Adaptive Time-Frequency Classification of Acoustic Backscatter

Brian A. Telfer, Harold H. Szu
Naval Surface Warfare Center – Dahlgren Division, Code B44
Silver Spring, MD 20903

Gerald J. Dobeck
Naval Surface Warfare Center – Dahlgren Division
Coastal Systems Station
Panama City, FL 32407

Abstract

An adaptive time-frequency classifier algorithm is detailed and tested on a data set of acoustic backscatter from a metallic manmade object and natural clutter with synthetic reverberation noise. The algorithm is improved over previous versions in that it operates directly on time signals rather than their wavelet transforms, and in that the features measure time-frequency energy and are insensitive to phase differences (due to signal variations). The time-frequency features are initialized by selecting wavelet transform coefficients with the highest Fisher ratios. Similarities and differences of this classification algorithm with matching pursuit, a representation algorithm, are discussed. This is done to show the need to treat classification differently than representation for realistic cases where class boundaries significantly overlap. The adaptive classifier is shown to have significantly higher detection rates than Fourier transform and wavelet transform methods, as compared by receiver operating characteristics.

1 Introduction

The ability of dolphins to recognize underwater objects with their active sonar is well known [1], and provides an existence proof for man-made active sonar classifiers. The goal of our work is to improve the abilities of these classifiers. Power spectral features have commonly been chosen as a feature space for this task [2, 1]. Power spectral features provide only frequency information. Features that are time-frequency atoms provide additional information of how frequency content changes over time, and thus would be expected to lead to higher classification rates for nonstationary applications. We demonstrate this with a case study.

A number of methods have been developed for representing (as opposed to classifying) signals based on adaptively selecting a small number of time-frequency atoms (i.e., regions used as building blocks) [3, 4, 5, 6, 7, 8, 9]. Matching pursuit [9], in particular, embodies attractive properties that carry over to classification. In matching pursuit, the idea is to represent a signal as an expansion of time-frequency atoms that are selected from a large and redundant dictionary. Choosing from a redundant set allows better matching of a signal's local structure with a smaller number of terms than would be required if selecting atoms from a single basis. An expansion in single basis, whether Fourier, wavelet or other, is not flexible enough to represent all signal structures. Matching pursuit selects time-frequency atoms by processing the dictionary parameters of frequency, window width, and time shift, rather than explicitly operating on a time-frequency representation. This allows significant savings in computation and memory.

For acoustic classification, we also desire features that are time-frequency atoms that are drawn from a redundant dictionary. However, the goals of classification are quite different from representation. A time-frequency atom that best represents a particular signal may contain little information for discriminating

0-8194-1844-7/95/$6.00

between two classes of signals. We have developed two earlier algorithms for adaptively computing time-frequency features [10, 11], and a related adaptive algorithm for detecting objects in images has also been developed [12]. For the classification applications we consider, we desire an adaptive algorithm that

1. operates on the dictionary parameters rather than explicitly on a time-frequency representation (faster and less memory needed in training),
2. computes features that respond to time-frequency energies, but are insensitive to phase (because irrelevant phase differences appear across multiple training signals from the same class),
3. computes time-frequency atoms that can be expressed as time-domain filters rather than as nonlinear combinations of numerous time-frequency coefficients (faster classification once trained).

The adaptive classifier described in [10] incorporated goals 1 and 3, but not 2. Another classifier incorporated goal 2, but not 1 and 3 [11]. The classifier described in this paper is the first to meet all three goals. The algorithm is described in Section 2, the data base in Section 3, and the results in Section 4.

2 Adaptive Wavelet Classifier Algorithm

By way of comparison, we first present an outline of the matching pursuit algorithm for signal representation. It is a greedy algorithm that sequentially finds time-frequency atoms that approximate a signal $s(t)$ by

$$\hat{s}(t) = \sum_i c_i g_i(t), \tag{1}$$

where $g_i(t)$ is a time-frequency atom drawn from a dictionary, which, e.g., consists of all possible Gabor functions at different time shifts, frequencies and window widths, and c_i is a coefficient. The goal is to find these parameters to minimize the approximation error $\| s(t) - \hat{s}(t) \|$. The algorithm proceeds conceptually according to the following steps:

1. Set $i = 0$ and the residual signal $R_0(t)$ to $s(t)$
2. For atom g_i, find initial approximate parameter values by searching small subdictionary,
3. Optimize the initial parameters with gradient descent,
4. Update the signal residue by $R_{i+1}(t) = R_i(t) - c_i g_i(t)$,
5. If desired representation accuracy has not been achieved, set $i = i + 1$ and go to step 2.
6. The approximation can be further refined by optimizing all weights with a back-projection algorithm.

This outline contains the general algorithm flow, but omits a number of the details that make the matching pursuit algorithm computationally attractive for representation (such as not explicitly computing the residual signal).

Now consider a two-class classifier formulated as

$$v = \sigma \left(\sum_i c_i \| \sum_t g_i(t) s(t) \| \right), \tag{2}$$

where v is the classifier output, σ is a sigmoidal function $\sigma(z) = 1/[1+\exp(-z)]$, and $g_i(t)$ is a Morlet wavelet (or equivalently, Gabor function) with any time shift, frequency and window width, given by

$$g_i(t) = exp\left[j\left(\frac{t-b_i}{a_i}\right) - 0.5\frac{(t-b_i)^2}{w_i^2 a_i^2}\right]. \tag{3}$$

The purpose of computing the magnitude term in Eq. 2 is to make the classifier insensitive to phase misalignments in the training signals and the signals to be classified. It is the time-frequency magnitudes that contain the discriminatory information in our applications, not phase differences. The classifier is formulated as having a single output that is a linear combination of magnitudes for simplicity and because that is sufficient for the data that we consider. It is straightforward to extend this formulation to a multiple layer feedforward network with multiple outputs, if that is needed.

The goal of training the classifier is to minimize the error

$$E = \frac{1}{2}\sum_{n=1}^{N}(d_n - v_n)^2 \tag{4}$$

where d_n and v_n are the desired and actual outputs for the n-th training signal $s_n(t)$, by optimizing c_i, a_i, b_i, and w_i in Eqs. 2 and 3.

A greedy training algorithm for the classifier can be formulated that is analogous in some respects to matching pursuit:

1. Set $i = 0$,
2. For atom g_i, find initial parameter values by evaluating small subdictionary with suitable measure of discriminatory information, e.g., Fisher ratios [13] of wavelet transform coefficients,
3. Optimize the initial parameters with gradient descent, with other atoms and weights fixed,
4. If desired classification accuracy has not been achieved, set $i = i + 1$ and go to step 2,
5. The classifier can be further refined by optimizing all weights together using gradient descent.

Like matching pursuit, this is a greedy algorithm, finds initial parameter values for each atom and optimizes them, and has a final optimization step for all parameters. It does not fit into the class of pursuit algorithms because it does not remove the effect of an atom on the training data before finding a new atom. It uses an evaluation function in step 2 that is suitable for classification, not representation. The algorithm is a type of constructive algorithm, which have received significant attention in the neural network community in the last several years.

In this paper, a greedy algorithm was not needed for particular data considered, which only requires a few atoms (13 were used). A greedy algorithm could be utilized more effectively in a more complex classification problem, and the outline provided above also serves to illustrate the similarities and differences between the representation and classification tasks. Instead, the number of atoms and their initial parameters were chosen by examining the Fisher ratios of the wavelet transform coefficients. The parameters were then optimized over a training set of data using a conjugate gradient method and a simple line search.

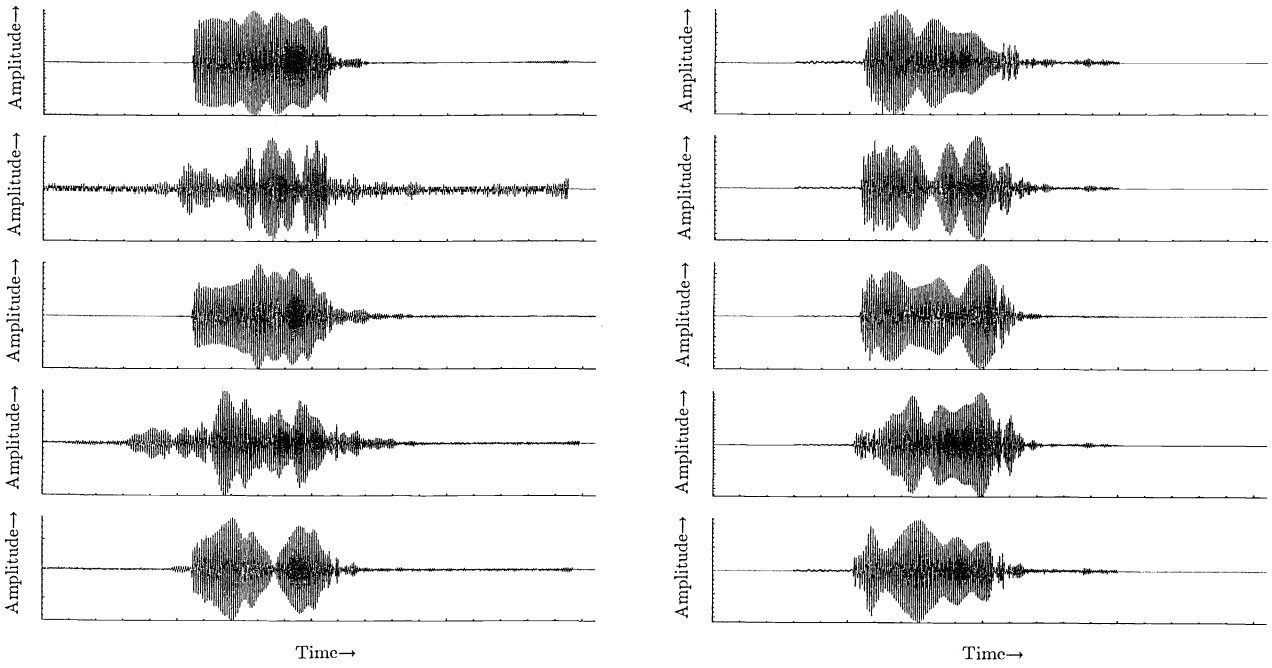

Figure 1: Acoustic returns from metallic object (left) and natural clutter (right) at different aspect angles.

3 Data Description

The goal is to discriminate between sonar returns from metallic objects and similar natural clutter. The data base is based on that described in [11], but modified extensively by more thorough signal alignment, by eliminating low-powered returns that could not be realistically detected, and by adopting a more realistic data normalization scheme. A summary is provided here.

Acoustic returns were collected from a metallic object at different aspect angles and from natural clutter, using a linear-FM transmit signal. The synthetic clutter described in [11] is not considered in this paper. Figure 1 shows representative returns of the metallic object and natural clutter. Note that the different signal classes have similar strengths and durations, and that there is significant intraclass variation. These characteristics make the data difficult to classify. In addition, 15dB synthetic reverberation noise was also added to the data (not shown in figure).

The data base considered here consists of 173 returns from the metallic object and 216 from the natural clutter. These were divided into 193 training samples (multiple aspects and multiple days of collection) and 196 test samples at aspects not included in the training set.

4 Classification Results

Three classifiers were tested. The first classified 41 Fourier transform (FT) magnitude features with a feedforward neural network with two layers of weights. The second classified 9600 wavelet transform (WT) magnitudes (16 scale $\times$ 600 time samples) with a neural network having a single layer of weights. These two classifiers serve as benchmarks.

The third classifier was the adaptive wavelet classifier. The initial parameter values and the number of atoms were selected from the Fisher ratios of the wavelet transform magnitude coefficients. The continuous wavelet transform (discretely sampled) was computed by

$$W(a,b) = \int g^* \left(\frac{t-b}{a} \right) s_n(t)\, dt/a, \tag{5}$$

with a complex Morlet wavelet given by

$$g(t) = \exp(-t^2/2 + j5t). \tag{6}$$

The Fisher ratio [13] is

$$f = \frac{(m_1^2 - m_2^2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{7}$$

where m_i and σ_i^2 are the mean and variance of a feature for class i. The Fisher ratio is a measure of class separability for a particular feature that is most applicable to unimodal distributions. For unimodal distributions, a large separation of class means and small variances leads to a large Fisher ratio. For multimodal distributions, the ratio is not necessarily a good indication of class separability. For example, consider one class with a bimodal distribution that straddles a central unimodal distibution for the other class, such that the class means are identical and that the classes can perfectly separated. In this case the Fisher ratio is 0. Although not an ideal indicator, the Fisher ratio provides easily computable clues as to what areas of time-frequency space provide good discrimination, which can serve as initial values for adaptation.

Figure 2a and b display the class means for the wavelet transform magnitudes of the metallic and natural returns. They exhibit the chirp of the transmit signal. The intraclass variation is considerable. The Fisher ratios are shown in Fig. 2c. The highest Fisher ratios are at the beginning and end of the average signals, implying a class difference in attack and decay or time extent. The initial a and b parameters for the adaptive classifier were selected to match local maxima of the Fisher ratios. The window width parameters w were all initially set to a constant (5). The number of atoms, 13, was selected by choosing only Fisher ratios that were larger than the largest Fisher ratio that was due to only reverb noise. The 13 highest Fisher ratios ranged from 0.19 to 0.037. These values are minuscule, either because of multimodal distributions or very little discriminatory content. For example, a feature with Gaussian distributions for the two classes, with a mean separation of 1 and standard deviations of 2, would have a Fisher ratio of 0.125 and a very high misclassification probability. Yet, the Fisher ratio does seem to pick out interesting initial values that result in good final classification performance.

The adaptive algorithm proceeded by updating the time-frequency atoms from their initial values using gradient descent. It was found necessary to update only a single type of parameter at a time, since the gradients of each parameter type had different magnitudes. Thus, the coefficients c_i were updated for 10 iterations with the other parameters fixed, then the a_i, b_i and w_i parameters were each updated separately for ten iterations in turn. The cycle was repeated until the algorithm was stopped. This was done for

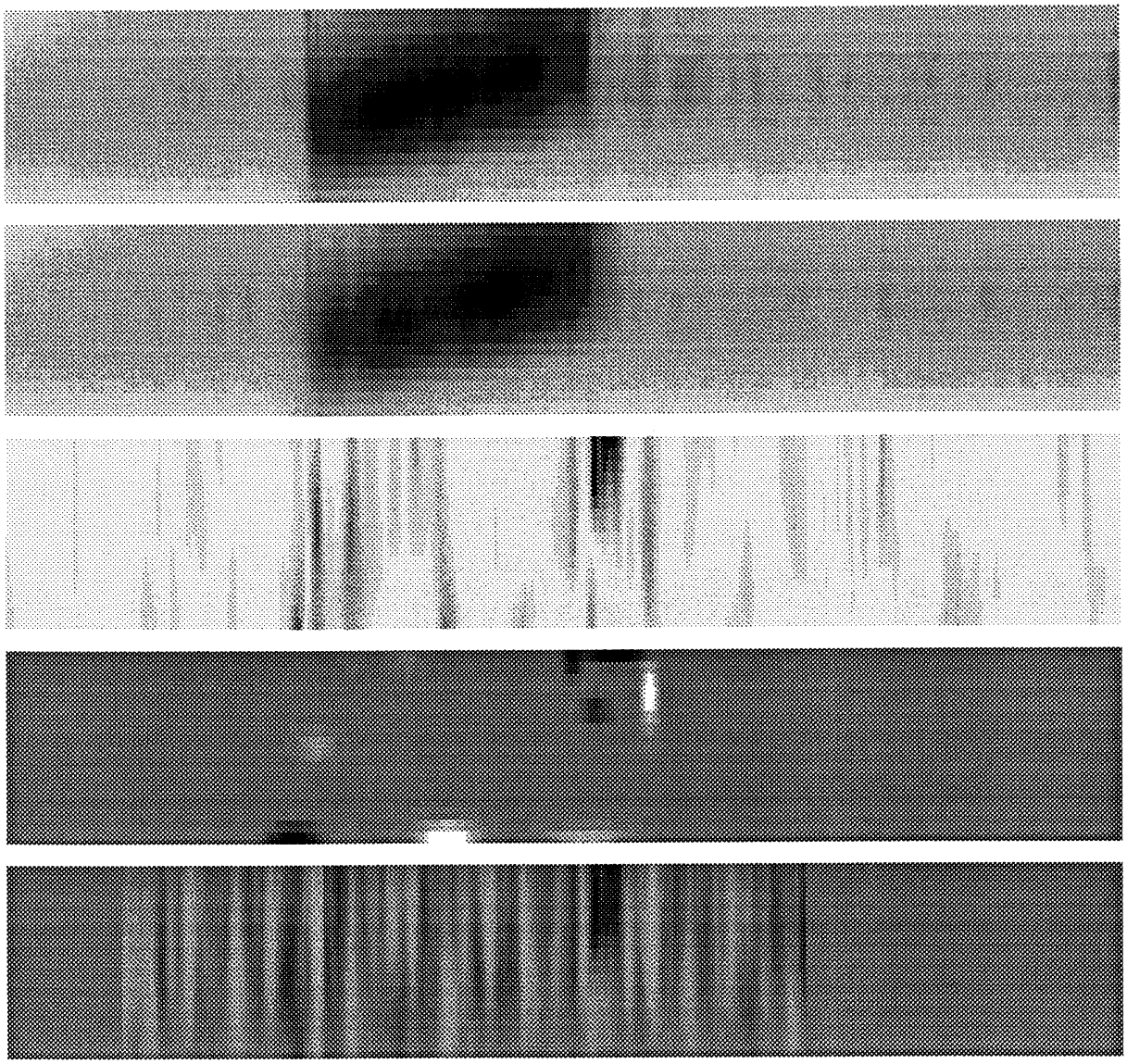

Figure 2: Time-frequency displays with time as horizontal axis and frequency (inverse scale) as vertical. From top: a) wavelet transform magnitudes averaged over all metallic object returns (black is largest magnitude), b) wavelet transform magnitudes averaged over all natural object returns (black is largest magnitude), c) Fisher ratios of WT magnitude coefficients (black is largest magnitude), d) weighted time-frequency features computed by adaptive 2 classifier (white is positive, black is negative) that overlay displays in a-c, d) weights of WT classifier (white is positive, black is negative).

50 iterations total. (The coefficients c were trained for a second cycle.) The FT and WT classifiers were each run for 25 iterations. The stopping point was selected according to the minimum test set error that occurred during training. We believe this is a reasonable way to determine a stopping point during algorithm development. Others have also used this method for comparing performance [14].

Training and test set classification errors and number of trainable parameters are given in Table 1. The two adaptive classifiers perform better than the benchmarks. The difference between the adaptive 1 and FT classifiers is significant at a 90% level. (The 95% confidence level for an error of 0.24 is ± 0.06, and for an error of 0.35 is ± 0.067 [15].) The "adaptive 1" classifier resulted in atoms with center frequencies higher and lower than those used for the FT and WT classifiers. (All atom center frequencies were initialized to be within bandwidth used in FT and WT classifiers, but some moved outside this bandwidth during training.) These frequencies had been excluded from the other classifiers because there was little energy at these frequencies, and it was uncertain whether class differences at these frequencies were due to artifacts. To be fair, the 10 of 13 atoms at these frequencies were reset to have center frequencies at the maximum and minimum values used in the benchmark classifiers. The adaptive classifier was then further trained with all frequency values fixed. This resulted in the "adaptive 2" classifier. That the adaptive 1 classifier found higher and lower frequency atoms that resulted in lower classification error reflects well on the algorithm. However, the adaptive 2 classifier provides for a better comparison with the benchmarks. The FT classifier consisted of two layers of weights and 4 hidden units. A linear FT classifier with 41 weights was also tried, resulting in a classification error of 0.38. A small number of trainable parameters does not guarantee better performance, as can be seen from the poor performance of the FT linear classifier with 41 weights. The excellent performance of the adaptive classifier is due to both the small number of trainable parameters and the ability to pick out discriminatory time-frequency features.

The weights and time-frequency features of the adaptive 2 classifier are compared with those from the WT magnitude classifier in Fig. 2d and e. Both classifiers tend to strongly weight time-frequency regions with high Fisher ratios (Fig. 2c). However, most of the weights in the WT magnitude classifier are extraneous and lead to poor generalization to the test data. Eight time-frequency features are weighted strongly enough to be visible in Fig. 2d. Most are readily interpretable from Figs. 2a-c, and the fact that the metallic and natural returns were specified to have desired outputs of 1 and 0, respectively. The right-most positive high-frequency atom overlies a time-frequency region where the average metallic return is slightly higher than the average natural return. Hence the positive weighting. The right-most negative high-frequency returns correspond to the decay region where the natural returns tend to be higher than the metallic. This is due to a different time extent of the two classes, or a slower decay for the natural returns. The more faint positive mid-frequency atom at the left corresponds to the attack region, where the metallic returns seem to have a sharper onset. The three visible low-frequency returns are harder to interpret, although they correspond to local maxima in the Fisher ratio. They appear strongly weighted in Fig. 2, but this is offset by the time-frequency magnitudes being small in those regions. It is interesting to note that none of the features occur in time-frequency regions in the main part of the returns, where a representation algorithm would place atoms. Also, note that the varying window widths of the atoms in Fig. 2d indicate that the atoms could not have been extracted from a single wavelet or windowed-Fourier basis, demonstrating the importance of drawing from a redundant dictionary.

The receiver operating characteristics (ROCs) shown in Fig. 3 were generated by varying the output threshold of each classifier. The adaptive classifier performs much better than the benchmarks.

A key advantage of the adaptive classifier is its ability to find highly discriminatory time-frequency atoms. This was investigated in two ways. First, the adaptive classifier was retrained from the initial values for 50

Classifier	Error	# Param
FT	0.35	168
WT	0.38	9600
Adaptive 1	0.24	52
Adaptive 2	0.29	52

Table 1: Test set error rates and number of trainable parameters.

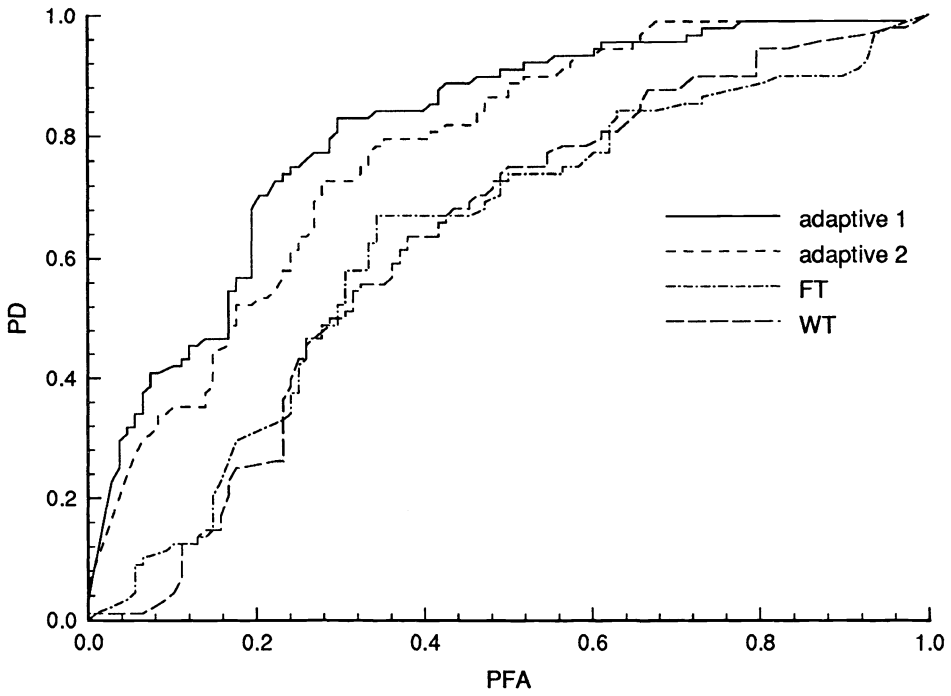

Figure 3: Receiver operating characteristics for the four classifiers, showing probability of detection vs. probability of false alarm.

iterations as before, but only allowing the combination coefficients c_i to change. This gave a best test set error of 0.265, only slightly worse than the error of 0.24 when the algorithm was allowed to vary the time and frequency of the atoms. However, tests on other data have shown a much for significant difference (e.g., 0.12 vs. 0.21). Second, the time-frequency parameters changed significantly during training: the largest change in a was 62% (from 1.81 to 0.68 in the adaptive 1 classifier), in n was 32% (from 5.0 to 6.6), and in b was 17% of the separation between two b values (from 2114 to 2103, with an adjacent initial b value of 2050).

5 Summary

A classification algorithm has been demonstrated that adaptively finds and weights time-frequency features. Attractive properties of the algorithm are that it: operates on the time-frequency parameters (time and frequency centers, window width) rather than explicitly on time-frequency representations, computes features that respond to time-frequency energies and is insensitive to phase, and computes features that can be expressed as short time-domain filters. For an acoustic backscatter application, the Fisher ratio of wavelet transform magnitudes was demonstrated to be an easily computable effective way to initialize the time-frequency features. The classifier was demonstrated to significantly outperform Fourier and wavelet benchmark classifiers in terms of receiver operating characteristics, and only utilized 13 time-frequency features.

Acknowledgements

This work was supported the Office of Naval Research, Code 32, under the Sensor Signal/Image Processing Project (RN15C83A) within the Mine Reconn/Hunter Program, and by the NSWCDD Independent Research Program.

References

[1] P. Moore, H. Roitblat, R. Penner, and P. Nachtigall, "Recognizing successive dolphin echoes with an integrator gateway network," *Neural Networks*, vol. 4, pp. 701–709, 1991.

[2] R. Gorman, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, pp. 75–89, 1988.

[3] J. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1169–1179, July 1988.

[4] Q. Zhang and A. Benveniste, "Approximation by nonlinear wavelet networks," *Proc. International Conf. Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3417–3420, May 1991.

[5] S. Mann and S. Haykin, "Adaptive 'chirplet' transform: An adaptive generalization of the wavelet transform," *Optical Engineering*, vol. 31, pp. 1243–1256, June 1992.

[6] B. Bakshi and G. Stephanopoulos, "Wavelets as basis functions for localized learning in a multi-resolution hierarchy," *Proc. International Joint Conf. Neural Networks - Baltimore*, vol. II, pp. 140–145, June 1992.

[7] Q. Zhang and A. Benveniste, "Wavelet networks," *IEEE Trans. Neural Networks*, vol. 3, pp. 889–898, November 1992.

[8] Y. Pati and P. Krishnaprasad, "Analysis and synthesis of feedforward neural networks using discrete affine wavelet transforms," *IEEE Trans. Neural Networks*, vol. 4, pp. 73–85, January 1993.

[9] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, December 1993.

[10] H. Szu, B. Telfer, and S. Kadambe, "Adaptive wavelets for signal representation and classification," *Optical Engineering*, vol. 31, pp. 1907–1916, September 1992.

[11] B. Telfer, H. Szu, G. Dobeck, J. Garcia, H. Ko, A. Dubey, and N. Witherspoon, "Adaptive wavelet classification of acoustic backscatter and imagery," *Optical Engineering*, vol. 33, pp. 2192–2203, July 1994.

[12] D. Casasent and J. Smokelin, "Neural net design of macro Gabor wavelet filters for distortion-invariant object detection in clutter," *Optical Engineering*, vol. 33, pp. 2264–2271, July 1994.

[13] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* New York: John Wiley and Sons, 1973.

[14] K. Lang, A. Waibel, and G. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural Networks*, vol. 3, pp. 23–43, 1990.

[15] W. Highleyman, "The design and analysis of pattern recognition experiments," *Bell System Technical Journal*, vol. 41, pp. 723–744, 1962.