TECHNICAL
RESEARCH
REPORT

SYSTEMS
RESEARCH
CENTER

*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
Industry and the University*

Analysis and Synthesis of Feedforward Neural Networks Using Discrete Affine Wavelet Transformations

by Y.C. Pati and P.S. Krishnaprasad

TR 90-44r1

Analysis and Synthesis of Feedforward Neural Networks Using Discrete Affine Wavelet Transformations *

Y. C. Pati † P. S. Krishnaprasad

Electrical Engineering Department and Systems Research Center
University of Maryland,College Park, MD 20742

Abstract

In this paper we develop a representation of a class of feedforward neural networks in terms of discrete affine wavelet transforms. It is shown that by appropriate grouping of terms, feedforward neural networks with sigmoidal activation functions can be viewed as architectures which implement affine wavelet decompositions of mappings. This result follows simply from the observation that standard feedforward network architectures possess an inherent translation-dilation structure and every node implements the same activation function. It is shown that the wavelet transform formalism provides a mathematical framework within which it is possible to perform both analysis and synthesis of feedforward networks. For the purposes of analysis, the wavelet formulation characterizes a class ($L^2(\mathbb{R})$) of mappings which can be implemented by feedforward networks as well as reveals the exact implementation of a given mapping in this class. Spatio-spectral localization properties of wavelets can be exploited in synthesizing a feedforward network to perform a given approximation task. Synthesis procedures based on spatio-spectral localization result in reducing the training problem to one of *convex* optimization. We outline two such synthesis schemes.

*This research was supported in part by the National Science Foundation's Engineering Research Centers Program: NSFD CDR 8803012, the Air Force Office of Scientific Research under contract AFOSR-88-0204 and by the Naval Research Laboratory.

†Also with Nanoelectronics Processing Facility, Code 6804, Naval Research Laboratories, Washington D.C.

in choosing the particular set of 'basis' functions which are used to implement the transform. In the case of discrete affine wavelet transforms, which we discuss in Section 3, the 'basis' functions are generated by translating and dilating a single function.

In Section 4 we demonstrate that affine wavelet decompositions of functions can be implemented within the standard architecture of feedforward neural networks. Sigmoidal functions have traditionally been used as 'activation' functions of nodes in a neural network. Section 4.1 is concerned with constructing a wavelet 'basis' using combinations of sigmoids. For simplicity, we restrict discussion to networks designed to learn one-dimensional maps. One of the main results of this paper is Theorem 4.1. In Section 4.2 we briefly describe extensions of these results to higher dimensions.

In Section 5 we outline two schemes in which spatio-spectral localization properties of wavelets are used to formulate synthesis procedures for feedforward neural networks. It is shown that such synthesis procedures can result in systematic definition of network topology and simplified network 'training' problems. Most of the weights in the network are determined via the synthesis process and the remaining weights may be obtained as a solution to a *convex* optimization problem. Since the resulting optimization problem is one of least squares approximation, the remaining weights can also be determined by solving the associated 'normal equations'.

A few simple numerical simulations of the methods of this paper are provided in Section 5.4.

2 Functional Approximation and Neural Networks

This section provides a brief introduction to the application of feedforward neural networks to functional approximation problems.

Let Θ be a set containing pairs of sampled inputs and the corresponding outputs generated by an unknown map, $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$, $m, n < \infty$, i.e. $\Theta = \{(x^i, y^i) : y^i = f(x^i);\ x^i \in \mathbb{R}^m,\ y^i \in \mathbb{R}^n, i = 1, \ldots, K,\ K < \infty\}$. We call Θ the *training set.* Note that the samples in Θ need not be uniformly distributed. In this context, the task of functional approximation is to use the data provided in Θ to 'learn' (approximate) the map f. Many existing schemes to perform this task are based on parametrically fitting a particular functional form to the given data. Simple examples of such schemes are those which attempt to fit linear models or polynomials of fixed degree to the data in Θ. More recently, nonlinear feedforward neural networks have been applied to the task of 'learning' the map f. In the interest of keeping this papaer self-contained, an overview of the neural network approach is given below.

2.1 Feedforward Neural Networks

The basic component in a feedforward neural network is the single 'neuron' model depicted in Figure 2(a). Where $u_1, \ldots, u_n$ are the inputs to the neuron, $k_1, \ldots, k_n$ are multiplicative weights applied to the inputs, I is a biasing input, $g : \mathbb{R} \rightarrow \mathbb{R}$, and y is the output of the neuron. Thus $y = g(\sum_{i=1}^{n} k_i u_i + I)$. The 'neuron' of Figure 2(a) is often depicted as shown in Figure 2(b) where the input weights, bias, summation, and function g are implicit. Traditionally, the *activation* function g has been chosen to be the sigmoidal nonlinearity shown in Figure 3. This choice of g

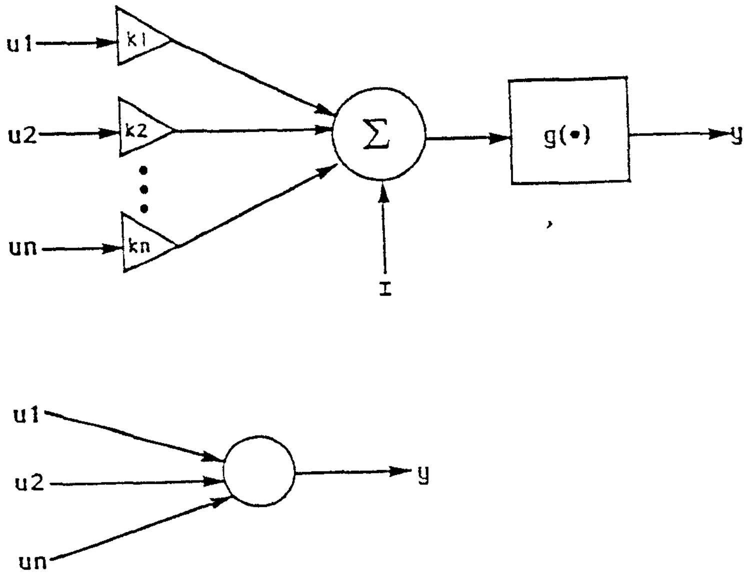

Figure 2: *(a)Single neuron model. (b) Simplified schematic of single neuron*

was initially based upon the observed firing rate response of biological neurons. A feedforward neural network is constructed by interconnecting a number of neurons (such as the one shown in Figure 2) so as to form a network in which all connections are made in the forward direction (from input to output without feedback loops) as in Figure 4. Neural networks of this form are usually comprised of an input layer, a number of hidden layers, and an output layer. The input layer consists of neurons which accept external inputs to the network. Inputs and outputs of the hidden layers are internal to the network, and hence the term 'hidden'. Outputs of neurons in the output layer are the external outputs of the network. Once the structure of a feedforward network has been decided, i.e the number of hidden layers and the number of nodes in each hidden layer has been set, a mapping is 'learned' by varying the connection weights, w_{ij}'s and the biases, I_j's so as to obtain the desired input-output response for the network[1].

One method often used to vary the weights and biases is known as the backpropagation algorithm in which the weights and biases are modified so as to minimize a cost functional of the form,

$$E = \sum_{(x^i,y^i)\in\Theta} \|O^i - y^i\|^2_{\mathbb{R}^n}, \tag{1}$$

where O^i is the output vector (at the output layer) of the network when x^i is applied at the input. Backpropagation employs gradient descent to minimize E. That is, the weights and biases are varied in accordance with the rules,

$$\Delta w_{ij} = -\epsilon \frac{\partial E}{\partial w_{ij}} \quad \text{and} \quad \Delta I_j = -\epsilon \frac{\partial E}{\partial I_j}.$$

Feedforward neural networks are known to have empirically demonstrated ability to approximate complicated maps very well using the technique just described. However, to date

[1] We will use w_{ij} to denote the weight applied to the output O_j of the j^{th} neuron when connecting it to the input of the i^{th} neuron. I_j is the bias input to the j^{th} neuron.

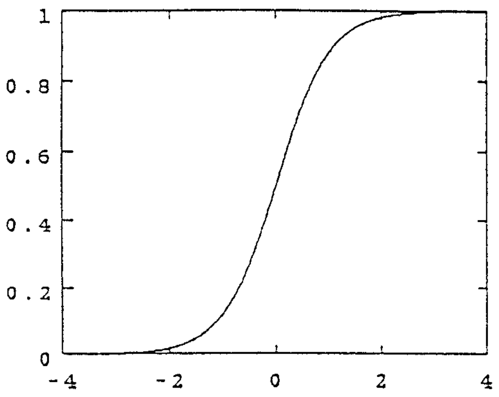

Figure 3: *Sigmoidal nonlinearity.*

there does not exist a satisfactory theoretical foundation for such an approach. We feel that a *satisfactory* theoretical foundation should provide more than just a proof that feedforward networks can indeed approximate certain classes of maps arbitrarily well. Some of the problems that one should be able to address within a good theoretical setting are the following:

(1) Development of a well-founded systematic approach to choosing the number of hidden layers and the number of nodes in each hidden layer required to achieve a given level of performance in a given application.

(2) Learning algorithms often ignore much of the information contained in the training data, and thereby overlook potential simplification of the weight setting problem. As we will show later, preprocessing of training data results in convexity of the training problem.

(3) An inability to adequately explain empirically observed phenomena. For example, the cost functional E may possess many local minima due to the nonlinearities in the network. A gradient descent scheme such as backpropagation is bound to settle to such local minima. However, in many cases, it has been observed that settling to a local minimum of E does not adversely affect overall performance of the network. Observations such as this demand a suitable explanatory theoretical framework.

The methods of this paper offer a framework within which it is possible to address at least the first two issues above.

3 Time-Frequency Localization and Discrete Affine Wavelet Transforms

In this section we review some basic properties of frames and discrete affine wavelet transforms. We also introduce some definitions to formalize the concept of time-frequency localization. To avoid confusion, we point out that throughout this paper we will refer to the domain of the map to be approximated as time or space interchangably.

Given a separable Hilbert space $\mathcal{H}$, we know that it is possible to find an orthonormal basis $\{h_n\}$ such that for any $f \in \mathcal{H}$ we can write the Fourier expansion $f = \sum_n a_n h_n$ where

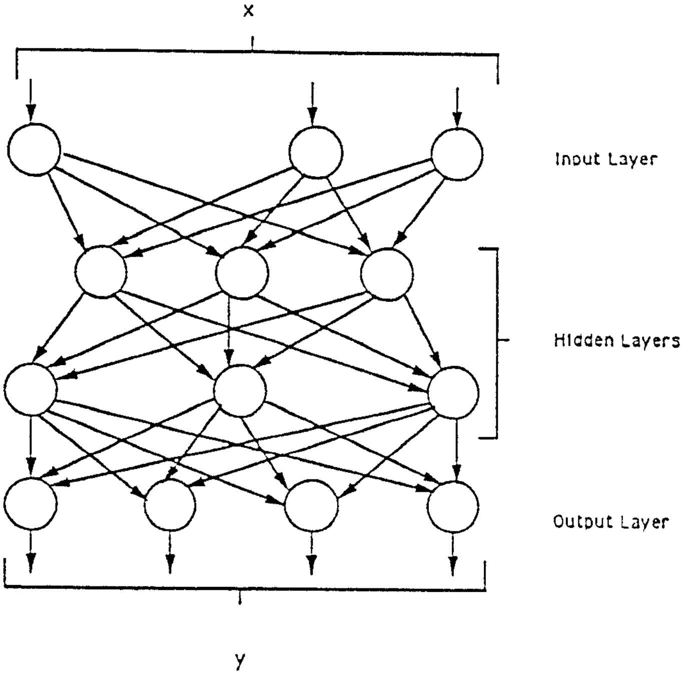

Figure 4: *Feedforward neural network.*

$a_n = < f, h_n >$. For example, the trignometric system $\left\{\frac{1}{\sqrt{2\pi}}e^{j2\pi nt}\right\}$ is an orthonormal basis for the Hilbert space $L^2[-\pi, \pi]$. The Fourier expansion of a signal with respect to the trignometric system is useful in frequency analysis of the signal since each basis element $\frac{1}{\sqrt{2\pi}}e^{j2\pi nt}$ is localized in frequency at $\omega = n$. Hence the distribution of coefficients appearing in the Fourier expansion provides information about the frequency composition of the original signal. In many applications it is desirable to be able to obtain a representation of a signal which is localized to a large extent in both time and frequency. The utility of joint time-frequency localization is easily illustrated by noting that the coefficients in the Fourier expansion of the signal shown in Figure 5 do not readily reveal the fact that the signal is mostly flat and that high

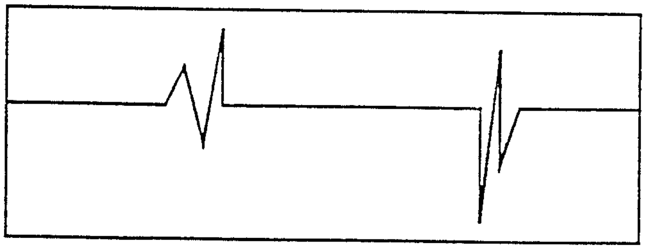

Figure 5: *Signal for which time-frequency localized representations are useful*

frequency components are localized to a short time interval. Examples of applications where time-frequency localization is desirable can be found for instance in image processing [17] [16] [7] [23], and analysis of sound patterns [12]. One method of obtaining such localization is the windowed Fourier transform . This involves taking the Fourier transform of a signal in small time windows which are defined by a window function. Hence the windowed Fourier transform provides information about the frequency content of a signal over a relatively short interval of time. Doubly 'localized' (well concentrated in both time and frequency) representation is one of the primary benefits of wavelet decompositions. However, in obtaining such a localized representation using 'nice' 'basis' functions, it is sometimes necessary to sacrifice the convenience of decomposing signals with respect to an orthonormal basis. Instead it becomes necessary to consider generalizations of orthonormal bases which are called *frames.*

3.1 Frames in Hilbert Spaces

Frames, which were first introduced by Duffin and Schaeffer in [8], are natural generalizations of orthonormal bases for Hilbert spaces.

Definition 3.1 *Given a Hilbert space $\mathcal{H}$ and a sequence of vectors $\{h_n\}_{n=-\infty}^{\infty} \subset \mathcal{H}$, $\{h_n\}_{n=-\infty}^{\infty}$ is called a frame if there exist constants $A > 0$ and $B < \infty$ such that*

$$A\|f\|^2 \leq \sum_n |< f, h_n >|^2 \leq B\|f\|^2, \tag{2}$$

for every $f \in \mathcal{H}$. A and B are called the frame bounds.

Remarks:

(a) A frame $\{h_n\}$ with frame bounds $A = B$ is called a *tight frame.*

(b) Every orthonormal basis is a tight frame with $A = B = 1$.

(c) A tight frame of unit-norm vectors for which $A = B = 1$ is an orthonormal basis.

Given a frame $\{h_n\}$ in the Hilbert space $\mathcal{H}$, with frame bounds A and B, we can define the *frame operator*, $S : \mathcal{H} \longrightarrow \mathcal{H}$ as follows. For any $f \in \mathcal{H}$,

$$Sf = \sum_n < f, h_n > h_n. \tag{3}$$

The following theorem lists some properties of the frame operator which we shall find useful. Proofs of these and other related properties of frames can be found in [9] or [6].

Theorem 3.1

(1) S is a bounded linear operator with $AI \leq S \leq BI$, where I is the identity operator in $\mathcal{H}$.

(2) S is an invertible operator with $B^{-1}I \leq S^{-1} \leq A^{-1}I$.

(3) Since $AI \leq S \leq BI$ implies that $\|I - \frac{2}{A+B}S\| \leq 1$, S^{-1} can be computed via the Neumann series,

$$S^{-1}g = \frac{2}{A+B}\sum_{k=0}^{\infty}\left(I - \frac{2}{A+B}S\right)^k g. \tag{4}$$

(4) *The sequence* $\{S^{-1}h_n\}$ *is also a frame, called the* dual frame, *with frame bounds* B^{-1} *and* A^{-1}.

(5) *Given any* $f \in \mathcal{H}$, f *can be decomposed in terms of the frame (or dual frame) elements as*

$$f = \sum < f, S^{-1}h_n > h_n = \sum < f, h_n > S^{-1}h_n. \tag{5}$$

(6) *Given* $f \in \mathcal{H}$, *if there exists another sequence of coefficients* $\{a_n\}$ *(other than the sequence* $\{< f, S^{-1}h_n >\}$*) such that* $f = \sum a_n h_n$, *then the* a_n*'s are related to the coefficients given in (5) by the formula,*

$$\sum |a_n|^2 = \sum | < f, S^{-1}h_n > |^2 + \sum | < f, S^{-1}h_n > - a_n|^2. \tag{6}$$

3.1.1 Definitions Pertaining to Time-Frequency Localization

In this paper we shall restrict discussion to the Hilbert space $\mathrm{L}^2(\mathbb{R})$ which is the space of all finite energy signals on the real line i.e $f \in \mathrm{L}^2(\mathbb{R})$ if and only if

$$\int_{\mathbb{R}} |f(x)|^2 dx < \infty.$$

If $f, g \in \mathrm{L}^2(\mathbb{R})$ then the inner product $< f, g >$ is defined by

$$< f, g > = \int_{\mathbb{R}} f(x)\overline{g(x)} dx,$$

where $\overline{g}$ denotes the complex conjugate of g, and the norm $\|\cdot\|$ on $\mathrm{L}^2(\mathbb{R})$ is defined by $\|f\|^2 = < f, f >$.

The following definitions are useful in formalizing the concept of time-frequency localization.

Definition 3.2 *Given a function* $f \in \mathrm{L}^2(\mathbb{R})$, $f : \mathbb{R} \to \mathbb{R}$, *with Fourier transform* $\widehat{f}$,

(1) *the center of concentration,* $x_c(f)$, *of* f, *is defined as*

$$x_c(f) = \frac{1}{\|f\|^2} \int_{\mathbb{R}} x|f(x)|^2 dx.$$

(2) *the center of concentration,* $\omega_c(|\widehat{f}|^2)$, *of* $|\widehat{f}|^2$, *(or center frequency of* f*) is defined as*

$$\omega_c(|\widehat{f}|^2) = \frac{1}{\pi\|f\|^2} \int_{[0,\infty)} \omega|\widehat{f}(\omega)|^2 d\omega.$$

Note that $\omega_c(|\widehat{f}|^2)$ is defined so as to account for the evenness of $|\widehat{f}|^2$ for real-valued f; so $\omega_c(|\widehat{f}|^2)$ is the *positive* center frequency of $|\widehat{f}|^2$.

Remark:

The center of concentration $x_c(f)$ can be thought of as the location parameter (in the sense of statistics) of the density $|f|^2/||f||^2$ on $\mathbb{R}$.

Definition 3.3
The support of a function f, denoted supp(f) is the closure of the set $\{x : f(x) > 0\}$.

Definition 3.4 *Given $f \in L^2(\mathbb{R})$, $f : \mathbb{R} \to \mathbb{R}$, with Fourier transform $\widehat{f}$, and centers of concentration $x_c(f)$ and $\omega_c(|\widehat{f}|^2)$,*

$$\mathcal{P}(f;\epsilon) = \left\{ [x_0, x_1] : \; |x_c(f) - x_0| = |x_c(f) - x_1| \text{ and } \int_{x \in \mathbb{R} \backslash [x_0,x_1]} |f(x)|^2 dx < \epsilon ||f||^2 \right\},$$

and,

$$\begin{aligned} \widehat{\mathcal{P}}(f;\widehat{\epsilon}) \;\; &= \;\; \Big\{ [\omega_0, \omega_1] : \; \omega_0 = \max(0, \widetilde{\omega_0}), \; |\omega_c(|\widehat{f}|^2) - \widetilde{\omega_0}| = |\omega_c(|\widehat{f}|^2) - \omega_1| \, , \\ & \text{and} \int_{\omega \in \mathbb{R} \backslash [\omega_0, \omega_1]} |\widehat{f}(\omega)|^2 d\omega < \widehat{\epsilon} ||f||^2 \Big\} . \end{aligned}$$

(1) The epsilon support (or time concentration) of f, denoted ϵ-supp(f, ϵ) is the set $[x_o(f), x_1(f)] \in \mathcal{P}(f;\epsilon)$ such that,

$$\mu([x_0(f), x_1(f)]) = \inf_{[x_0,x_1] \in \mathcal{P}} \mu([x_0, x_1]).$$

(2) The epsilon support of $|\widehat{f}|^2$ (or frequency concentration of f) denoted ϵ-supp$(|\widehat{f}|^2, \widehat{\epsilon})$ is the set $[\omega_0(f), \omega_1(f)] \in \widehat{\mathcal{P}}(f;\widehat{\epsilon})$ such that

$$|\omega_1(f) - \omega_0(f)| = \inf_{[\omega_0,\omega_1] \in \widehat{\mathcal{P}}} |\omega_1 - \omega_0| \, .$$

Remark:

The ϵ-support of f is the smallest (symmetric about $x_c(f)$) interval containing $(1 - \epsilon)\times$ the total signal energy. We further note that the notion of ϵ-support introduced here is used later in Section 5 to formulate a synthesis procedure for feedforward neural networks. In particular, the ϵ-support affects the number of hidden layer nodes needed to acheive a given quality of function approximation.

3.2 Discrete Affine Wavelet Transforms

Given a function $g \in L^2(\mathbb{R})$, consider the sequence of functions $\{g_{mn}\}$ generated by dilating and translating g in the following manner,

$$g_{mn}(x) = a^{n/2} g(a^n x - mb), \tag{7}$$

where, $a > 0$ and $b > 0$ and m and n are integers. Let us assume that $g \in L^2(\mathbb{R})$ is real-valued, concentrated at zero with sufficient decay away from zero, and that ϵ-supp$(g, \epsilon) = [-L, L]$, where ϵ is small and chosen such that the energy contribution of g outside $[-L, L]$ is negligible.

In addition, suppose that the Fourier transform $\widehat{g}$ of g is compactly supported, with $\mathrm{supp}(\widehat{g}) = [-\omega_1, -\omega_0] \cup [\omega_0, \omega_1]$ and concentrated at $\omega_c(|\widehat{g}|^2)$, $0 < \omega_0 < \omega_c(|\widehat{g}|^2) < \omega_1 < \infty$. Recalling the dilation property of the Fourier transform,

$$f(ax) \xrightarrow{\mathcal{F}} a^{-1}\widehat{f}(a^{-1}\omega),$$

we see that $\mathrm{supp}(\widehat{g}_{mn}) = [a^n\omega_0, a^n\omega_1] \cup [-a^n\omega_1, a^n\omega_0]$, $\omega_c(|\widehat{g_{mn}}|^2) = a^n\omega_c(|\widehat{g}|^2)$, and that g_{mn} is concentrated about the point $a^{-n}mb$ with ϵ-$\mathrm{supp}(g_{mn}) = [a^{-n}(-L+mb), a^{-n}(L+mb)]$. Hence if we could write an expansion of any $f \in \mathrm{L}^2(\mathbb{R})$ as

$$f = \sum_{mn} c_{mn}(f) g_{mn} \tag{8}$$

then each coefficient $c_{mn}(f)$ provides information about the frequency content of f in the frequency range $\omega \in [a^n\omega_0, a^n\omega_1] \cup [-a^n\omega_1, -a^n\omega_0]$ during the time interval $[a^{-n}(-L+mb), a^{-n}(L+mb)]$ about $x_c(f)$.

Discrete affine wavelet transforms provide a framework within which it is possible to understand expansions of the form given in (8). In a general setting, discrete affine wavelet transforms are based upon the fact that it is possible to construct frames for $\mathrm{L}^2(\mathbb{R})$ using translates and dilates of a single function. That is, for certain functions g it is possible to determine a dilation stepsize a and a translation stepsize b such that the sequence $\{g_{mn}\}$ as defined by (7) is a frame[2] for $\mathrm{L}^2(\mathbb{R})$. In this case (8) is referred to as the wavelet expansion of f. To form an affine frame the *mother wavelet* [3] g must satisfy an admissibility condition,

$$\int \frac{|\widehat{g}(\omega)|^2}{|\omega|} d\omega \; < \; \infty. \tag{9}$$

For a function g with adequate decay at infinity, (9) is equivalent to the requirement $\int g(x)dx = 0$ (see [6]). Since $\widehat{g}(0) = \int g(x)dx$, admissbility (for functions with adequate decay) is equivalent to requiring that $\widehat{g}(0) = 0$. Furthermore $g \in \mathrm{L}^2(\mathbb{R})$ together with admissibility implies the g must have certain approximate 'bandpass' characteristics.

Remarks

- The term *discrete affine* wavelet transform, is derived from the fact that the functions g_{mn} are generated via sampling of the continuous orbit of the left regular representation of the affine $(ax+b)$ group associated to the function g. A review of the implications of group representation theory in wavelet transforms is given in [9].

- Windowed Fourier transforms (of which the Gabor transform [7] [23] is a special case) are obtained via a representation of the group of translations and complex modulations (the Weyl-Heisenberg group) on $\mathrm{L}^2(\mathbb{R})$. An essential difference between windowed Fourier transforms and affine wavelet transforms arises due to the particular group action involved. For windowed Fourier transforms, the window size remains constant as higher frequencies are analyzed using complex modulations. In affine wavelet transforms the higher frequencies are analyzed over narrower windows due to the dilations, thereby providing a mechanism for 'zooming' in on fine details of a signal.

[2] In this case we say that the triplet (g, a, b) generates an affine frame for $\mathrm{L}^2(\mathbb{R})$.

[3] Also referred to as the fiducial vector or analyzing waveform.

4 Dilations and Translations in SISO Neural Networks

In this section we shall demonstrate how affine wavelet decompositions[4] of $L^2(\mathbb{R})$ can be implemented within the architecture of SISO feedforward neural networks. Consider the single-input-single-output (SISO) feedforward neural network shown in Figure 6. Input and output layers of this network each consist of a single node, whose activation function is linear with unity gain.

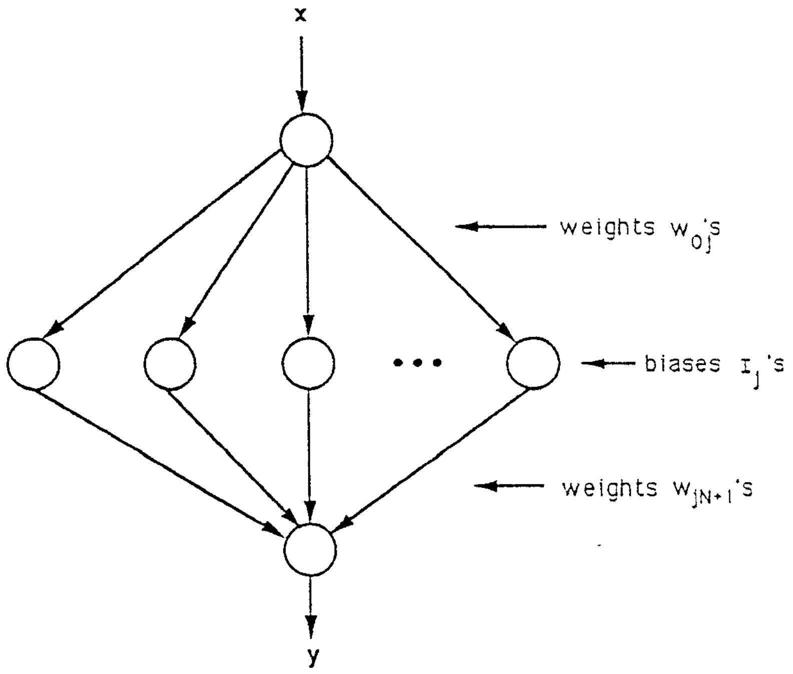

Figure 6: *SISO feedforward neural network*

In addition, the network has a single hidden layer with N nodes, each with activation function $g(\cdot)$. Hence the output of this network is given by

$$y = f(x) = \sum_{j=1}^{N} w_{j,N+1} g(w_{0,j}x - I_j), \tag{10}$$

where we have labeled the input node 0 and the output node $N+1$. It is clear that (10) is of the form in (8) with two key differences: (i) The summation in (10) is finite, and (ii) Even if we permit infinitely many hidden layer nodes, and let $g_j = g(w_{0,j}x - I_j)$, the infinite sequence $\{g_n\}$ will not necessarily be a frame. Since it is our intent to stay within the general framework of feedforward neural networks, let us first consider the sigmoidal function, $s(x) = (1 + e^{-x})^{-1}$ shown in Figure 3 as a possible mother wavelet candidate. Since $s \notin L^2(\mathbb{R})$, it is impossible to construct a frame for $L^2(\mathbb{R})$ using individual translated and dilated sigmoids as frame elements. However, we note that the difference of two translated sigmoids is in $L^2(\mathbb{R})$ for finite translations

[4] Throughout the rest of this paper we will use the term wavelet transform to mean discrete affine wavelet transform unless otherwise indicated

and that in general if we let

$$\varphi(x) = \sum_{n=1}^{M} s_{a_n b_n}(x) - \sum_{n=1}^{M} s_{c_n d_n}(x) \tag{11}$$

where $M < \infty$ and $s_{ab}(x) = s(ax - b)$, $a, b < \infty$ then $\varphi \in \mathrm{L}^2(\mathbb{R})$. With this observation, we show that it is possible to construct frames using combinations of sigmoids as in (11).

4.1 Frames From Sigmoids

Let $s(x) = (1 + e^{-qx})^{-1}$, where $q > 0$ is a constant which controls the 'slope' of the sigmoid . To obtain a function in $\mathrm{L}^2(\mathbb{R})$, we combine two sigmoids as in (11). Let

$$\varphi(x) = s(x + d) - s(x - d), \ 0 < d < \infty. \tag{12}$$

So, $\varphi(\cdot)$ (see Figure 7) is an even function which decays exponentially away from the origin.

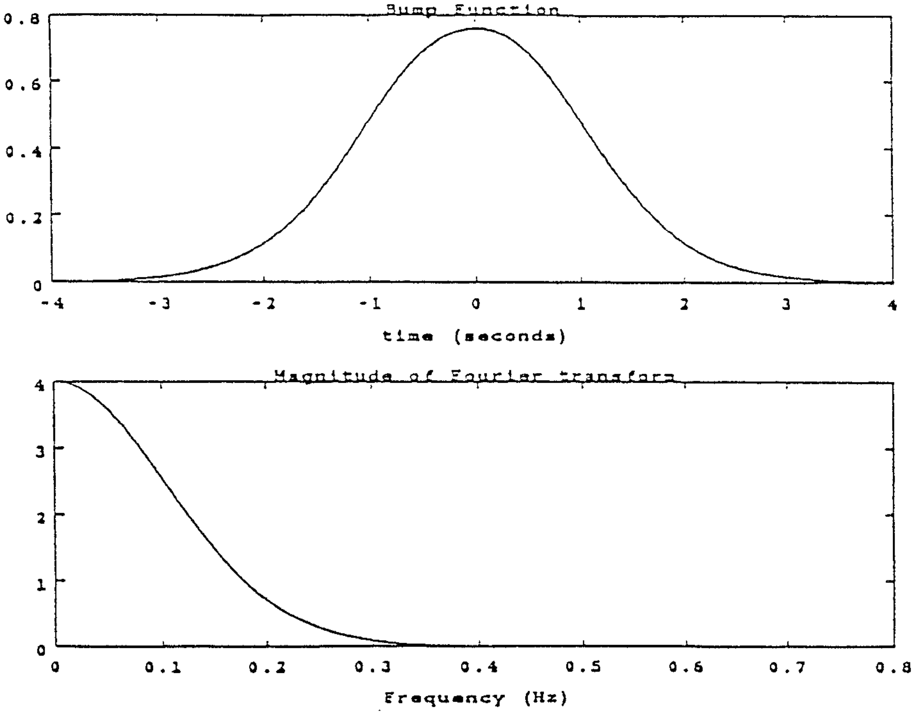

Figure 7: $\varphi(x)$ - *Sum of two sigmoids, and the magnitude of its Fourier transform*

Now, let

$$\psi(x) = \varphi(x + p) - \varphi(x - p). \tag{13}$$

$\psi(\cdot)$ (Figure 8) is an odd function, with $\int \psi(x)dx = 0$, which is dominated by a decaying exponential, and it can be shown that ψ satisfies the admissibility condition (9). The Fourier transform of φ is given by

$$\widehat{\varphi}(\omega) = \int_{\mathbb{R}} \varphi(x) e^{-j\omega x} dx = \frac{2\pi}{q} \frac{\sin(\omega d)}{\sinh(\frac{\pi\omega}{q})} \ , \tag{14}$$

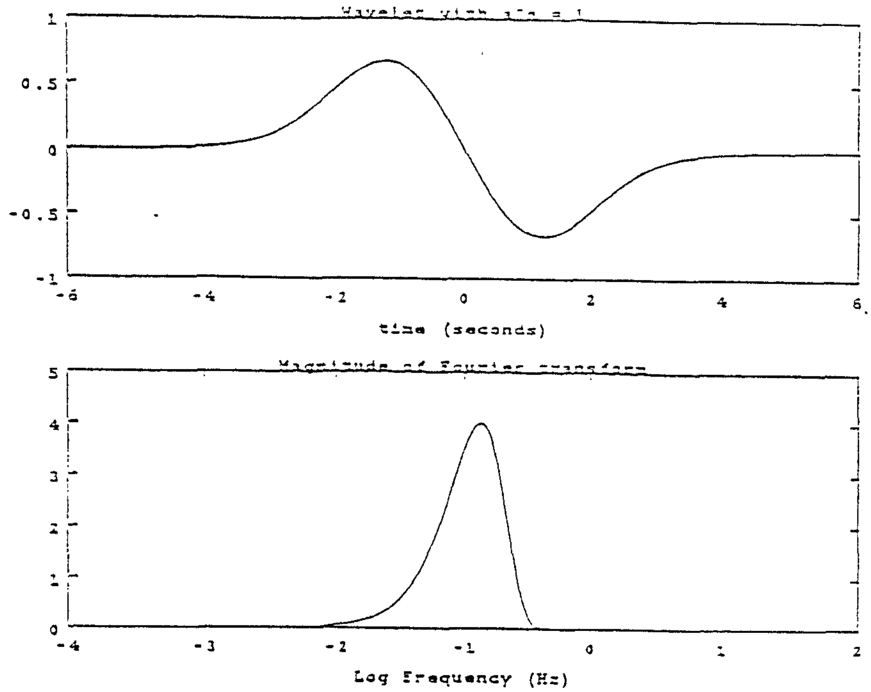

Figure 8: $\psi(x)$ - *Mother Wavelet candidate constructed from sigmoids and magnitude of Fourier Transform of* ψ

which is shown in Figure 7. Therefore the Fourier transform of ψ is,

$$\begin{aligned}\widehat{\psi}(\omega) &= e^{-jp\omega}\widehat{\varphi}(\omega) - e^{jp\omega}\widehat{\varphi}(\omega)\\ &= -j2\sin(p\omega)\widehat{\varphi}(\omega)\\ &= -j\frac{4\pi}{q}\frac{\sin(p\omega)\sin(d\omega)}{\sinh(\frac{\pi\omega}{q})}\end{aligned} \tag{15}$$

which is shown in Figure 8. Note that the function ψ is reasonably well localized in both the time and frequency domains. For the moment, we will concentrate on the specific case where $p = 1$, $d = 1$, and $q = 2$ (which is the case used for the plots shown in Figures 7–8). Table 1 lists some relevant parameters describing the (numerically determined) localization properties of ψ. For this choice of (p, d, q) (and in general whenever $p = d$) ψ is a linear combination of three sigmoids, $\psi(x) = s(x + 2) - 2s(x) + s(x - 2)$. Figure 9 shows the implmentation of ψ in a feedforward network.

ϵ	$\widehat{\epsilon}$	$x_c(\psi)$	$\omega_c(\lvert\widehat{\psi}\rvert^2)$	ϵ-supp(ψ, ϵ)	ϵ-supp$(\lvert\widehat{\psi}\rvert^2, \widehat{\epsilon})$
0.1	0.1	0.0	0.9420	$[-2.15, 2.15]$	$[0.2920, 1.5920]$

Table 1: *Time-frequency localization properties of* ψ *for* $(p, d, q) = (1, 1, 2)$

It is our goal to construct a frame for $L^2(\mathbb{R})$ using ψ as the mother wavelet. That is, we wish to find, if possible, a dilation stepsize a and a translation stepsize b such that the triplet

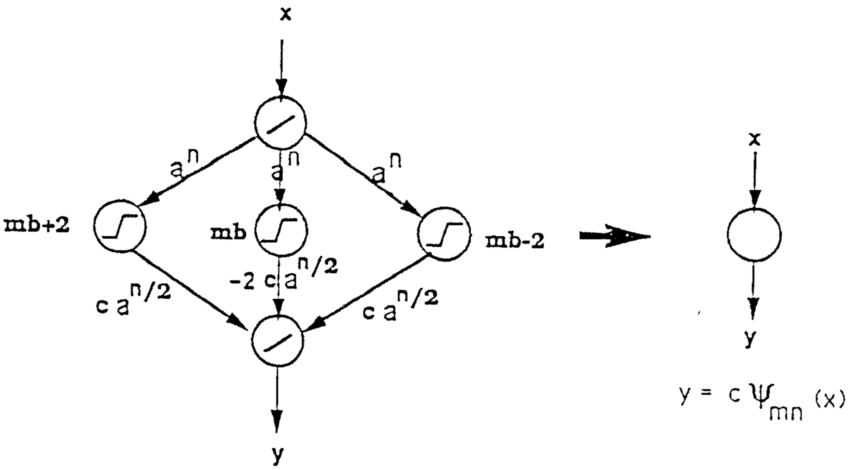

Figure 9: *Feedforward network implementation of ψ*

(ψ, a, b) generates an affine frame for $L^2(\mathbb{R})$. Recall that, we say (ψ, a, b) generates an affine frame for $L^2(\mathbb{R})$ if the sequence $\{\psi_{mn}\}$ is a frame for $L^2(\mathbb{R})$ where, $\psi_{mn} = a^{n/2}\psi(a^n x - mb)$. For the mother wavelet ψ constructed from sigmoids as above, it is possible to determine values of a and b for which (ψ, a, b) generates a frame for $L^2(\mathbb{R})$ (See Appendix A).

It follows that we have constructively proved the following analysis result.

Theorem 4.1 *Feedforward neural networks with sigmoidal activation functions and a single hidden layer can represent any function $f \in L^2(\mathbb{R})$. Moreover, given $f \in L^2(\mathbb{R})$, all weights in the network are determined by the wavelet expansion of f,*

$$f(x) = \sum_{m,n} \langle f, S^{-1}\psi_{mn}\rangle \, \psi_{mn}(x).$$

Remarks:

(a) In this section we have concentrated on wavelets constructed from sigmoids. We would however, like to point out that nonsigmoidal activation functions are also of considerable interest and we refer the reader to [24]. The techniques of wavelet theory should be applicable to such activation functions also.

(b) Among other activation functions used in neural networks, is the discontinuous sigmoid (step) function. Note that using such a step function together with the methods of this section results in a mother wavelet ψ which is the Haar wavelet. Dilates and translates of the Haar function generate an orthonormal basis for $L^2(\mathbb{R})$. The Haar transform is the earliest known example of an affine wavelet transform.

4.2 Wavelets For $L^2(\mathbb{R}^n)$ Constructed From Sigmoids

Although we shall primarily restrict attention to the one-dimensional setting ($L^2(\mathbb{R})$), wavelets for higher dimensional domains ($L^2(\mathbb{R}^n)$) can also be constructed within the standard feedforward network setting with sigmoidal activation functions. In applications such as image processing it is desirable to use wavelets which exhibit orientation selectivity as well as spatio-spectral selectivity. In the setting of Multiresolution Analysis [17] for example, wavelet bases for $L^2(\mathbb{R}^2)$ are constructed using tensor products of wavelets for $L^2(\mathbb{R})$ and the corresponding 'smoothing' functions. This method results in three mother wavelets for $L^2(\mathbb{R}^2)$ each with a particular orientation selectivity. However neural network applications do not necessarily require such orientation selective wavelets. In this case, it is possible to use translates and dilates of a single 'isotropic' function to generate wavelet bases or frames for $L^2(\mathbb{R}^n)$ (c.f. [16]). Figure 4.2 shows both an isotropic mother wavelet and an orientation selective mother wavelet for $L^2(\mathbb{R}^2)$ which are implemented in a standard feedforward neural network architecture with sigmoidal activation functions. The wavelets of Figure 4.2 are implemented by taking differences of 'bump' functions which are generated using a construction given by Cybenko in [1].

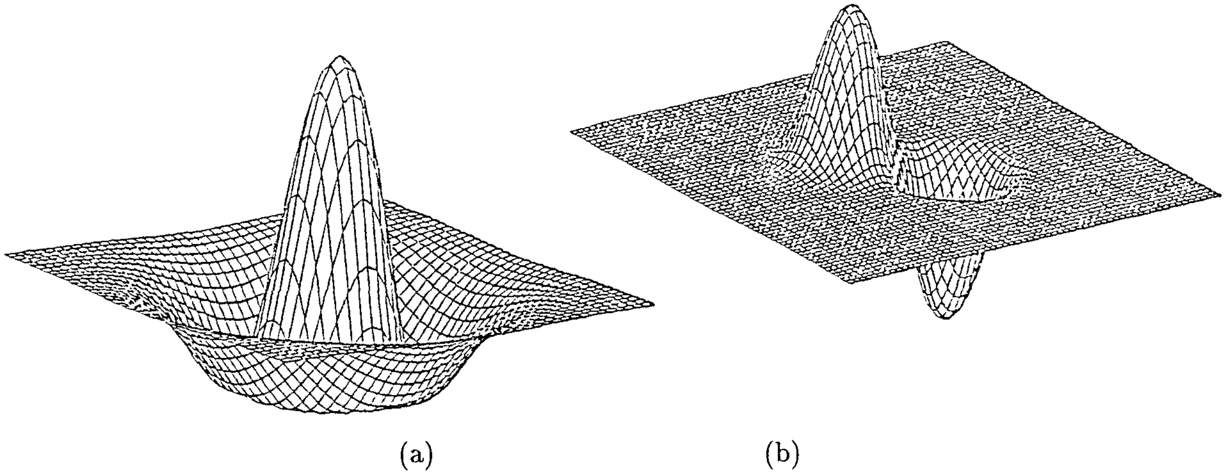

Figure 10: *Two-Dimensional wavelets constructed from sigmoids: (a)Isotropic wavelet, (b)Orientation selective wavelet.*

5 Synthesis of Feedforward Neural Networks Using Wavelets

In the last section, it was shown that it is possible to construct an affine frame for $L^2(\mathbb{R})$ using a function ψ which is a linear combination of three sigmoidal functions. In this section, we shall examine some implications of the wavelet formalism for functional approximation based on sigmoids, in the synthesis of feedforward neural networks. As was described in Section 2.1, sigmoidal functions have served as the basis for functional approximation by feedforward neural networks. However, in the absence of an adequate theoretical framework, topological definitions of feedforward neural networks have for the most part been trial-and-error constructions. We will demonstrate, by means of the simple network discussed in Section 4, how, it is possible to incorporate the joint time and frequency domain characteristics of any given approximation problem into the initial network configuration.

Let $f \in L^2(\mathbb{R})$ be the function which we are trying to approximate. In other words, we are provided a set Θ of sample input-output pairs under the mapping f,

$$\Theta = \{(x^i, y^i) : y^i = f(x^i);\; x^i, y^i \in \mathbb{R}\},$$

and we would like to obtain a good approximation of f. To perform the approximation using a neural network, the first step is to decide on a network configuration. For this problem, it is clear that the input and output layers must each consist of a single node. The remaining questions are how many hidden layers should we use and how many nodes should there be in each hidden layer. These questions can be addressed using the wavelet formulation of the last section. We consider a network of the form in Figure 6, i.e. with a single hidden layer. At this point, a traditional approach would entail fixing the number of nodes N, in the hidden layer and then applying a learning algorithm such as backpropagation (described in Section 2.1) to adjust the three sets of weights, input weights $\{w_{0,j}\}_{j=1}^{N}$, output weights $\{w_{j,N+1}\}_{j=1}^{N}$, and the biases $\{I_j\}$. We would like to use information contained in the training set Θ to, (1) decide on the number of nodes in the hidden layer, and (2) reduce the number of weights that need to be adjusted by the learning algorithm.

Here we describe two possible schemes for use of the wavelet transform formulation in the synthesis of feedforward networks. The first scheme captures the essence of how time-frequency localization can be utilized in the synthesis procedure. However, this scheme is difficlut to implement when considering high dimensional mappings and in most cases will result in a network that is far larger than necessary. We also outline a second method which further utilizes the time-frequency localization offered by wavelets to reduce the size of the network. This second method is conceivably a more viable option in the case of higher dimensional mappings.

5.1 Network Synthesis: Method I

Assume f, the function which we are trying to approximate, is such that $\epsilon\text{-supp}(|\widehat{f}|^2, \widehat{\epsilon}) = [\omega_{\min}, \omega_{\max}]$ where $\omega_{\min} \geq 0$ [5]. Also assume that there exists a finite interval $[x_{\min}, x_{\max}]$ in which we wish to approximate f. Our network synthesis procedure is described in algorithmic form below.

Synthesis Algorithm:

Step I Our first step is to perform a frequency analysis of the training data. In this step we wish to obtain an estimate of the 'bandwidth' $\epsilon\text{-supp}(|\widehat{f}|^2, \widehat{\epsilon})$ of f based on the samples of f provided in Θ. A number of techniques can be considered for performing this estimate. We will not elaborate on such techniques here. Let $\widetilde{\omega}_{\min}$ be our estimate of $\omega_{\min}$, and $\widetilde{\omega}_{\max}$ be our estimate of $\omega_{\max}$.

Step II We now use the knowledge of $\widetilde{\omega}_{\min}$, $\widetilde{\omega}_{\max}$, $x_{\min}$, and $x_{\max}$ to choose the particular frame elements to be used in the approximation. The main idea in this step is to choose only those elements of the frame $\{\psi_{mn}\}$ which 'cover' the region Q_f of the time-frequency plane defined by

$$Q_f(\epsilon, \widehat{\epsilon}) = [x_{\min}, x_{\max}] \times ([\widetilde{\omega}_{\min}, \widetilde{\omega}_{\max}] \cup [-\widetilde{\omega}_{\max}, -\widetilde{\omega}_{\min}]).$$

which represents the concentration of f in time and frequency as determined from the data Θ. Recall that $\epsilon\text{-supp}(|\widehat{\psi}|^2, \widehat{\epsilon}) = [\omega_0(\psi), \omega_1(\psi)]$ and $\epsilon\text{-supp}(\psi, \epsilon) = [x_0(\psi), x_1(\psi)]$ (see Table 1).

[5] Since f is real-valued, we need only consider positive frequencies

Thus the concentration of the mother wavelet ψ in the time-frequency plane is in the region $[x_0(\psi), x_1(\psi)] \times [\omega_0(\psi), \omega_1(\psi)]$. Hence the concentration of ψ_{mn} in the time-frequency plane is

$$\mathcal{Q}_{mn}(\epsilon,\widehat{\epsilon}) = [a^{-n}(x_0(\psi)+mb), a^{-n}(x_1(\psi)+mb)] \times ([a^n\omega_0(\psi), a^n\omega_1(\psi)] \cup [-a^n\omega_1(\psi), -a^n\omega_0(\psi)]),$$

which is centered at $(x_c(\psi_{mn}), \omega_c(|\widehat{\psi}_{mn}|^2) = (x_c(\psi) + a^{-n}mb, a^n\omega_c(|\widehat{\psi}|^2)$. Figure 11 shows the location of $\mathcal{Q}_f$, and the $\mathcal{Q}_{mn}$'s together with the time-frequency concentration centers $(x_c(\psi_{mn}), \omega_c(|\widehat{\psi}_{mn}|^2)$ of the frame elements. Therefore to 'cover' $\mathcal{Q}_f(\epsilon,\widehat{\epsilon})$ we need to determine the index set $\mathcal{I}$ of pairs (m,n) of integer translation and dilation indices such that,

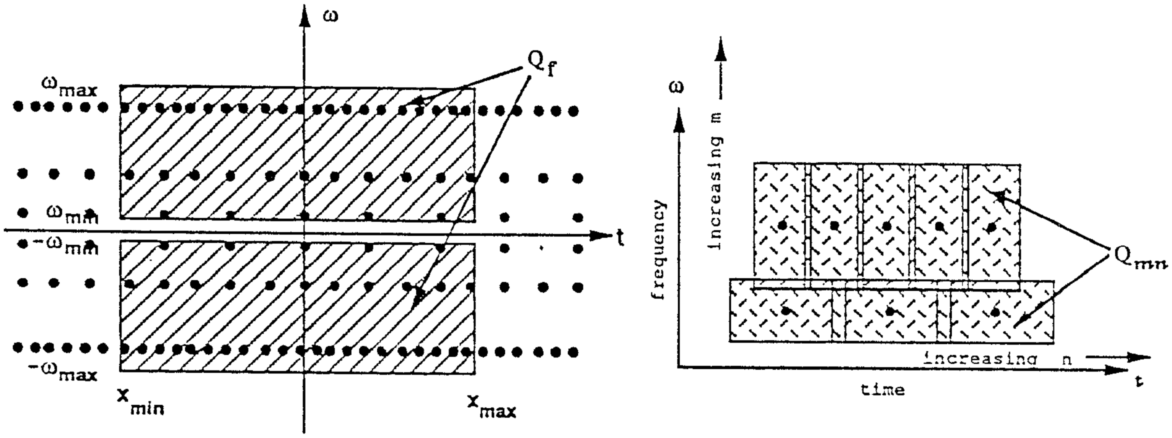

Figure 11: *Time-frequency concentrations $\mathcal{Q}_f$ and $\mathcal{Q}_{mn}$'s together with time-frequency concentration centers $(x_c(\psi_{mn}), \omega_c(|\widehat{\psi}_{mn}|^2)$ of the frame elements.*

$$\mathcal{Q}_{mn} \cap \mathcal{Q}_f \neq \emptyset, \text{ for } (m,n) \in \mathcal{I}.$$

Daubechies in [6] discusses the existence of a bounding box $\mathcal{B}_\epsilon$ surrounding the time-frequency concentration $\mathcal{Q}_f$ of f such that the f can be approximated to any desired precision ϵ by including in the approximation, all frame elements with concentration centers in $\mathcal{B}_\epsilon$.

Step III Given $\mathcal{I}$, it is now possible to configure the network. From the manner in which $\mathcal{I}$ is defined, we expect to be able to obtain an approximation to f of the form

$$f(x) \approx \sum_{(m,n)\in\mathcal{I}} c_{mn}\psi_{mn}(x) \;=\; \widetilde{f}(x). \tag{16}$$

for $x \in [x_{\min}, x_{\max}]$. The approximation error in (16) can be made arbitrarily small by allowing ϵ and $\widehat{\epsilon}$ to go to zero in the computation of the various ϵ-supports used to define the sets $\mathcal{Q}_f$ and $\mathcal{Q}_{mn}$. This is because we know that $\{\psi_{mn}\}$ is a frame and therefore it is possible to write f as

$$f(x) = \sum_{m,n\in\mathbb{Z}} c_{mn}(f)\psi_{mn} \tag{17}$$

for some coefficients $\{c_{mn}(f)\}$. Returning to the single-hidden layer feedforward network shown in Figure 6, choose the number of nodes in the hidden layer to be equal to the number of elements in $\mathcal{I}$, i.e. $N = \#(\mathcal{I})$ where the activation function of each node is taken to be[6] ψ. Now if we set the weights form the input node to the hidden layer and the biases on each hidden layer node to the dilation and translation coefficients indexed by $(m, n) \in\mathcal{I}$, then the output of the network can be written as

$$y = \sum_{(m,n)\in\mathcal{I}} c_{mn}\psi_{mn}(x) \tag{18}$$

where x is the input of the network and c_{mn}'s are the weights form the hidden layer to the output node. We have therefore obtained a network configuration which defines an output function (18) that is exactly of the form required to approximate the function f (Equation (16)).

It remains to determine the coefficients c_{mn}'s in (18) that will result in the desired approximation.

5.2 Network Synthesis: Method II

The synthesis algorithm described above in Section 5.1 uses identification of an 'important' region Q_f of the time-frequency plane. Critical to identification of this region is the 'bandwidth' estimate made in Step I. There are two significant drawbacks of making such a bandwidth estimate:

(1) Estimation of spectral concentration of signals in high dimensions is computationally expensive.

(2) Any estimate of spectral concentration which relies on Fourier techniques is going to generate a generalized rectangle in joint time-frequency space. For many functions such a rectangular concentration in time-frequency is simply an artifact of the spatial nonlocality of the Fourier basis. For example, an estimate of the frequency concentration of the signal in Figure 5 will generate a rectangle in time-frequency as the concentration of the signal. If we then use this rectangle to choose which elements of a wavelet basis to use to approximate the signal, the time-frequency rectangle will dictate that large dilations (corresponding to high frequencies) of the wavelets be used over the entire time interval. However, since each wavelet is also localized in time, and high frequency components of the signal are localized as well, this is clearly an excessive number of wavelets. Large dilations can be used locally where needed.

Spatio-spectral localization properties of wavelets can be further exploited to reduce the number of network nodes (wavelets) used in the approximation. The basic idea is that since wavelets are well-suited to identify spatially local regions of fine scale (high frequency) features in a signal, locations and values local maxima of the wavelet approximation coefficients at one scale (dilation) indicate whether or not it is necessary to locally refine the approximation by the use of wavelets at finer scales (c.f. [18]). A network synthesis algorithm using this idea would be an adaptive procedure of the following form.

[6] Recall that ψ is a linear combination of three sigmoids.

(1) Construct and train a network to approximate the mapping at some scale a^n over the entire spatial region of interest.

(2) Identify local maxima of the wavelet coefficients and locally refine the approximation by adding new dilations (nodes) to the network where needed.

(3) Repeat (2) until some stopping crterion has been satisfied.

Using a scheme such as this would result in approximations being performed over regions of time-frequency of the form shown in Figure 5.2. Some aspects of this scheme are discussed in [22].

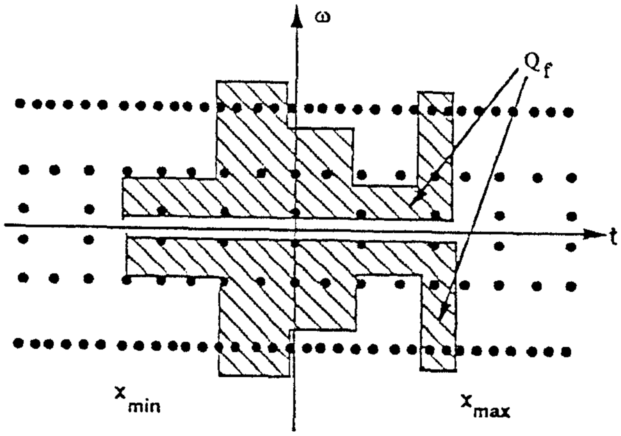

Figure 12: *Form of time-frequency coverage from approximation scheme of Section 5.2*

5.3 Computation of Coefficients

In the case of an infinite expansion via frame elements, there exists (at least in theory) a method of determining the expansion coefficients in terms of the inverse of the frame operator S defined in (3). From (5), we see that given the frame $\{\psi_{mn}\}$, the coefficients in (17) are given by,

$$c_{mn} = < f, S^{-1}\psi_{mn} > . \tag{19}$$

From Theorem 3.1, we see that in principle $S^{-1}\psi_{mn}$ can be computed from the series expansion given in (4). However rate of convergence of this series is governed by how close the frame is to being a tight frame i.e by how close the ratio B/A is to 1. So for 'loose' frames explicit computation of wavelet expansion coefficients may prove overly demanding of computational resources.

Considering now the case of a finite approximation to f as in (16), let Span$\{h_n\}$ denote the closed linear span of the vectors $\{h_n\}$. It is clear that f can be represented exactly by the expansion in (16) *if and only if* $f \in$ Span$\{\psi_{mn},\ (m,n) \in \mathcal{I}\}$. If $f \notin$ Span$\{\psi_{mn},\ (m,n) \in \mathcal{I}\}$ then the 'best'[7] approximation to f in terms of the finite subset of frame elements with indices in $\mathcal{I}$ is the projection of f onto Span$\{\psi_{mn},\ (m,n) \in \mathcal{I}\}$. In this case, we would like to compute the coefficients of expansion of the projection of f onto Span$\{\psi_{mn},\ (m,n) \in \mathcal{I}\}$.

[7] With respect to the $L^2(\mathbb{R})$ norm.

5.3.1 Variational computation of wavelet coefficients based on training data

Although the problem of determining the wavelet coefficients in a finite approximation can be well formulated, we know of no analytic solution to the problem of explicitly computing the coefficients, given only (possibly irregularly spaced) samples of the function. We can however formulate the coefficient computation problem as a variational principle in a fashion analogous to learning algorithms such as backpropagation. We define our cost functional to be

$$E = \sum_{(x^i,y^i)\in\Theta} \|O^i - y^i\|^2 = \sum_{(x^i,y^i)\in\Theta} \mid \sum_{(m,n)\in\mathcal{I}} c_{mn}\psi_{mn}(x^i) - y^i|^2, \tag{20}$$

where O^i is the output of the network when x^i is the input as in Section 2.1. We choose the wavelet coefficients as those which minimize E. As a result of the wavelet formulation, the weights to be determined appear linearly in the output equation of the network. Thus E is a *convex* function of the coefficients $\{c_{mn}\}$ and therefore any minimizer $c^* = \{c^*_{mn}\}_{(m,n)\in\mathcal{I}}$ of E is a *global minimizer*. Simple iterative optimization algorithms such as gradient descent can be used to minimize E.

5.3.2 Normal Equations

There exists however an alternative formulation of the above optimization problem which provides a non-iterative solution. Minimization of E as defined in (20) defines a 'least squares'problem. Therefore solutions can be determined by solving the system of linear equations constructed via the first order optimality condition (which is both necessary and sufficient in this case) $\frac{\partial E}{\partial c^*_{kj}} = 0,\ (k,j) \in \mathcal{I}$ at any minimizer c^*. By choosing an ordering of the wavelet terms $\{\psi_{mn},\ (m,n) \in \mathcal{I}\}$ the normal equations can be written as

$$P\,C \;=\; W \tag{21}$$

where, P is the $\#(\mathcal{I}) \times \#(\mathcal{I})$ matrix defined by,

$$P = [P_{kj}] = [\sum_{(x^i,y^i)\in\Theta} \Psi_k(x^i)\Psi_j(x^i)], \tag{22}$$

and

$$W = [\sum_{(x^i,y^i)\in\Theta} \Psi_1(x^i)y^i, \ldots, \sum_{(x^i,y^i)\in\Theta} \Psi_{\#(\mathcal{I})}(x^i)y^i]^T. \tag{23}$$

and C is the coefficient vector which needs to be solved for. Typically solutions of (21) will not be unique and stabilizing methods such as use of the generalized inverse, $P^\dagger = (P^*P)^{-1}P^*$ must be applied.

Remark

Given a frame $\{\psi_{mn}\}$, and $f \in \mathrm{L}^2(\mathbb{R})$ let $c(f)$ be the vector in l^2defined by the wavelet expansion coefficients $\{< f, S^{-1}\psi_{mn} >\}$ of f. From Theorem 3.1 (6), it is clear that if the wavelet expansion of $f \in \mathrm{L}^2(\mathbb{R})$ is not unique, then all sequences $a(f)$ in $l^2(\mathbb{Z}^2)$ of wavelet expansion coefficients of f must be such that $\|a(f)\|^2 = \|c(f)\|^2 + \|c(f) - a(f)\|^2$. Therefore $c(f)$

is an optimal sequence of expansion coefficients in the sense of being minimum (l^2) norm. It can easily be shown that any finite number of vectors form a frame for their span (c.f. [22]). It is also well known that use of the generalized inverse, $P^\dagger$, of P results in the minimum l^2 norm solution. Thus the generalized inverse $P^\dagger$ is a sensible choice for use in solving (21).

5.4 Simulations

As a test of the neural network synthesis procedure described above, we simulated a few simple examples. As a first test we chose the bandlimited function comprised of two sinusoids at different frequencies, specifically $f(x) = sin(2\pi 5x) + sin(2\pi 10x)$ which is shown in Figure 13. Taking $x_{\text{min}} = 0.0$ and $x_{\text{max}} = 0.3$, 50 randomly spaced samples of the function were included in the training set Θ. A single dilation of the mother wavelet was chosen ($n = 6$) which covered the frequency range adequately (see Figure 14). Translations[8] of this dilation of ψ which contributed significantly in the interval $[x_{\text{min}}, x_{\text{max}}]$ were used, resulting in 40 hidden units. Applying a simple gradient descent scheme to minimize E, an approximation to f was obtained. The resulting approximation is shown in Figure 13 along with the original function.

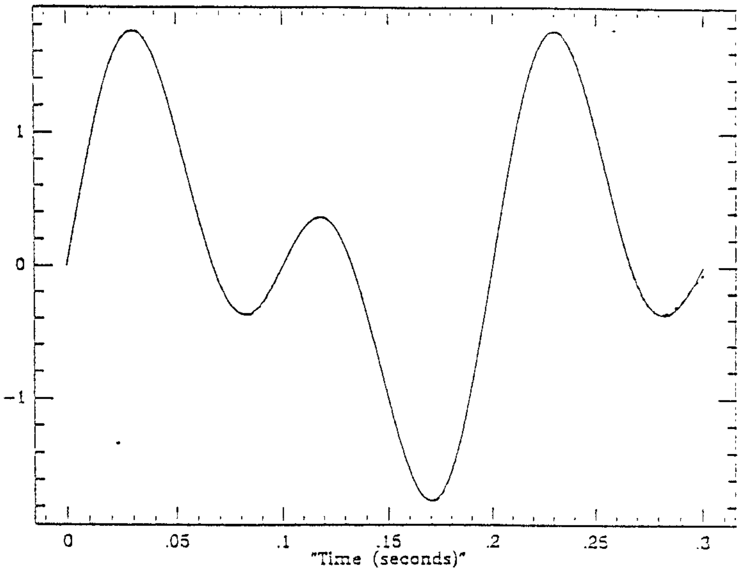

Figure 13: *Original bandlimited function* $f(x) = \sin(2\pi 5x) + \sin(2\pi 10x)$ *and finite wavelet approximation (dashed line).*

A second, slightly more complicated, example was simulated by first generating a random spectrum (Figure 15) which is concentrated in frequency and then sampling the corresponding function in the time domain. The result of this simulation using again just one dilation of the mother wavelet is shown in Figure 16.

[8] These translations were integer multiples of the translation stepsize b.

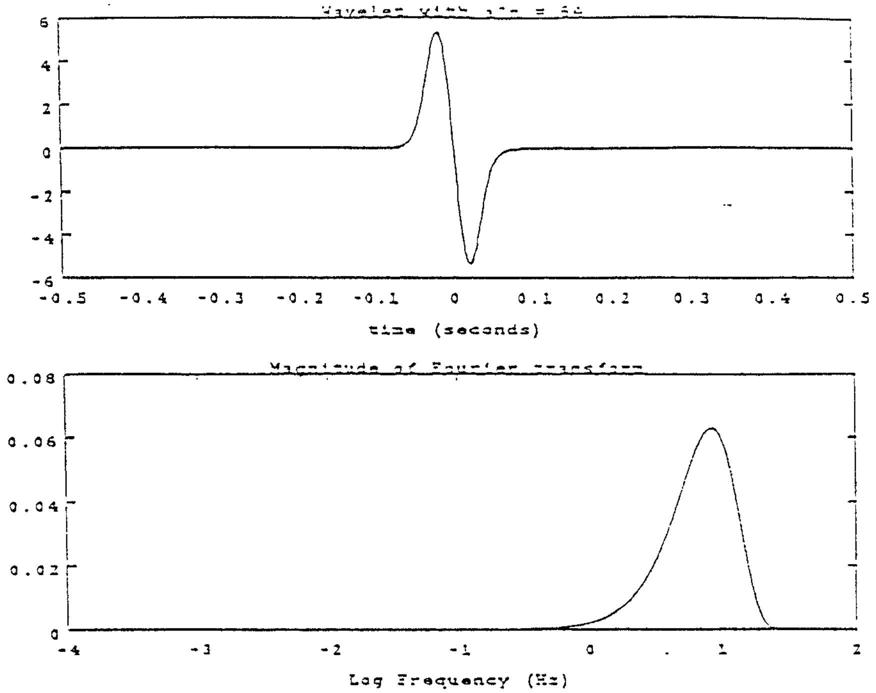

Figure 14: *Wavelet at dilation $n = 6$ and magnitude of Fourier transform.*

6 Conclusions and Discussion

We have demonstrated that it is possible to construct a theoretical description of feedforward neural networks in terms of wavelet decompositions. This description follows naturally from the inherent translation and dilation structure of such networks. The wavelet description of feedforward networks easily characterizes the class of mappings which can be implemented in such architectures. Although such characterizations have been previously provided in a number of different forms [2, 1, 10], to our knowledge, no previous characterization using sigmoidal activation functions is capable of defining the exact network implementation of a given function. What is distinctly different about the wavelet viewpoint is that it provides an extremely flexible (not necessarily orthogonal) transform formalism. This flexibility has been utilized in this paper to construct a transform based upon combinations of sigmoids. We would like to point out that in general there is nothing special about sigmoidal functions and that a variety of different activation functions, including e.g. orthogonal wavelets can be of significant interest. Sigmoidal functions however hold one attraction; such functions can be easily implemented in analog integrated circuitry (see e.g.[19]). Aside from this, we have chosen to work with sigmoidal functions only to demonstrate the general methodology that can be applied in the context of feedforward neural networks.

In addition to providing a theoretical framework within which to perform analysis of feedforward networks, the wavelet formalism supplies a tool which can be used to incorporate spatio-spectral information contained in the training data in structuring of the network. Two possible schemes to perform this task were described in Section 5. Minimality in terms of the

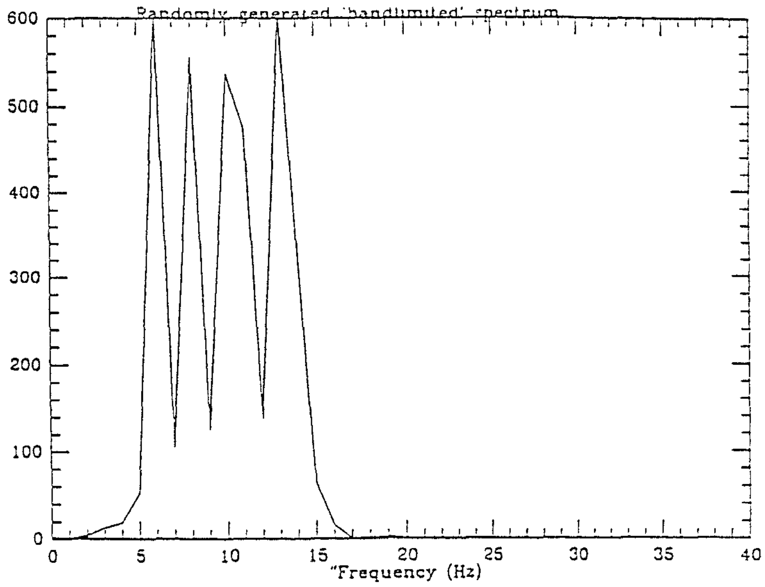

Figure 15: *Random spectrum*

number of nodes in the network cannot be guaranteed using these methods[9]. However, it is possible to estimate the approximation error ([6]) in terms of the signal energy lying outside the chosen spatio-spectral region.

In this paper, attention has been primarily restricted to approximating functions in $L^2(\mathbb{R})$. Most applications where neural networks are particularly useful involve mappings in higher dimensional domains (e.g. in vision, robot motion control, etc). Although extensions of the methods of this paper to higher dimensions are possible (as described in Section 4.2), such extensions have the potential to be computationally expensive. We are currently studying the formulation of more computationally viable synthesis techniques for approximation of higher dimensional mappings using feedforward neural networks.

Using the wavelet formalism to synthesize networks results in a greatly simplified training problem. Unlike the situation in traditional feedforward neural network constructions, the cost functional is convex and thereby admits *global* minimizing solutions only. Convexity of the cost functional is a result of fixing the weights in the arguments of the nonlinearities so as to provide the required dilations and translations. Simple iterative solutions to this problem such as gradient descent are thus justifiable and are not in danger of being trapped in local minima.

[9] This problem of large networks is particularly limiting when considering mappings in higher dimensions

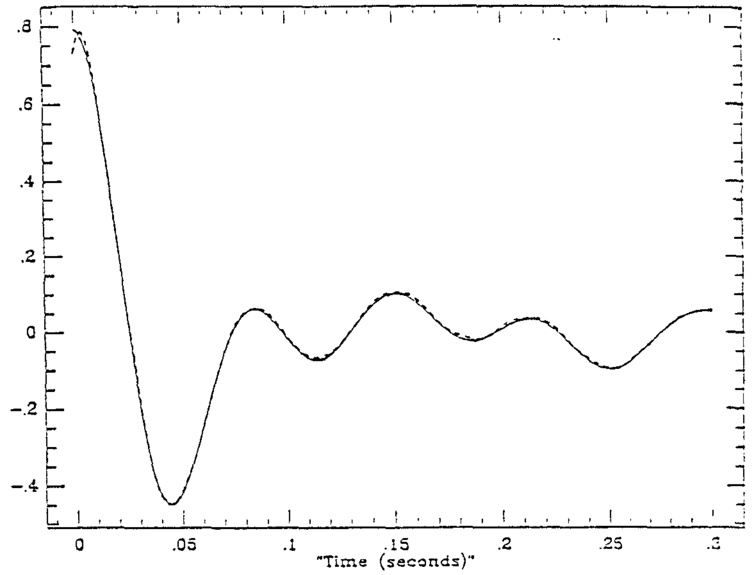

Figure 16: *Frequency-concentrated signal corresponding to random spectrum in Figure 17 and finite wavelet approximation (dashed line).*

Appendix

A Determining Translation and Dilation Stepsizes

Given an admissible mother wavelet $g \in \mathrm{L}^2(\mathbb{R})$, the following theorem by Daubechies [6] can be used to numerically determine values of the parameters a and b for which (g, a, b) generates an affine frame for $\mathrm{L}^2(\mathbb{R})$.

Theorem A.1 (Daubechies[6]) *Let $g \in \mathrm{L}^2(\mathbb{R})$ and $a > 1$ be such that:*

(1)

$$m(g; a) = \operatorname{ess} \inf_{|\omega| \in [1,a]} \sum_n |\widehat{g}(a^n \omega)|^2 > 0 \tag{24}$$

(2)

$$M(g; a) = \operatorname{ess} \sup_{|\omega| \in [1,a]} \sum_n |\widehat{g}(a^n \omega)|^2 < \infty \tag{25}$$

(3)

$$\lim_{b \to 0} 2 \sum_{k=1}^{\infty} \beta(k/b)^{1/2} \beta(-k/b)^{1/2} = 0, \tag{26}$$

where

$$\beta(s) = \operatorname{ess} \sup_{|\omega| \in [1,a]} \sum_n |\widehat{g}(a^n \omega)| |\widehat{g}(a^n \omega - s)|.$$

Then there exists $B_c > 0$ such that (g, a, b) generates an affine frame for each $0 < b < B_c$.

Proof of the following corollary, can also be found in [6].

Corollary A.1 *If $g \in L^2(\mathbb{R})$ and $a > 1$ satisfy the hypotheses of Theorem A.1 then,*

$$B_c \geq b_c = \inf\{b|\ m(g;a) - 2\sum_{k=1}^{\infty} \beta(k/b)^{1/2}\beta(-k/b)^{1/2} \leq 0\} \tag{27}$$

and for $0 < b < b_c$, the frame bounds A and B can be estimated as,

$$\begin{aligned} A &\geq b^{-1}(m(g;a) - 2\sum_{k=1}^{\infty}\beta(k/b)^{1/2}\beta(-k/b)^{1/2}) \\ B &\leq b^{-1}(M(g;a) + 2\sum_{k=1}^{\infty}\beta(k/b)^{1/2}\beta(-k/b)^{1/2}) \end{aligned} \tag{28}$$

A.1 Dilation and Translation Stepsizes for the Wavelet ψ Constructed From Sigmoids

For the task of constructing an affine frame based on the mother wavelet candidate ψ of Section 4.1 with dilation stepsize $a = 2$, we can check conditions (24) and (25) numerically. Figure 17 shows a plot of the sum in (24) using the mother wavelet candidate ψ with dilation step size a=2. From the plot in Figure 17 the minimum value of the sum in (24) is approximately

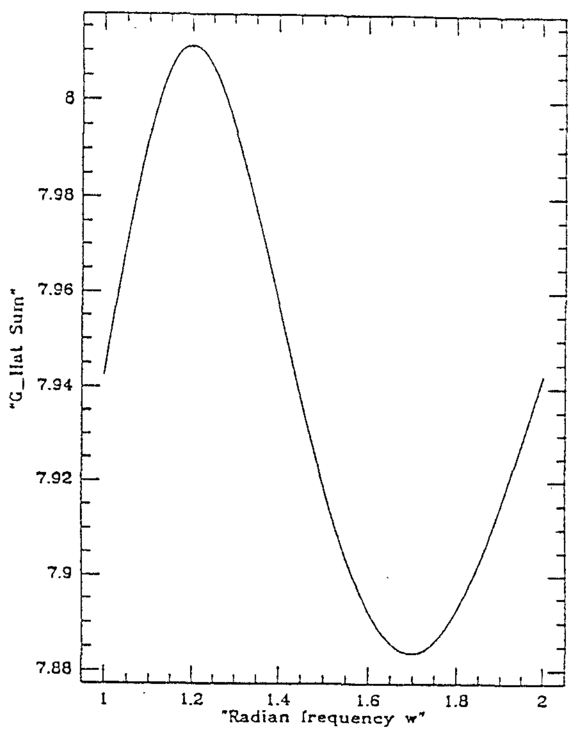

Figure 17: *Plot of $\sum_n |\widehat{g}(a^n\omega)|^2$ for $\omega \in [1,a] = [1,2]$*

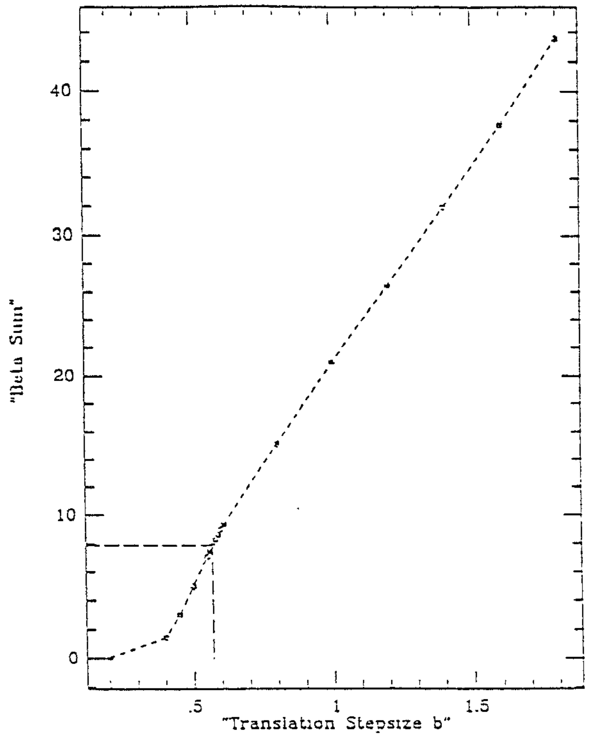

Figure 18: *Plot of* $2\sum_{k} \neq 0\, \beta(k/b)^{1/2}\beta(-k/b)^{1/2}$, *for various values of* b

$m(g;2)$ =7.88, and the maximum value is $M(g;2)$ =8.01. Figure 18 is a plot showing the value of $2\sum_{k\neq 0}\beta(k/b)^{1/2}\beta(-k/b)^{1/2}$, for various values of b. From this, we see that for $a = 2$ and $0 < b \leq 0.57$, (ψ, a, b) generates an affine frame for $L^2(\mathbb{R})$.

Remark

The conditions in Theorem A.1 and subsequently those in Corollary A.1, are in general very conservative since the theorem relies on the *Cauchy-Schwartz inequality* to establish bounds. Therefore although it may be possible to determine values of a and b for which (g, a, b) generates an affine frame, for a given mother wavelet g, it is almost certain that these are not the 'best' possible values of a and b which can be used with g to obtain a frame. For very small values of a and b, a large number of frame elements will be required to 'cover' any given time interval and frequency band. Thus it is desirable to use the largest values of a and b for which (g, a, b) generates a frame. That is, we would like the frame elements to be as sparsely distributed in joint time-frequency space as possible.

Acknowledgements

The authors are grateful to Professor Hans Feichtinger of the University of Vienna for many helpful discussions and numerous suggestions regarding this paper and to Dr. Jim Gillis for discussions on network synthesis techniques. We also wish to thank Professor Halbert White of the University of California, San Diego for helpful comments, and Professor John Benedetto of the University of Maryland, College Park for discussions and the many references he provided us on the subject of wavelet transforms.

References

[1] G. Cybenko. Continuous valued neural networks with two hidden layers are sufficient. Technical report, Department of Computer Science, Tufts University, Medford, MA, March 1988.

[2] G. Cybenko. Approximations by superpositions of a sigmoidal function. Technical Report CSRD 856, Center for Supercomputing Research and Development, University of Illinois, Urbana, February 1989.

[3] I. Daubechies, Grossmann A., and Y. Meyer. Painless nonorthogonal expansions. *Journal of Mathematical Physics*, 27(5):1271–1283, May 1986.

[4] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41:909–996, 1988.

[5] I. Daubechies. Time-frequency localization operators: A geometric phase space approach. *IEEE Transactions on Information Theory*, 34(4):605–612, July 1988.

[6] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5), September 1990.

[7] J. Daugman. Six formal properties of two-dimensional anisotropic visual filters: Structural principles and frequency/orientation selectivity. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):882–887, September/October 1983.

[8] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[9] C. E. Heil and D. F. Walnut. Continuous and discrete wavelet transforms. *SIAM Review*, 31(4):628–666, December 1989.

[10] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[11] K. Hornik, M. Stinchcombe, and H. White. Universal approximations of an unknown mapping and its derivatives using multilayer feedforward networks. Preprint, January 1990.

[12] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound patterns through wavelet transforms. *Internat. J. Pattern Recog. Artif. Int.*, 1:273–302, 1987.

[13] M. Kuperstein. Generalized neural model for adaptive sensory-motor control of single postures. In *Proceedings IEEE Inter. Conf. On Robotics and Automation*, pages 134–139, Philadelphia, Pa., April 1988.

[14] M. Kuperstein and Jyhpyng Wang. Neural controller for adaptive movements with unforseen payloads. *IEEE Trans. on Neural Networks*, 1(1):137–142, March 1990.

[15] H. C. Lueng and V. W. Zue. Applications of error backpropagation to phonetic classification. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems*, pages 206–231. Morgan Kaufman Publisher, 1989.

[16] S. G. Mallat. Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions On Acoustics Speech and Signal Processing*, 37(12):2091–2110, December 1989.

[17] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[18] S. G. Mallat and W. Hwang. Singularity detection and processing with wavelets. Preprint.

[19] C. Mead. *Analog VLSI and Neural Systems.* Addison Wesley, 1989.

[20] K. S. Narendra and K. Parthasarathy. Identification and control of dynamical systems using neural networks. *IEEE Trans. on Neural Networks*, 1(1):4–27, March 1990.

[21] Y. C. Pati and P. S. Krishnaprasad. Discrete affine wavelet transforms for analysis and synthesis feedforward neural networks. In R. Lippman, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems III*, pages 743–749, San Mateo, CA, 1990. Morgan Kaufmann, Publishers.

[22] Y. C. Pati. Frames generated by subspace addition. Technical Report SRC TR 91-55, University of Maryland, Systems Research Center, 1991.

[23] M. Porat and Y. Y. Zeevi. The generalized Gabor scheme of image representation in biological and machine vision. *IEEE Transactions on Pattern Analysis and Machine Intellegence*, 10(4):452–468, July 1988.

[24] M. Stinchcombe and H. White. Universal approximations using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings Inter. Joint Conf. on Neural Networks (IJCNN)*, pages 613–617, Washington D.C, 1989.