 runs in one modality. The results for the other modality are similar. The size of the diamonds corresponds to the fraction of classifiers from the 20 runs that misclassified the pattern. For both stages of the algorithm there are several patterns that are surrounded by many patterns of the same class and are always classified correctly as well as many patterns that are surrounded by patterns from other classes that are always misclassified. However, after the Minimizing-Disagreement stage many of the patterns between the two extremes, especially those at the edges of the class distributions, are classified more reliably (correctly by more of the classifiers). This shows as a decrease in diamond size from Figure 9 to Figure 10.

This difference between the two figures is made explicit in Figure 11 where the size of the diamonds corresponds to the difference in percentage of classifiers that correctly classified the pattern. Filled diamonds represent positive differences or better classification after the complete Minimizing-Disagreement algorithm and open diamonds better classification after only the initial labeling. The improvements from the M-D stage are in the border regions between classes indicating that minimizing the disagreement tends to find better border placements.

3.2 Cycling Helps

The final performance figures were positively correlated with the performance after the label initialization step, which in turn was positively correlated with (and bounded by) the best performance possible with the randomly chosen initial codebook vectors (as measured independently with optimal labels). This suggested that improved methods of choosing the initial codebook vector positions and/or labels might result in improved final performance. Thus, we tried using the final codebook vectors from a run of the Minimizing-Disagreement algorithm as the initial codebook vectors for another run (replacing the first step in Figure 7). This resulted in improved performance ($73 \pm 4\%$

resonant frequencies are characteristic of different configurations of the vocal tract and are useful features for vowel recognition.

[5] Each speaker repeated each vowel twice except 3 speakers that were each missing one vowel. The raw data were linearly transformed to have zero mean and fall within the range $[-3, 3]$ in both components.

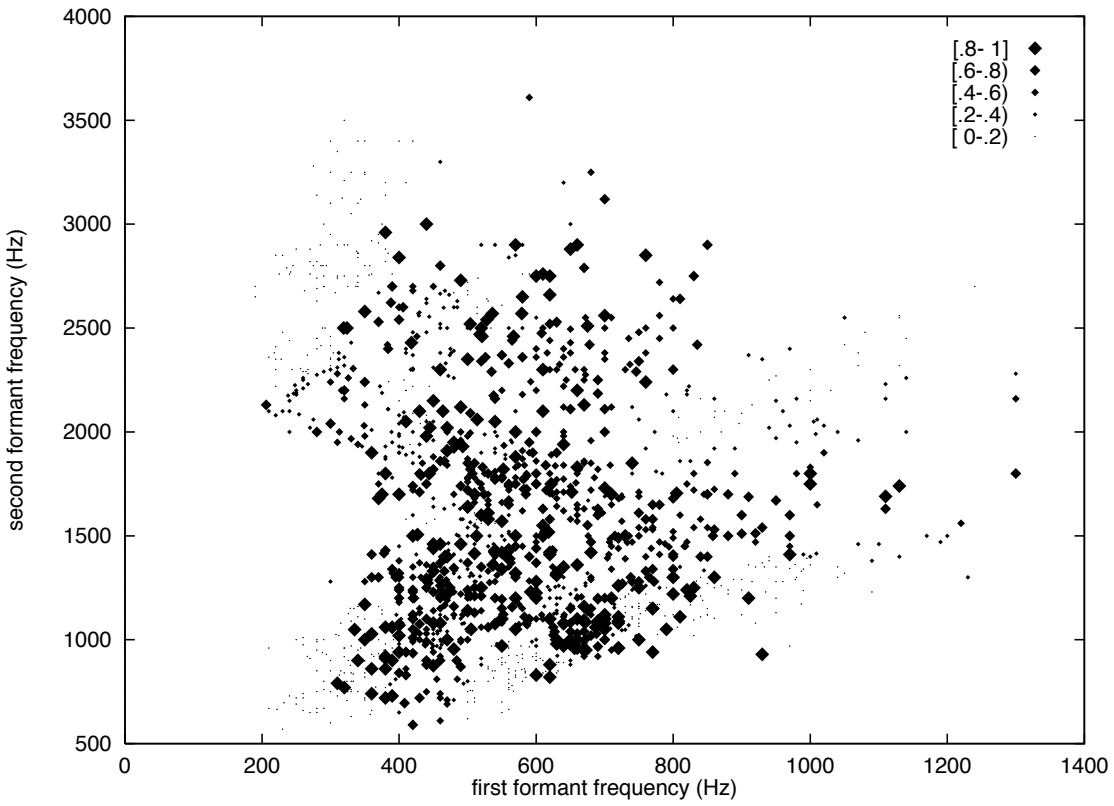

Figure 9: **Performance of initial labeling algorithm on the Peterson-Barney dataset.** The size of the diamonds corresponds to the fraction of classifiers that misclassified the pattern.

after step 2 and $76 \pm 4\%$ after step 3). Figure 12 shows performance after the initial labeling, first application of the M-D algorithm, and second application of the M-D algorithm for 30 different trials. This cycling of steps 2 and 3 was repeated several more times with no further significant increase in performance (see Table 1).

The improvement can be seen in the difference plot in Figure 13. Again the size of the diamonds corresponds to the difference in performance. The main area of improvement is in the crowded and very overlapped middle area.

3.3 Improvement when the Confusable Classes are Different in the Two Modalities

One feature of cross-modality information is that classes that are easily confusable in one modality may be well separated in another. This is evident in lip-reading. For example, consonants /b/ and /p/ which differ only in voicing (the presence of vocal cord vibration) are easily acoustically distinguished even in the presence of noise (Miller & Nicely, 1955). Visually however they are indistinguishable. On the other hand /b/ and /d/, which differ in their place of articulation, are visually distinct but more acoustically confusable in the presence of noise. This difference in confusable classes between the modalities should lead to improved performance with the Minimizing-Disagreement algorithm. As the "labeling" signal for separating the overlapping classes is likely determined between two non-overlapping distributions in the other modality, it will tend to be more reliable.

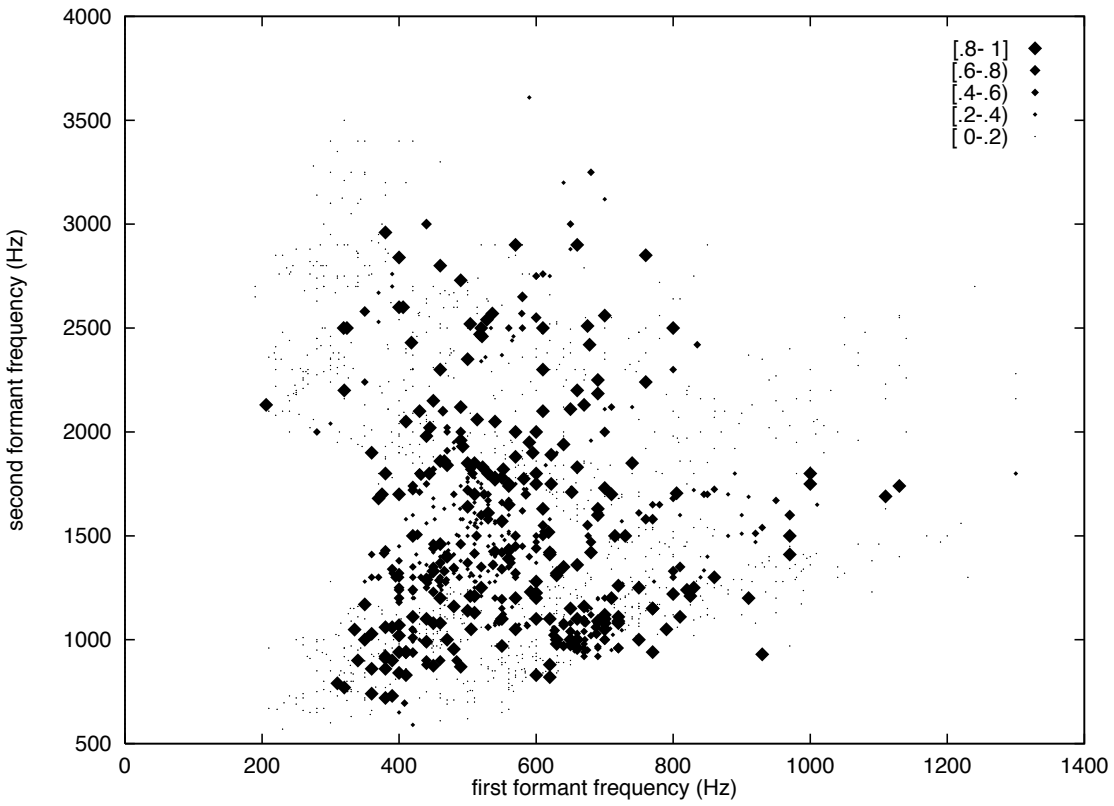

Figure 10: **Performance after Minimizing-Disagreement algorithm on the Peterson-Barney dataset.** The size of the diamonds corresponds to the fraction of classifiers that misclassified the pattern.

To explore this, more tests were conducted with random pairing of the vowels between the modalities for each run. Vowel patterns from one class to one network were paired with vowel patterns from another class to the other network. This pairing of classes was chosen randomly before the experiments but kept consistent for each experiment. For example, presentation of patterns of /a/ vowels to one modality would be paired with presentation of patterns of /i/ vowels to the other. That is $p(x_1|C_j) = p(x_2|C_{\alpha_j})$ for a random permutation $\alpha_1, \alpha_2..\alpha_{10}$. For the labeling stage the performance was as before (60±4%) as the difficulty within each modality has not changed. However after the Minimizing-Disagreement algorithm the results were better as expected. After 1 and 2 iterations of the algorithm, $77 \pm 3\%$ and $79 \pm 2\%$ were classified correctly.

3.4 Comparison to Other Algorithms

The algorithm performs better than a hybrid unsupervised-supervised algorithm — Kohonen feature mapping algorithm (with 30 codebook vectors) followed by optimal labeling of the codebook vectors, which achieved 72%. (In fact if the same optimal labeling algorithm is applied to the codebook vectors resulting from the M-D algorithm, an average performance of 76% and 78% (for applying after one or two iterations respectively) results.) Performance is not as good as that of the related fully supervised algorithm LVQ 2.1, which achieved an average performance of 80%, but is comparable to the performance of (supervised) back-propagation (with 25-200 hidden layer neurons) which obtained average performances of 73.4-78.5% (Nowlan 1991). Nowlan's (1991) more

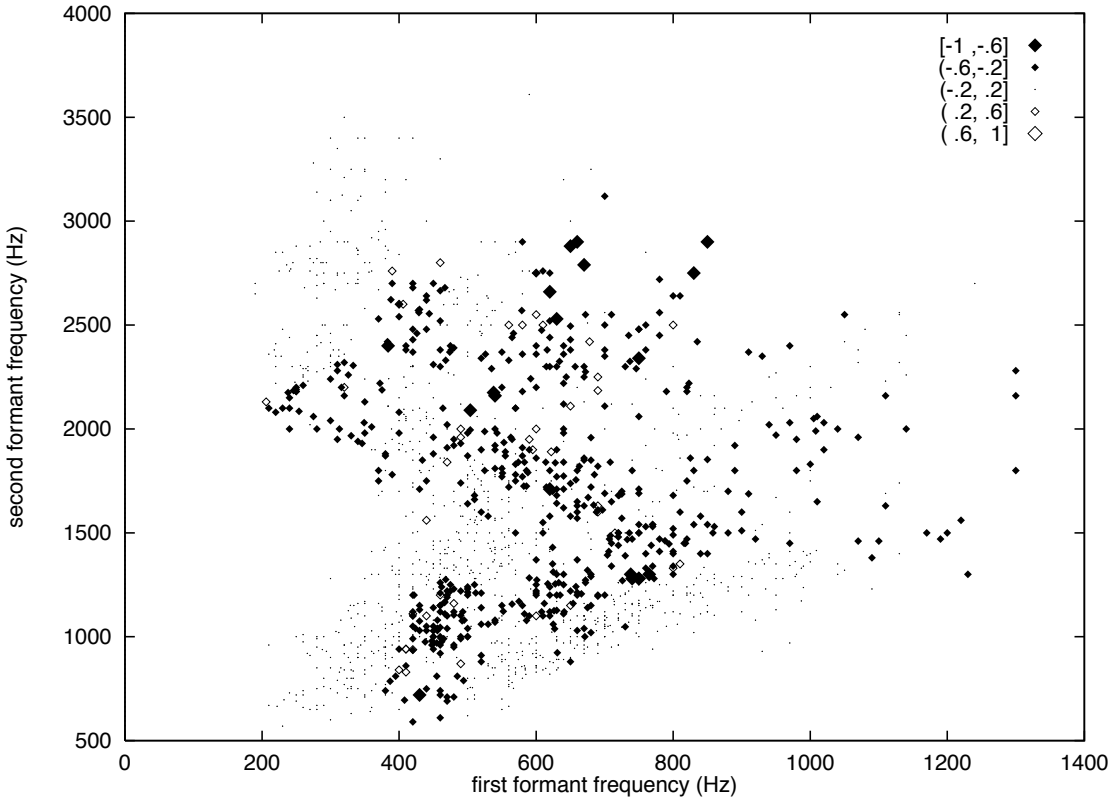

Figure 11: **Improvement from running the M-D stage.** The size of the diamonds corresponds to the difference in percentage of classifiers that correctly classified the pattern. Filled diamonds represent better classification after the Minimizing-Disagreement algorithm and open diamonds better classification after only the initial labeling.

complicated mixture model (supervised) achieved an average performance of 86.1%. These results are all shown in Figure 14.

Figure 15 shows a comparison between the results from the 2-stage Minimizing-Disagreement algorithm (with same-vowel pairs) and the results from using the supervised LVQ2.1 algorithm. Again the size of the diamonds reflects the magnitude of the difference in performance and the color reflects the sign. The larger number of open diamonds reflects the slight superiority of the supervised algorithm. More strikingly though, this plot emphasizes the borders between the regions. The two algorithms differ so slightly in the resulting classifiers that the only patterns classified differently are right at the borders. The two algorithms tend to pick slightly different borders resulting in some border patterns being better classified by one algorithm and others by the other.

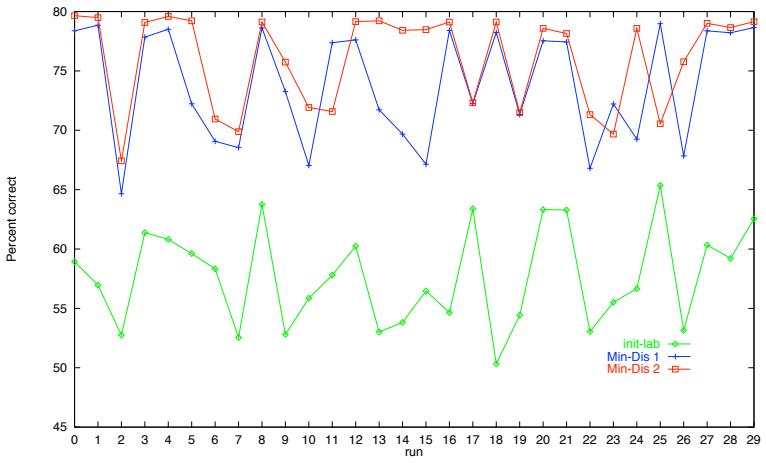

Figure 12: **The performance (for 30 different initial configurations).** Performance was measured after: initial labeling (init-lab), one application of the M-D algorithm (Min-Dis 1), and two applications of the M-D algorithm (Min-Dis 2)

Table 1: **Tabulation of performance figures (mean percent correct and sample standard deviation over 30 trials and 2 modalities).** The heading $i-j$ refers to performance measured after the j^{th} step during the i^{th} iteration. (Note Step 1 is not repeated during the multi-iteration runs).

	1-2(IL)	1-3(M-D)	2-2	2-3(M-D2)	3-3	4-3	5-3
same-paired vowels	60 ± 5	$\mathbf{75 \pm 4}$	73 ± 4	$\mathbf{76 \pm 4}$	76 ± 4	76 ± 4	76 ± 4
random pairing	60 ± 4	$\mathbf{77 \pm 3}$	77 ± 3	$\mathbf{79 \pm 2}$	79 ± 2	79 ± 2	79 ± 2

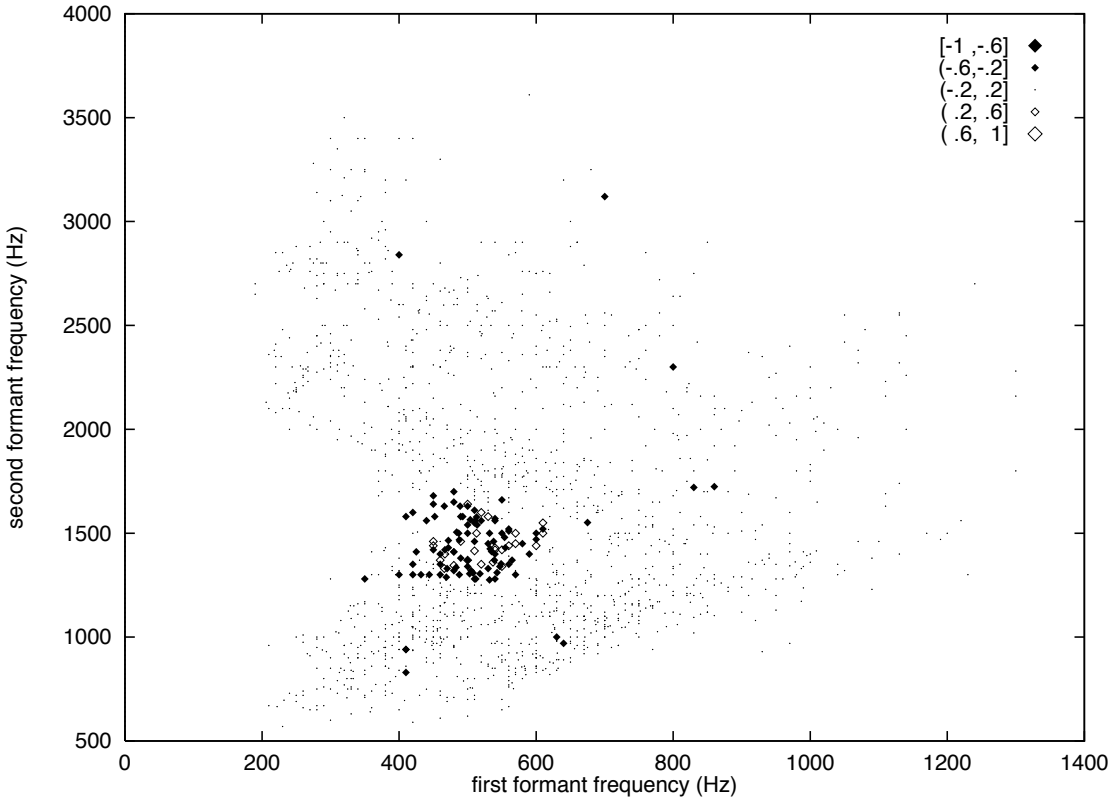

Figure 13: **Difference in Performance after 1 or 2 iterations of the M-D algorithm.** The size of the diamonds corresponds to the difference in percentage of classifiers that correctly classified the pattern. Filled diamonds represent better classification after the two iteration algorithm and open diamonds better classification after the first iteration.

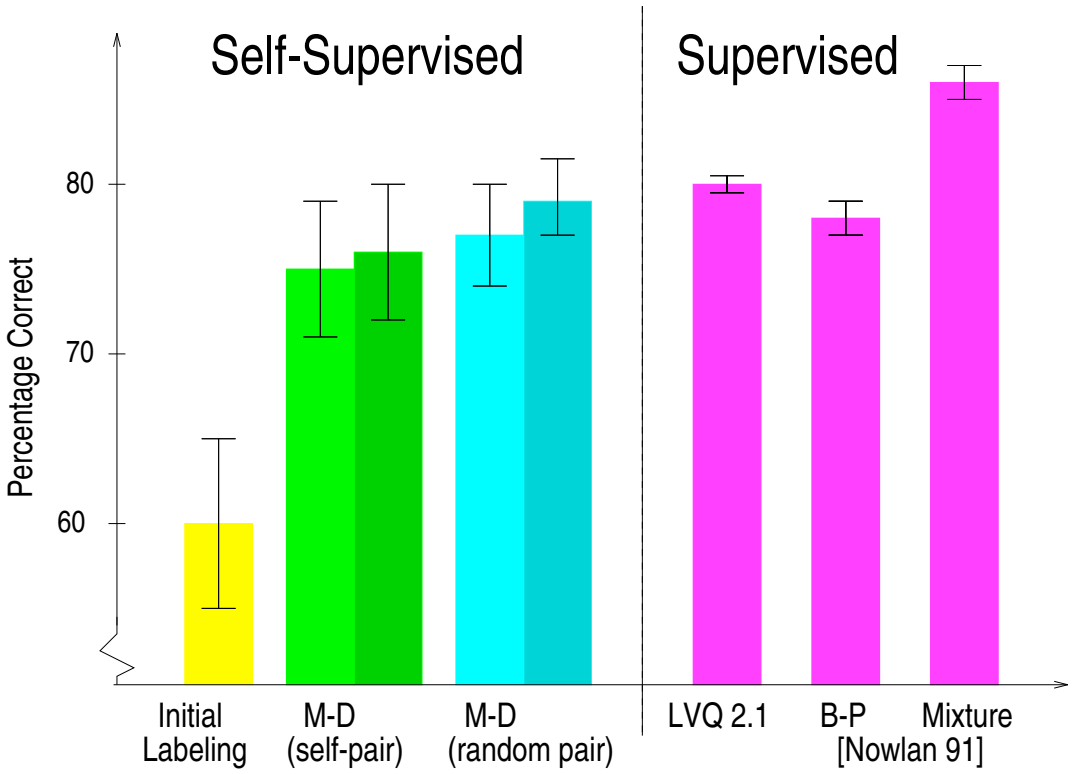

Figure 14: **Results from different algorithms on the Peterson-Barney Vowel Dataset.** The double bars for the M-D algorithms represent results from 1 iteration and 2 iterations. The error bars represent 1 standard deviation.

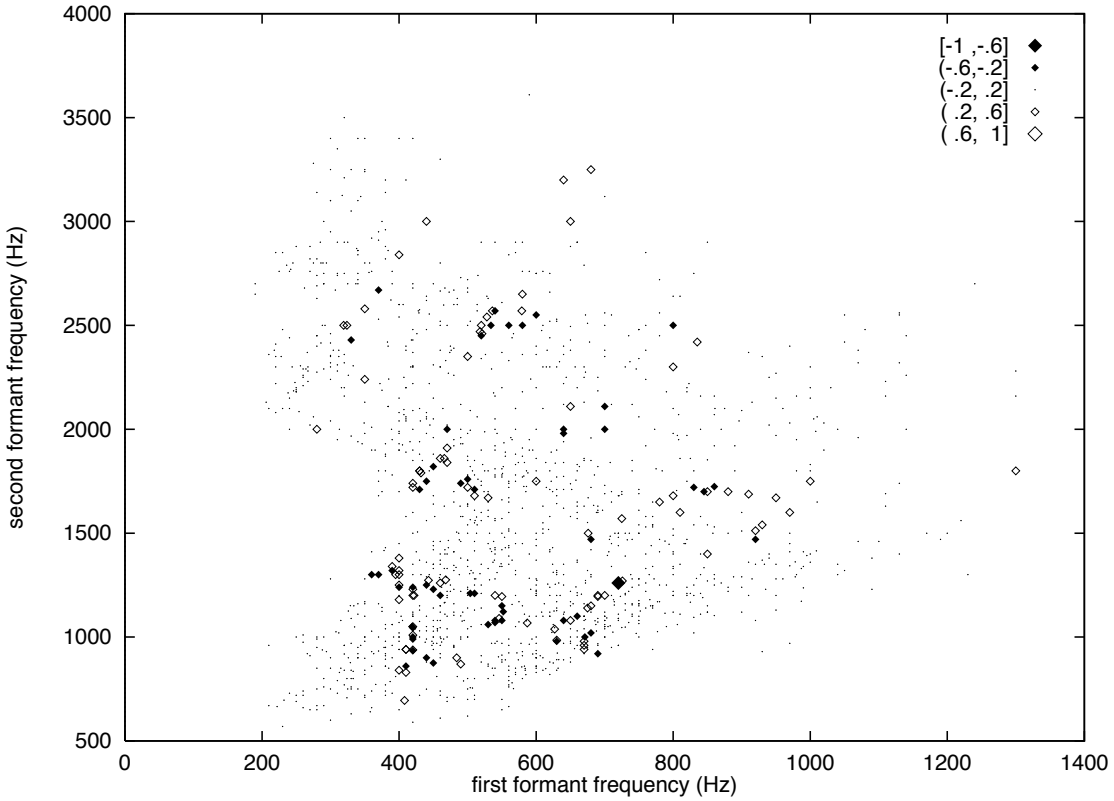

Figure 15: **Difference in Performance between the supervised and M-D algorithm.** The size of the diamonds corresponds to the difference in percentage of classifiers that correctly classified the pattern. Filled diamonds represent better classification with the Minimizing-Disagreement algorithm (with two iterations) and open diamonds better classification with the supervised algorithm.

4 Cross-Modal Experiments

Data Collection

Data were collected using an 8mm camcorder from 5 male English speakers as they spoke 26 iterations of /ba/ /va/ /da/ /ga/ /wa/ [6]. Each set of 10 utterances (twice through the set) was preceded by a clap using a clapboard arrangement similar to that used in commercial movie production for matching the visual and auditory signals. The camera recorded 30 frames a second and was roughly positioned to view the tip of the nose through chin of the speaker. The audio was recorded through a directional microphone positioned approximately 12cm from the speaker's mouth.

The acoustic data were transferred to a Sparc LX and low-pass filtered. Utterances were detected using threshold crossings of the smoothed time-domain waveform (using the ESPS software from Entropic Research Laboratory, Inc.). As some of the consonantal information is low amplitude (before the threshold crossing), each utterance was taken from 50msec before the automatically detected utterance start to 50msec after. These utterances were then encoded using a 24 channel mel code [7] over 20msec windows overlapped by 10msec. This is a coarse short time frequency encoding, which crudely approximates peripheral auditory processing. The final auditory code is a (24*9 = 216) dimension vector for each utterance.

The visual data were processed using software designed and written by Ramprasad Polana (1994). Visual frames were digitized as 64×64 8 bit gray-level images using the Datacube MaxVideo system. The video and auditory tracks were aligned using the clapboard arrangement and visual detection of the clap was performed manually which allowed alignment to within 1 video frame (1/30 second). (For an example of a video sequence showing a clap frame see Figure 16). The frame of the clap was matched to the time of the acoustically detected clap allowing the automatic segmentation obtained from the acoustic signal to be used to segment the video. Segments were taken as 6 frames before the acoustically determined utterance onset and 4 after. The normal flow was computed using differential techniques between successive frames. Each pair of frames was then averaged resulting in 5 frames of motion over the 64×64 pixel grid. The frames were then divided into 25 equal areas (5×5) and the motion magnitudes within each frame were averaged within each area. This gave a final visual feature vector of dimension (5 frames * 25 areas)= 125.

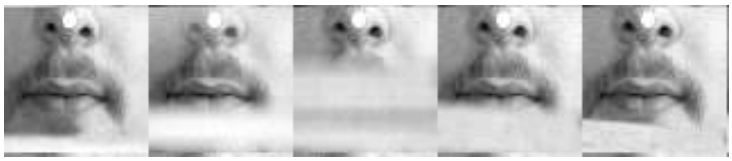

Figure 16: **Example frames showing the clap detection.** The clap is in the center frame.

Figure 17: **Example /ba/ utterance.**

[6] A few speakers spoke more iterations.

[7] A frequency transform in which the frequency bands have linear spacing below 1000Hz and logarithmic above 1000Hz.

Figure 18: **Example /va/ utterance.**

Figure 19: **Example /da/ utterance.**

4.1 Results

The data for one speaker were unusable due to a problem with the video tape. The training set was made up of each of the other speakers' first 20 cycles through the utterances (minus a few cycles that could not be used due to lost frames during digitization). The test set was made up of the next 6 cycles [8]. For all experiments 30 codebook vectors were used in the auditory pattern space and 60 in the visual pattern space. As the auditory signal contains more reliable information, the relative contributions from the auditory and visual networks during the labeling algorithm were weighted (1.5 to 1) in favor of the auditory choice.

We first benchmarked the dataset by running the supervised LVQ2.1 algorithm. Using 30 codebook vectors for the auditory patterns we achieved an accuracy of 99% on the training set and 97% on the test set. Using 60 codebook vectors for the visual patterns the performance was 83% on the training set and 60% on the test set. We also ran the unsupervised-supervised algorithm of Kohonen feature mapping followed by optimal labeling of the codebook-vectors. This gave accuracies of 84% and 55% on the two training sets.

Ideally we would like to test the M-D algorithm by presenting to the auditory and visual networks the pairs of patterns that occurred together. However, to get a good covering of the spaces, many utterances need to be collected. Due to the time involved in the current method of synchronizing the audio and video (they are processed separately and synchronized manually through the visual clap detection) it was decided to run preliminary experiments that artificially expand the dataset using the technique employed with the vowel dataset. This technique makes the assumption that within an utterance class the exact auditory and visual patterns are independent and thus each auditory pattern can be paired with each visual pattern from the same class (not just the one with which it actually co-occurred). For example, an individual acoustic pattern from a /ba/ utterance is randomly paired with a visual sample from a randomly chosen /ba/ utterance.

For these experiments, the Minimizing-Disagreement algorithm was applied to codebook vectors resulting from the unsupervised Kohonen learning algorithm instead of randomly initialized ones

[8] For some speakers there were a few extra cycles that were also included.

Figure 20: **Example /ga/ utterance.**

Figure 21: **Example /wa/ utterance.**

in the respective spaces [9]. The initial labeling algorithm on the codebook vectors resulting from the Kohonen learning algorithm resulted in 72% on the training set and 68% on the test set for the auditory network and 48% (training) and 36% (test) for the visual network. The Minimizing-Disagreement stage was able to greatly increase the classification performance from this initial state to 97% and 92% for the auditory network and 82% and 58% for the visual network. The performance results are summarized in Figure 22.

While the previous results were encouraging, it was important to demonstrate the algorithm in the fully unsupervised way, making no assumptions about independence between the modalities, and using only the cross-modality information sampled from the environment. In order to accurately sample the space, we restricted the problem to that of a single speaker. This speaker repeated 120 cycles of /ba/, /va/,/da/,/ga/,/wa/. The first 100 cycles (minus two that lost frames during the digitization) were used as the training set and the last 20 were used as the test set. Again for this cross-modal dataset the M-D algorithm was applied using the results of the Kohonen learning algorithm as the initial codebook vector positions. The algorithm achieved accuracies of 92% (train), 92% (test) on the auditory data and 91% (train),78% (test) on the visual data. For comparison the supervised LVQ2.1 algorithm, as well as the M-D algorithm using full-pairings as before, were also run on this dataset. The supervised results were 99% (aud-train) ,95% (aud-test) and 96% (vis-train), 82% (vis-test). The M-D algorithm using the artificial increased pairing resulted in 98% (aud-train), 95% (aud-test) and 95% (vis-train), 80% (vis-test). These results are displayed in Figure 23.

The results demonstrate that for one speaker, the natural lip-sound co-occurrences were enough to give performance within 7% percent of the supervised results on the training set and within 4% of the supervised results on the test set. More data collection is needed to determine if the fully unsupervised algorithm would work on the multi-speaker problem.

[9]Initial experiments suggested that this might provide better results but later tests indicated there was not much difference.

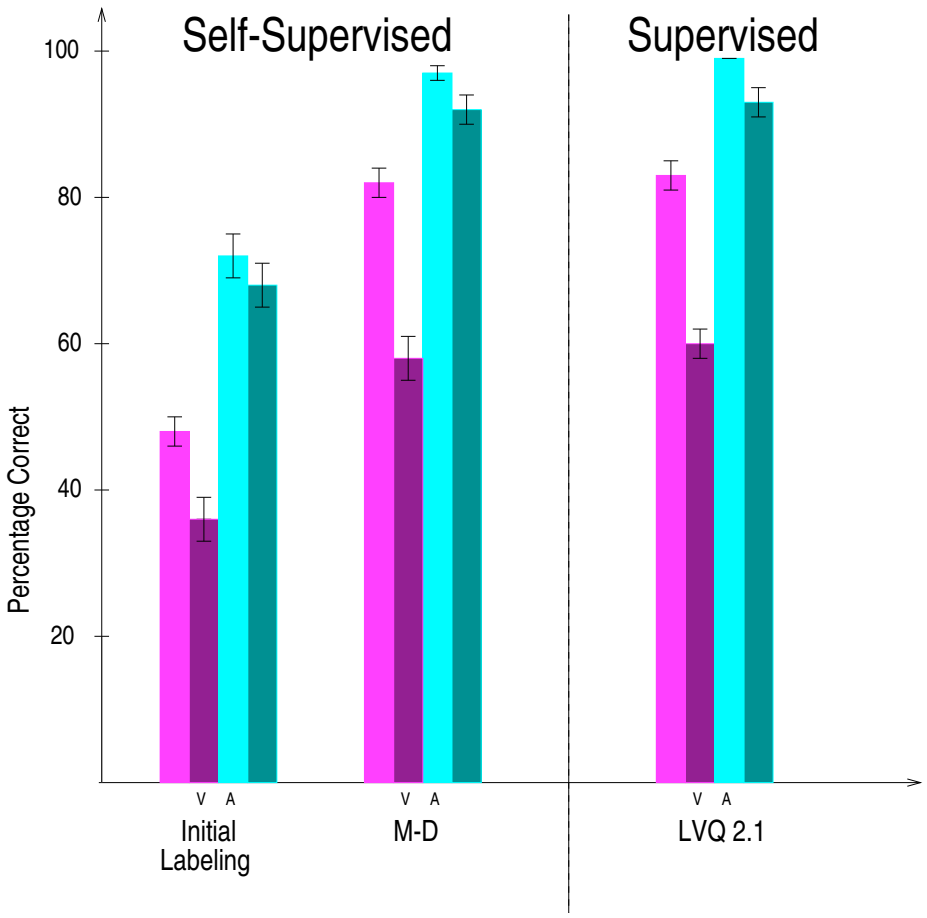

Figure 22: **Results on the preliminary cross-modal dataset.** The two leftmost bars in each set of four give the performance of the visual network and the rightmost bars show the auditory network's performance. Within the two bars for each modality, the lighter and leftmost bar represents performance on the training set. The darker, rightmost bars give results on the test set. The error bars represent 1 standard deviation.

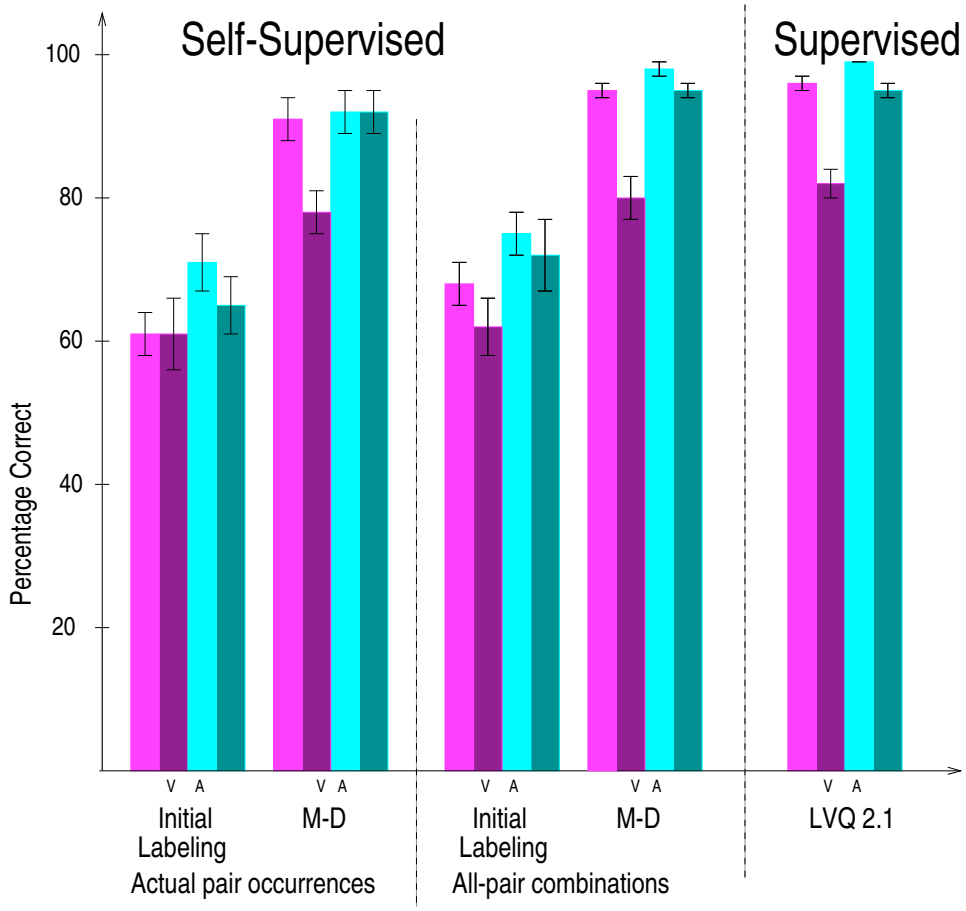

Figure 23: **Results on the single-speaker cross-modal dataset.** The two leftmost bars in each set of four give the performance of the visual network and the rightmost bars show the auditory network's performance. Within the two bars for each modality, the lighter and leftmost bar represents performance on the training set. The darker, rightmost bars give results on the test set. The error bars represent 1 standard deviation.

5 What makes a modality?

Often, as in the previous example, input data for a classification task arise from separate sources (e.g. visual and auditory sources for speech). Although these inputs from multiple sources can be considered as one long input vector for training, we have seen that having separate modalities for processing different sets of input dimensions allows the subnetworks to teach each other and avoids density modeling in high dimensional spaces. In this section we investigate the importance of, and requirements for properly choosing the input subsets (or "modalities").

Consider splitting the auditory and visual data to produce two "submodalities" consisting of part visual or part auditory data. The data are divided as shown in Figure 24. These new sets of dimensions (or submodalities) A1,A2,V1 and V2 can be combined in various ways to make up "pseudo-dimensions". In this section, all performance figures are calculated by dividing the data 10 times into a training, validation and test set [10].

In order to not bias performance towards one division due to the learning parameters, a variety of reasonable parameter ranges were tried. For the supervised algorithms, the best parameters were discovered using the validation data and then trained with these parameters on the combined training and validation sets. For the M-D experiments the networks were trained on the combined training and validation sets; classification performance on this set was averaged across all 10 divisions to find the best parameters[11]. The performance with the best parameters is reported on the test sets. As with the earlier cross-modal experiments, the same parameters were used for both modalities. It is possible that all the networks would benefit from different parameters on each side, however we don't believe that this would change the general trend of our results.

First consider the submodality A1, being trained by A2 or V1 or V2. Which of the other submodalities when trained using the M-D algorithm with A1 will lead to the best subsequent A1 classifier? One could try to discover how much potential is in the various subsets by comparing the performance of supervised algorithms on each one. These results for A1, A2, V1 and V2 are reported in Table 2.

Given that A2 seems to have more potential than either set of the visual dimensions, one might think that A2 would be the best "co-modality" for training A1. The actual results however are surprising. The performance of A1 when trained with A2, V1, and V2 are shown in Table 3. V1 and A2 are equally good co-modalities for A1. The results for training A2 are even more surprising. V1 is a better co-modality than A1, and V2 and A1 are equivalent.

This relative improvement of the visual submodalities as co-modality for the auditory submodalities stems from the benefit of having uncorrelated dimensions in the other modality. If the relationships between the various dimensions does not differ much between classes (i.e. if the conditional correlational structure is the same as the correlational structure across the whole dataset) then the clusters in the joint space are not as obvious. A simple demonstration of this is shown in Figure 26. It is shown that as the conditional dependence between the two dimensions is increased, the bimodal shape in the joint distribution is less obvious. This means that the joint structure is less able to help the individual modalities to find the appropriate borders. For the M-D algorithm, the best co-modalities are those that have a lot of separation between the different classes

[10] Due to the performance averaging across the 10 divisions, the test set for some divisions appears in the training and validation sets for others. This might tend to inflate the performance results but we don't believe it would bias them towards any particular division.

[11] We realize that using the classification metric to pick the learning parameters is not the best for evaluating unsupervised learning algorithms (though one can make evolutionary arguments for this), but it allowed us to quickly compare the relative capabilities of the different divisions. Also as our intent was to compare the relative benefits of divisions, we did not use the improved double iteration technique.

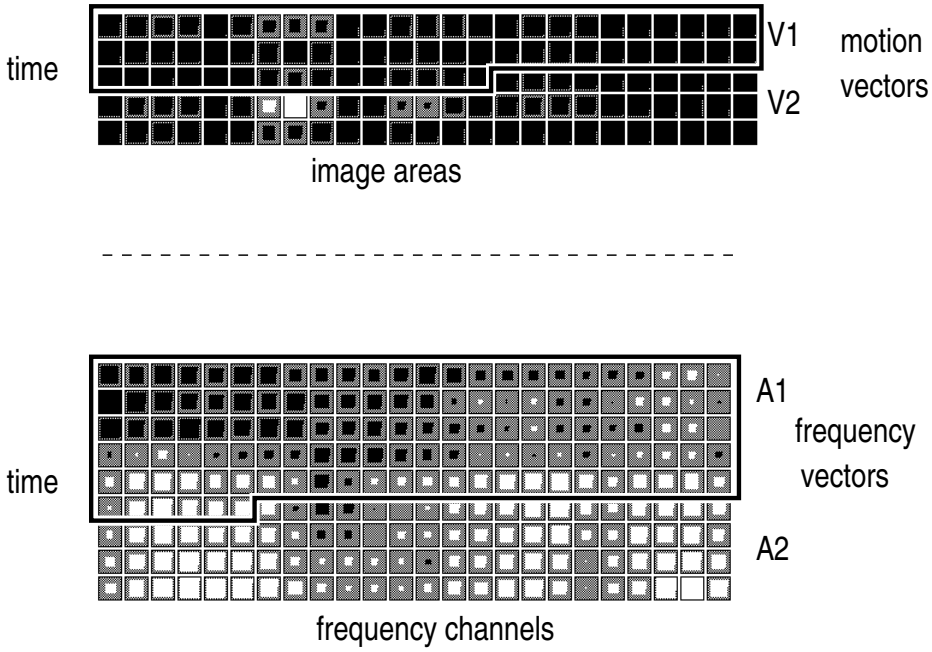

Figure 24: **Dividing up the Auditory and Visual Data.** For the mixed-modality experiments the auditory and visual data are divided into two parts. V1 refers to the first 65 dimensions of the visual feature vector and A1 refers to the first 126 dimensions of the auditory feature vector as shown.

Table 2: **Table of supervised performance for the submodalities.** The numbers give the percentage correct performance on the test sets (and standard deviation across the 10 dataset divisions).

Pseudo-Modality	Supervised Performance
A1	89 ± 2
A2	91 ± 2
V1	83 ± 2
V2	77 ± 3

Table 3: **Table of performance figures for A1 and A2 when trained by the other submodalities** The numbers give the percentage correct performance on the test sets (and standard deviation across the 10 dataset divisions).

Trained By	A1	A2	V1	V2
Performance of A1	N/A	69 ± 5	69 ± 3	63 ± 3
Performance of A2	74 ± 5	N/A	80 ± 4	74 ± 5

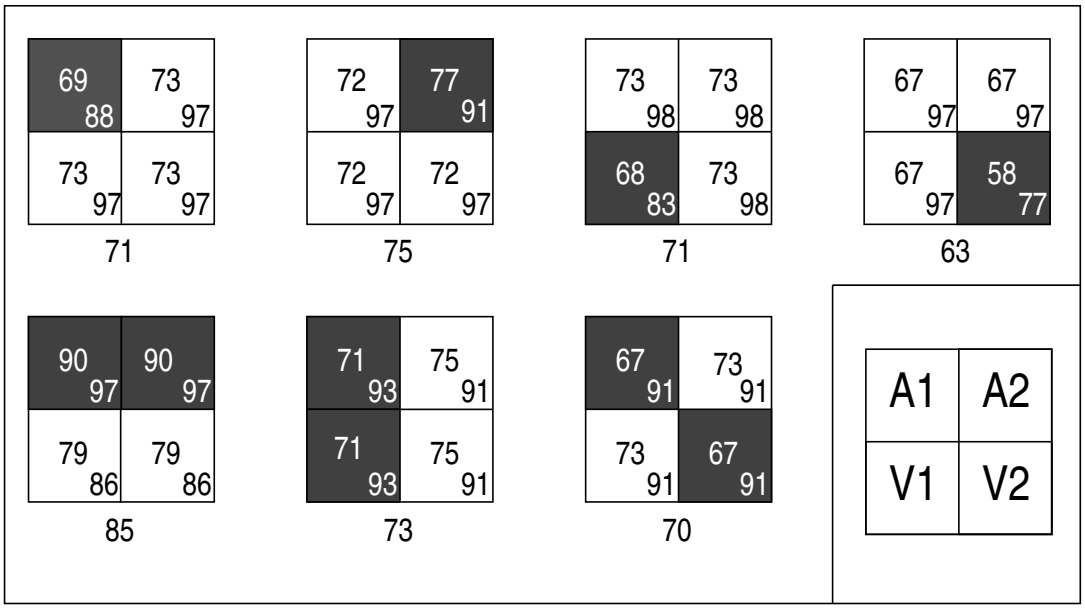

Figure 25: **The Pseudo-Modality Results.** All numbers represent the percentage of correct classifications on the test sets. Each large square represents the results from one division of modalities. Each quadrant represents one of the data subsets A1, A2, V1 and V2 as shown in the key in the lower right. All quadrants colored the same color within a square were in the same modality for training. The number in the middle of the quadrant gives the results of the pseudo-modality of that color after M-D training using the other colored quadrants as the other modality. The numbers in the lower right corner of each quadrant give the corresponding supervised performance for a network given all the quadrants of the same color (in that square). The standard deviations (across the 10 runs) vary from ± 4 to ± 7 for the M-D runs and ± 1 to ± 3 for the supervised runs.

and are conditionally uncorrelated (or even better have opposite correlational structure within the class than between classes) with the first modality. The auditory features for similar frequencies and time are correlated with each other, as are sets of similar spatio-temporal visual features; the auditory and visual features are reasonably uncorrelated allowing them to help significantly in the M-D architecture.

Now, let's compare the performance of the various pseudo-modalities trained with the M-D algorithm. The results are shown in Figure 25; they are arrayed in a graphical form to display many numbers at once. By comparing the numbers in the middle of the quadrants with those in the right corners we can measure the teaching ability of a M-D network given that division of modalities. The numbers below each big square give the numerical average of the M-D trained performances of each of the modalities represented by that division. This also gives a rough idea of the utility of the particular modality division. Also, looking in one quadrant across all the squares (network architectures), one can look at the best distribution for teaching that particular quadrant. So for example if the goal was to have an A1 classifier at the end, and the other dimensions A2, V1, V2 were available to distribute, looking at the numbers in the A1 quadrants gives the combination of these dimensions that would be the best for the performance of the modality that includes A1.

The results reveal that the best division is to keep the auditory dimensions together, and separate from the visual dimensions. Also observe that the A1 dimensions are best helped with the M-D algorithm by training with A2 and using V1 and V2 as the teaching modality. This may

not be that surprising as the A2 dimensions are very informative (as measured by their supervised performance) and by adding them to the A1 side, they are available during testing. What may be more surprising, is that for V1, the best combination is to have V2 on it's side and A1 and A2 on the *other* side. Again this results from the relative suitabilities of the input division. There is an increased benefit obtained by having the teaching network's input conditionally *uncorrelated* with those of its pair (uncorrelated within a class). (Note that another totally correlated input would provide very little new information).

Finally we note that though the M-D algorithm was described as an algorithm to minimize the disagreement between the output of two sensory modalities it can be viewed more generally (keeping in mind the advantage of independence between the modalities). Ideally the algorithm could be applied hierarchically with lower levels minimizing disagreement between submodalities such as color and form. For example as the lighting changes, the color of an object may change while its form may stay relatively constant. On the other hand, while moving around the form will change but the color will stay relatively constant. The M-D algorithm could also be applied hierarchically in a spatial sense. Low-level processing could minimize disagreement between small areas while high-level processing would be minimizing disagreement over larger areas.

The algorithm can also be generalized to apply to patterns near in time. The two networks would represent delayed versions of the same input stream. The signal to the other "modality" could be a temporally close sample from the same modality. As sensations change slowly over time, the network could learn to classify objects by minimizing the disagreements between outputs for input patterns received close in time. This approach is more powerful than that of (Foldiak, 1991) as signals close in time need not be mapped to the *same* codebook vector but the closest codebook vector of the *same class*.

6 Concluding Remarks

6.1 Summary

Many problems are better classified when the labels of the data points are available during training and classifiers can be further improved if the network attempts to minimize the number of misclassifications. We showed how to approach the performance of the minimizing misclassifications classifier without requiring labeled input patterns. The labels were replaced by an assumption that the world present patterns to different modalities in such a way that patterns from Modality 1's Class A occur with patterns from Modality 2's class A distribution. By iterating the algorithm, the labeling algorithm is able to take advantage of the better codebook vector placement and produce better results allowing the Minimizing-Disagreement algorithm to perform even better. With this two iteration algorithm the performance on the benchmark dataset was within 5% of the analogous supervised algorithm. The results were better when the confusable classes were different for the two modalities as they could provide better labeling where it was needed most. In this case the performance was within 2% of the supervised algorithm. The experiments on the visual-auditory speech data showed that the algorithm works on higher dimensional "real" patterns and the single-speaker experiments demonstrated the ability of the algorithm to work with co-occurrences present in natural cross-modal signals. The experiments with the pseudo-modalities indicate that performance with the M-D algorithm is much better when the modalities are divided with correlated dimensions kept together and those that are independent (given the class label) separated (as for example in the visual/auditory separation).

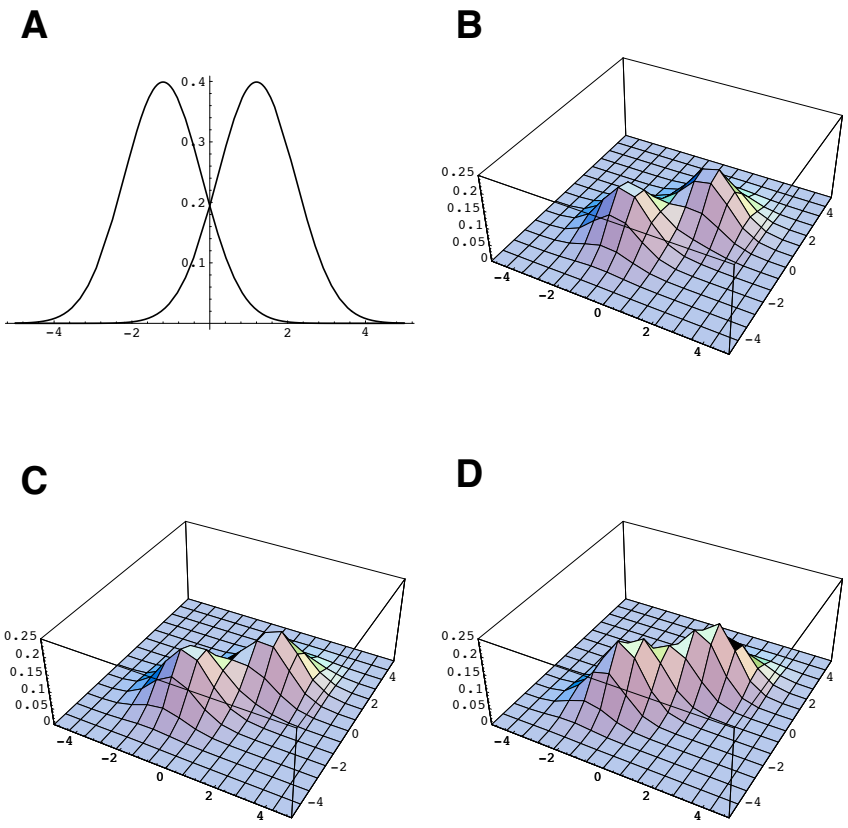

Figure 26: **The Effect of Correlated Dimensions.** A) Example distributions for two classes in one modality
B) Joint distribution for two modalities as in A with no cross-modal correlation within the class
C) Joint distribution for two modalities as in A with stronger cross-modal correlations
D) Joint distribution for the two modalities as in A with even stronger correlations.

6.2 Discussion

The work is shaped by the belief that to some extent, and particularly at higher levels, perceptual learning is shaped by the tasks for which it will be needed. In other words, the perceptual development of features is not an unsupervised bottom-up process but involves top-down feedback about the utility of the current features. This idea is also expressed in (Goldstone & Schyns, 1994) and is implicit in the work of Becker and Hinton (1992; Becker, 1993).

From a biological viewpoint, this work offers an explanation for why cells in one sensory area also respond to inputs to another sensory modality. We have shown that without connecting neurons to all sensory input we can still take advantage of the greater structure available in the higher dimensional total space of inputs. This occurs through integrating the modalities at a higher level and using feedback connections. The work provides an explanation for the ubiquitous back projections in cortex whose purpose is not yet well understood. We suggest that these back-projections are important for the organization of the incoming sensory stimuli during learning.

It has long been argued that the reliable relationships between the sensations to different modalities may be an important source of "teaching" information; in this chapter we have embodied the argument in a powerful working model. We have developed an abstract algorithm for making use of the cross-modal information, examined its properties mathematically and demonstrated its performance empirically on real data.

Acknowledgements

This chapter was largely excerpted from (de Sa, 1994b). In preparing this chapter, VdS was supported by postdoctoral fellowships from the Sloan Foundation and from the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank Robert Goldstone, Philippe Schyns and an anonymous reviewer for comments on an earlier version of this chapter.

References

Artola, A., Bröcher, S., & Singer, W. (1990, September). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*, 69–72.

Becker, S. (1993). Learning to categorize objects using temporal coherence. In C. Giles, S.J.Hanson, & J. Cowan (Eds.), *Advances in Neural Information Processing Systems 5* (pp. 361—368). Morgan Kaufmann.

Becker, S., & Hinton, G. E. (1992, January). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161–163.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Buser, P., & Borenstein, P. (1959). Responses somesthesiques, visuel et auditives, recuellies, au niveau du cortex "associatif" infrasylvien chez le chat curarise non anesthesie. *Electroencephalog. Clin. Neurophysiol.*, *11*, 285–304.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588.

Coultrip, R., Granger, R., & Lynch, G. (1992). A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, *5*, 47–54.

Desimone, R., & Ungerleider, L. G. (1989). Neural mechanisms of visual processing in monkeysIn F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology* (Vol. 2). Elsevier Science Publishers B. V.

Diamantini, C., & Spalvieri, A. (1995, November). Pattern classification by the bayes machine. *Electronics Letters*, *31*(24), 2086–2088.

Durgin, F. H. (1995). *Contingent aftereffects of texture density: Perceptual learning and contingency.* Unpublished doctoral dissertation, Department of Psychology, University of Virginia.

Durgin, F. H., & Proffitt, D. R. (1996). Visual learning in the perception of texture: Simple and contingent aftereffects of texture density. *Spatial Vision*, *9*(4), 423–474.

Edelman, G. M., Jr., G. N. R., Gall, W. E., Tononi, G., & Williams, D. (1992, August). Synthetic neural modeling applied to a real-world artifact. *Proc. Natl. Acad. Sci.*, *89*, 7267–7271.

Fishman, M. C., & Michael, C. R. (1973). Integration of auditory information in the cat's visual cortex. *Vision Research*, *13*, 1415–1419.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

George N. Reeke, J., Sporns, O., & Edelman, G. M. (1990, September). Synthetic neural modeling: The "darwin" series of recognition automata. *Proceedings of the IEEE*, *78*, 1498–1530.

Goldstone, R., & Schyns, P. (1994). Learning new features of representation. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society* (pp. 974–978). Erlbaum, Hillsdale NJ.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*, 187–202.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.

Hefferline, R. F., & Perera, T. B. (1963, march). Proprioceptive discrimination of a covert operant without its observation by the subject. *Science*, *139*, 834–835.

Howells, T. (1944). The experimental development of color-tone synesthesia. *Journal of Experimental Psychology*, *34*(2), 87—103.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Kohonen, T. (1990). Improved versions of learning vector quantization. In *IJCNN International Joint Conference on Neural Networks* (Vol. 1, pp. I–545–I–550).

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*, 361—381.

Lehky, S., & Sejnowski, T. (1988). Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature*, *333*, 452–454.

Linsker, R. (1986a, November). From basic network principles to neural architecture: Emergence of orientation-selective cells. *Proc. Natl. Acad. Sci. USA*, *83*, 8390–8394.

Linsker, R. (1986b, November). From basic network principles to neural architecture: Emergence of orientation columns. *Proc. Natl. Acad. Sci. USA*, *83*, 8779–8783.

Linsker, R. (1986c, October). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proc. Natl. Acad. Sci. USA*, *83*, 7508–7512.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–257.

Maunsell, J., Sclar, G., Nealey, T., & DePriest, D. (1991). Extraretinal representations in area V4 of macaque monkey. *Visual Neuroscience*, *7*(6), 561–573.

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

Merigan, W. H., & Maunsell, J. H. R. (1993). How parallel are the primate visual pathways? In *Annual Review of Neuroscience* (Vol. 16, pp. 369–402).

Meulders, M., Colle, J., & Biosacq-Schepens, N. (1965). Macro and microelectrode studies of somatic responses in the lateral geniculate body. In *Proceedings, XXIII International Congress of Physiological Sciences* (p. 364).

Miikkulainen, R. (1991). Self-organizing process based on lateral inhibition and synaptic resource redistribution. In T. Kohonen, K. Makisära, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks* (pp. 415–420). Elsevier Science Publishers.

Miller, G. A., & Nicely, P. E. (1955, March). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989, August). Ocular dominance column development: Analysis and simulation. *Science*, *245*, 605–615.

Morrell, F. (1972, July). Visual system's view of acoustic space. *Nature*, *238*, 44–46.

Munro, P. (1988, January). *Self-supervised Learning of Concepts by Single Units and "Weakly Local" Representations* (Report No. LIS003/IS88003).

Murata, K., Cramer, H., & Bach-y-Rita, P. (1965). Neuronal convergence of noxious, acoustic and visual stimuli in the visual cortex of the cat. *Journal of Neurophysiology*, *28*, 1233–1239.

Nowlan, S. J. (1991). *Soft Competitive Adaptation: Neural Network Learning Algorithms based on Fitting Statistical Mixtures.* Unpublished doctoral dissertation, School of Computer Science, Carnegie Mellon University.

Obermayer, K., Ritter, H., & Schulten, K. (1990, November). A principle for the formation of the spatial structure of cortical feature maps. *Proc. Natl. Acad. Sci.*, *87*, 8345–8349.

Obermayer, K., Schulten, K., & Blasdel, G. (1992). A comparison between a neural network model for the formation of brain maps and experimental data. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in Neural Information Processing Systems 4* (pp. 83–90). Morgan Kaufmann.

Peterson, G., & Barney, H. (1952). Control methods used in a study of vowels. *The Journal of the Acoustical Society of America*, *24*, 175–184.

Phillips, W., Kay, J., & Smyth, D. (1995). The discovery of structure by multi-stream networks of local processors with contextual guidance. *Network: Computation in Neural Systems*, *6*, 225—246.

Polana, R. (1994). *Temporal Texture and Activity Recognition.* Unpublished doctoral dissertation, Department of Computer Science, University of Rochester.

Quittner, A., Smith, L., Osberger, M., Mitchell, T., & Katz, D. (1994). The impact of audition on the development of visual attention. *Psychological Science*, *5*(6), 347—353.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, *5*, 289–304.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Math. Stat.*, *22*, 400—407.

Rolls, E. (1989). The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampusIn R. Durbin, C. Miall, & G. Mitchison (Eds.), *The Computing Neuron* (). Addison-Wesley.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533—536.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1, pp. 151–193). MIT Press.

Sa, V. de, & Ballard, D. (1992). Top-down teaching enables task-relevant classification with competitive learning. In *IJCNN International Joint Conference on Neural Networks* (Vol. 3, pp. III–364—III–371).

Sa, V. R. de. (1994a). Minimizing disagreement for self-supervised classification. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, & A. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School* (pp. 300—307). Erlbaum Associates.

Sa, V. R. de. (1994b). *Unsupervised Classification Learning from Cross-Modal Environmental Structure*. Unpublished doctoral dissertation, Department of Computer Science, University of Rochester. also available as TR 536 (November 1994)

Sams, M., Aulanko, R., Hämääinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*, 141–145.

Sklansky, J., & Wassel, G. N. (1981). *Pattern Classifiers and Trainable Machines*. Springer-Verlag.

Spinelli, D. (1967). Receptive field organization of ganglion cells in the cat's retina. *Experimental Neurology*, *19*, 291–315.

Spinelli, D., Pribram, K., & Weingarten, M. (1965). Centrifugal optic nerve responses evoked by auditory and somatic stimulation. *Experimental Neurology*, *12*, 303–319.

Spinelli, D., Starr, A., & Barrett, T. W. (1968). Auditory specificity in unit recordings from cat's visual cortex. *Experimental Neurology*, *22*, 75–84.

Spinelli, D., & Weingarten, M. (1966). Afferent and efferent activity in single units of the cat's optic nerve. *Experimental Neurology*, *15*, 347–362.

Stork, D. G., Wolff, G., & Levine, E. (1992). Neural network lipreading system for improved speech recognition. In *IJCNN International Joint Conference on Neural Networks* (Vol. 2, pp. II–286—II–295).

Sumby, W., & Pollack, I. (1954, March). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), 212–215.

Tan, A.-H. (1995). Adaptive resonance associative map. *Neural Networks*, *8*(3), 437–446.

Wassel, G. N., & Sklansky, J. (1972). Training a one-dimensional classifier to minimize the probability of error. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(4), 533—541.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Unpublished doctoral dissertation, Harvard University.

Yuhas, B., Jr., M. G., & Sejnowski, T. (1988). Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, *78*(10), 1658—1668.

Zellner, D. A., & Kautz, M. A. (1990). Color affects perceived odor intensity. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 391–397.

Zipser, D., & Andersen, R. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*, 679–684.