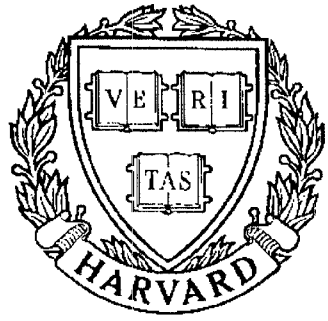
TECHNICAL RESEARCH REPORT

SYSTEMS RESEARCH CENTER



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

Classification of Transient Signals via Auditory Representations

by A. Teolis and S. Shamma

TR 91-99

Classification of Transient Signals via Auditory Representations

Anthony Teolis * Shihab Shamma †

Abstract

We use a model of processing in the human auditory system to develop robust representations of signals. These reduced representations are then presented to a neural network for training and classification.

Empirical studies demonstrate that auditory representations compare favorably to direct frequency (magnitude spectrum) representations with respect to classification performance (i.e. probabilities of detection and false alarm). For this comparison the Receiver Operating Characteristic (ROC) curves are generated from signals derived from the standard transient data set (STDS) distributed by DARPA/ONR.

1 Introduction

To a human subject the sound of a car door slamming or a finger snapping is easily recognizable. Humans demonstrate the ability to classify transient signals like these even in active (noisy) acoustic environments with little or no effort. Yet to an automaton or computer the task of transient recognition is far more formidable.

With the notion that the auditory system inherently represents acoustic signals in a robust manner, a promising approach to the goal of automatic transient classification seems to lie in the mimicking of that process. Fortunately the auditory system has been studied in detail. Recent research [9] has indicated an analytically appealing model of human auditory processing in its early stages.

In this paper we study two different approaches to the transient classification problem. First is a conventional approach based solely on spectrum analysis of the transient signals. That is signals are classified based on their respective spectral

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

†Electrical Engineering Department and Systems Research Center, and the Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742

content. Second is an approach based on the processing of acoustic signals in the auditory system. We look to the auditory system for signal representations which must in some sense be optimally designed by evolution.

In both approaches we maintain an identical classifier structure which we will refer to heretofore simply as the classifier system. The classifier system may be thought of simply as a cascade of two main components: (i) a transformation component and (ii) a neural associator component which takes as input the patterns generated in (i). The two approaches differ fundamentally only in their stage (i) transformations.

Both neural systems are trained on data derived from the standard transient data set (STDS). It is a compilation of transient signals of differing character in various noise environments (provided by DARPA/ONR) in order to standardize classification attempts. Basic training and testing sets are generated from this data.

This paper is organized as follows. In Section 2 we describe the STDS. An overview of the entire neural classification scheme is presented in section 3. Section 4 presents a detailed description of the auditory-like model on which the pattern transformations are based. Implementation details are covered in Section 5. A comparison of the auditory versus a spectral based classifier is presented in Section 6. Performance of the two classifier schemes is examined in Section 7. And finally Section 8 presents a discussion of the results.

2 The Standard Transient Data Set

Provided by DARPA, the standard transient data set consists of a compilation of six different transient signals as recorded in different (noisy) environments. Depicted in Figure 1 are examples of the six basic transient signals in a 40ms window. The sampling rate of this data is 25kHz. In addition the data set also contains files with pure ocean noise.

Table 1 reports where in time and frequency each of the basic six transient signals resides.

In the interest of (computer) time, we restrict ourselves to the first three of the basic transient signals. However, we believe that our approach can be applied to the entire set.

3 Classification Overview

Figure 2 depicts the basic classifier scheme to which we adhere. In the figure it can be seen that the classifier system consists of two main components. It

Signal	*Frequency (kHz)*	*Duration (ms)*
A	3	13
B	>3	8
C	3	10
D	3	100
E	0.1	1000-8000
F	0.15	1000-8000

Table 1: Transient Data

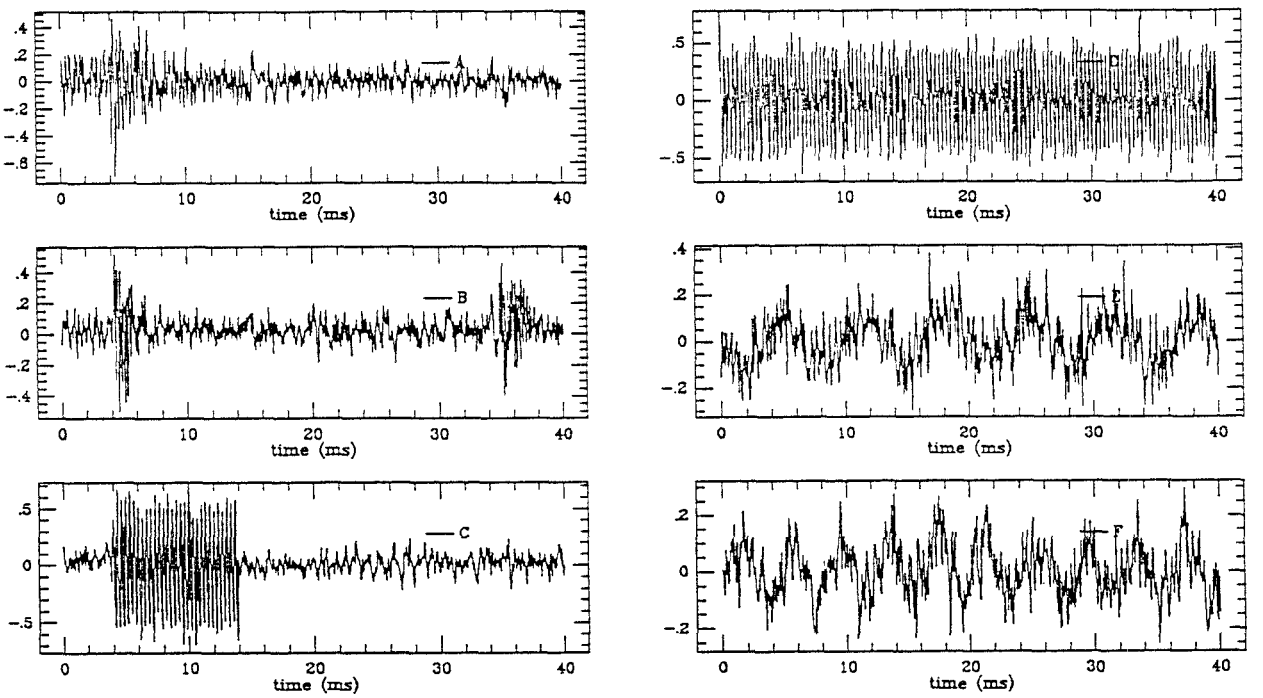


Figure 1: The six basic transient signals (in noise)

is the first component, the pattern transformation component, which takes the raw input data from the environment and transforms via sensory apparatus to an alternative representation. This representation is then presented to a the second component of the system for higher level processing. As depicted in the diagram we implement this high level processor as a feed forward neural network. Acting as a pattern associator, this neural network learns (during training) to perform the desired classification.

Being learned associations, all the classifications proposed here are required to have a training set of data on which to learn. All the neural learning implemented here is accomplished by standard gradient descent learning methods (i.e. backpropagation). For a complete reference the reader may consult [6].

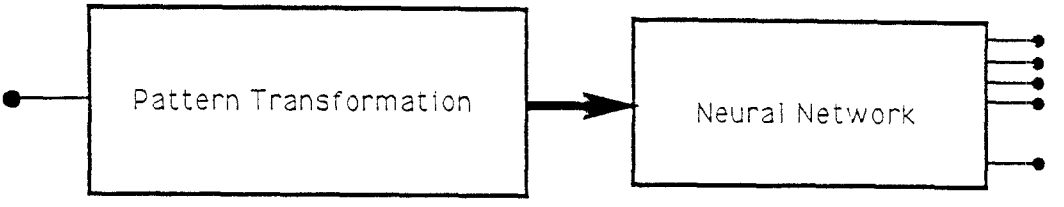


Figure 2: Classification System

4 Auditory Representation

In this section we present a brief review of the auditory model on which our pattern transformations are based. The interested reader is referred to [9] for the details.

When a signal enters the ear it creates a pattern of activity (displacement) along the basilar membrane. This membrane has a length and hence a spatial axis associated with it which extends from its base ($s = 0$) towards its apex ($s > 0$). The magnitude of activity at any particular spatial location, s, on the membrane may be modeled as the output of a linear (bandpass) filter with impulse response $h_s(\cdot)$.

Because of the physical nature of the membrane the filters $\{h_s : s > 0\}$ are related to one another in a special way. Physically this is reflected in the fact that the basilar membrane actually provides a tonotopic (logarithmic) mapping of the acoustic spectrum. One might think of the basilar membrane as acting like a bank of bandpass filters where the center frequency of the filter $H_s(\cdot)$ is related logarithmically to s. As it will be seen, this type of relationship can be described nicely in terms of wavelet transformations.

Figure 3 depicts the auditory processing model in block form. In the first block the basilar membrane action is modeled as a wavelet transformation. Later stages incorporate complex nonlinear transformations which culminate in a greatly abbreviated representation of the input 'acoustic' signal.

4.1 Stage I: Wavelet Transformation

The wavelet transform (WT) provides a mapping of one dimensional (e.g. time) functions into a two-dimensional space (e.g. time and frequency). A detailed

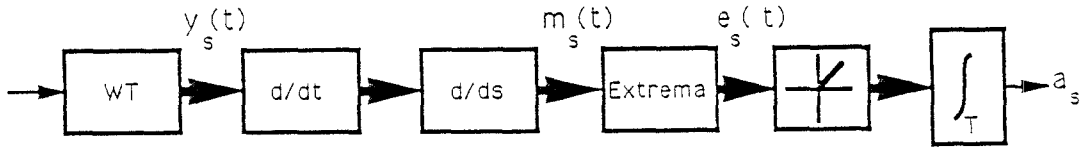


Figure 3: Auditory Model

discussion of wavelets and wavelet theory can be found in [3]. Simply, the WT can be thought of as a parallel bank of linear filters in which the filter responses share a special relationship. Namely they are related by dilations of a nominal response called the basic analyzing wavelet or mother wavelet. [1]

In these terms the wavelet output of the auditory model is a continuous ensemble of signals indexed by s which can be simply described as

$$y_s(t) = (x * h_s)(t). \tag{4.1}$$

Discretizing the s-axis the filter impulse responses, $h_k(t)$ are related by dilations of the 'mother' wavelet denoted by $h_0(t)$:

$$h_k(t) = a_0^k \, h_0(a_0^k t) \tag{4.2}$$

where a_0 is a given discretization parameter.

4.2 Later Stages

The later stages of auditory processing consist of several non-linear and complex transformations. It is out of the scope of this paper to describe in detail the justification for each of these processing blocks. Rather we include a functional description merely for consistency. The reader should consult [9] for further insight.

After the WT the next two operations are again linear operations, differentiation in time and then along the s axis, to yield the mixed partial derivative

$$m_s(t) = \frac{\partial^2 y_s(t)}{\partial t \partial s}. \tag{4.3}$$

Next the mixed partial ensemble is placed through an extrema (associated with the wavelet transform $y_s(t)$ in t) filter. That is the mixed partial derivative is sampled at the location of extrema in $y_s(t)$ as a function of t. In mathematical notation we may write the output of the extrema filter as

$$e_s(t) = \begin{cases} m_s(t), & t \in \mathcal{E} \\ 0, & else \end{cases} \tag{4.4}$$

[1] The mother wavelet is subject to the admissibility condition $\int_{-\infty}^{\infty} \frac{|H_0(\omega)|^2}{|\omega|} d\omega < \infty$.

where $\mathcal{E}$ is the set of extrema associated with $y_s(\cdot)$:

$$\mathcal{E} \triangleq \left\{ t : \frac{\partial y_s(t)}{\partial t} = 0 \right\}. \tag{4.5}$$

After the extrema filter the resulting pattern is then half-wave rectified and integrated over a small window of time to achieve the final pattern. Denoting $a_s(t)$ the resulting auditory pattern (in s) we write

$$a_s = \int_{T_w} \max\{e_s(t),\ 0\}\ dt, \tag{4.6}$$

where T_w is the length of the window of integration.

It should be noted that the original signal, $x(t)$, can be reconstructed from the acoustic representation a_s [9] . So that we can think of the acoustic representation as a data compaction; that is, there is no loss of information in the transformation.

5 Implementation

Presumably a real classification system would take an incoming data stream and continuously perform its classification. Because of the amount and expense of computation this would require, however, we do not fully process every frame of data. Instead we first pass each frame of data through a crude signal detector. The purpose of the detector is only to determine the existence of a signal and provide its starting position in the given frame. Its purpose is not to classify it. Should the detector decide that there is indeed a signal present then the the frame will be subject to further processing else it is discarded.

5.1 Detection on Sectors

A region S in the wavelet time-frequency plane can be chosen based on knowledge of where in frequency the signals to be detected lie. From Table 1 it can be seen that the regions for the signals of interest, namely A, B, and C, can be taken to be identical. For simplicity we take S to be a rectangle around the frequencies of interest (e.g. 3kHz).

For an arbitrarily long input signal, $x(t)$, a uniform partition in time is established with a frame or sector length of T_F so that

$$x_i(t) = \begin{cases} x(t), & t \in [iT_F, (i+1)T_F] \\ 0, & else \end{cases}. \tag{5.1}$$

So for the transient signals A,B and C the length of the rectangle S is T_F.

We form the detector function $d(\cdot) : L_2(\mathbb{R}) \to L_2(\mathbb{R})$ by

$$d(x_i)(t) = \int_S y_s^2(t) ds \tag{5.2}$$

where $y_s(\cdot)$ is as in Equation 4.1. A signal is deemed present by a simple thresholding of the detector function:

$$\exists\ t\ \ni d(x_i)(t) > t_{\text{thresh}} \implies \text{signal present.} \tag{5.3}$$

The set of potential starting times of a signal is , $\mathcal{T}_{\text{start}}$, is determined as

$$\mathcal{T}_{\text{start}} = \left\{ t\ :\ d(x_i)(t) = t_{\text{thresh}},\ \left.\frac{d}{dt} d(x_i)\right|_t > 0 \right\}. \tag{5.4}$$

Figure 4 illustrates the detector function for one of the signals (A) in the STDS. The figure displays in the bottom graph the wavelet transformation computed on S as well as the detector function itself plotted at the very top of graph. In the top graphs is displayed the signal on T_F and the mother wavelet $h_0(t)$.

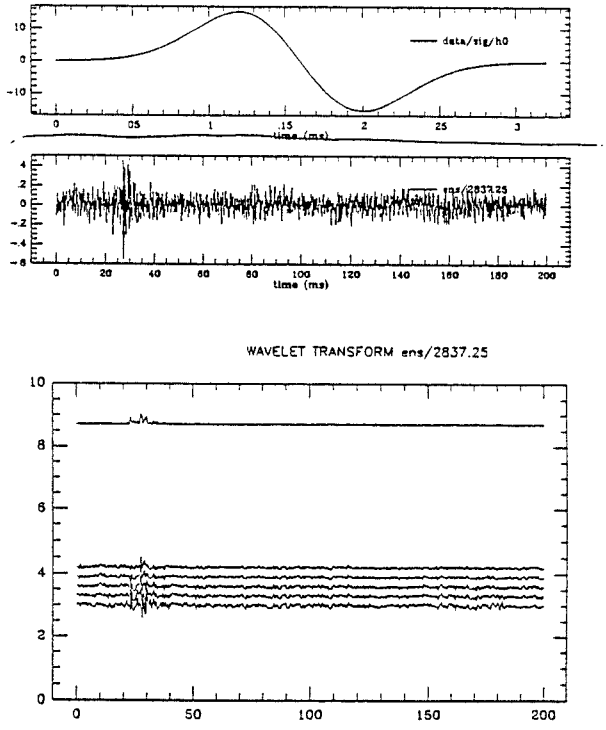


Figure 4: Detection on S

5.2 Auditory Model

The choice of mother wavelet in all the simulations is fixed as

$$h_0(t) = te^{-t^2/2}\cos(\omega_0 t) \tag{5.5}$$

where ω_0 is a modulation constant. This choice is not strictly faithful to the auditory model since it is well know that the auditory filters are asymmetric in the frequency domain (and $H_0(\omega)$ is symmetric). We note that the shape of these filters is a key area of further investigation. For the purposes of this paper we shall not address this issue. The choice of the modulation constant ω_0 is discussed in Section 6.

Another implementation issue is the choice of the fundamental processing window size, T_w. Choosing T_w too large would have the effect of smoothing out fast transitions on the final output, a_s of the auditory model (equation 4.6). Choosing T_w too small might cause necessary information about the signal to be lost. Clearly the window size must be chosen to be on the scale of the events in which one is interested. Since we are interested in the range of around f =3kHz a good choice of window size might be $T_w = 10T_f$ = 3ms. (An order of magnitude larger than the period of the phenomenon in which we are interested).

Outputs of various stages of the auditory model can be found in the Appendix. The Appendix shows the wavelet transformations of the three example transients A, B, and C taken from the STDS. In the figures the parameter $a_0 = 1.15$ and k takes integer values in the range $[-7, 7]$.

For convenience we list in Table 2 the values of all parameters used in the classification schemes we have discussed.

Parameter	*Description*	*Value*
T_F	sector window size	200ms
t_{thresh}	detector threshold	0.15
T_w	fundamental window size	3ms
a_0	dilation base	1.15

Table 2: Parameters of Implementation

5.3 Neural Network

The function of the neural network is to learn an association which performs the desired classification of the input signals. Mapping abilities of feed-forward networks have been well established and studied in the literature [4], [1].

In the simulations we use three layer (one hidden) neural networks with the topology as n_i input nodes, $n_i/2$ hidden nodes, and six output nodes for both the spectral and auditory based patterns. The number of input nodes, n_i, is determined by the number of points in the representation of interest. We choose six output nodes to accommodate each of the six basic transient signals (A-F) although we

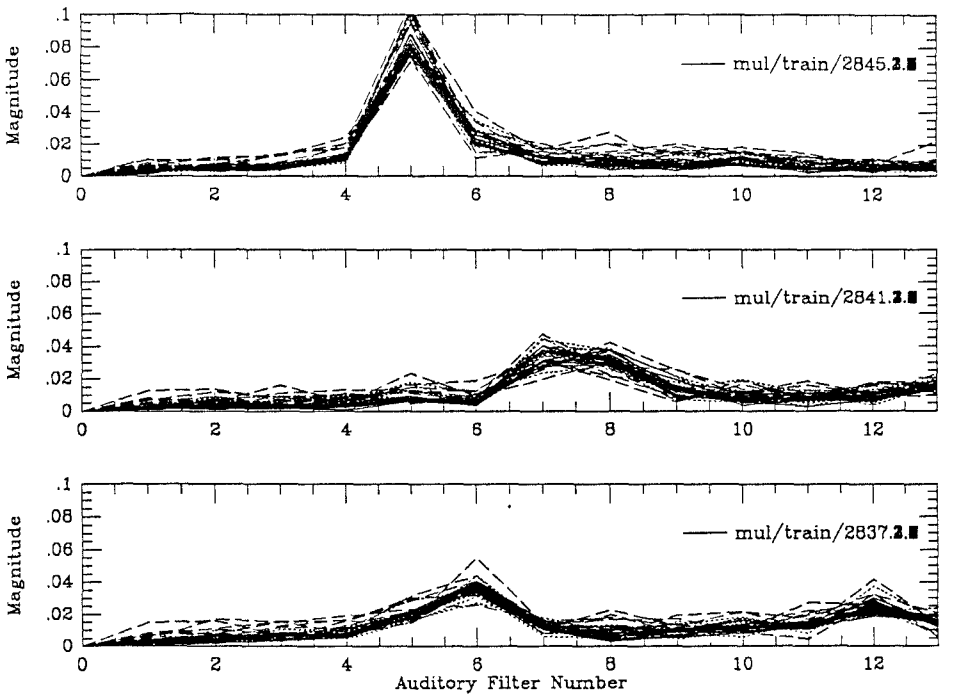


Figure 5: Example Patterns for Auditory Classification

only use the first three in this study. The training set is designed so that only one output (corresponding to the signal's true class) is active.

6 Spectral vs. Auditory Representations

In order to make a fair assessment of the relative merits of between spectral and auditory representations it is necessary to insure that the underlying filters have equivalent frequency resolution at the frequency of interest. In the following we compare the spectral and auditory representations and explain the details of 'matching' bandwidths at the frequency of interest.

6.1 Spectral Representation

Suppose we have a continuous bandlimited signal x with Fourier transform X. Further suppose we have sampled version of a x denoted x^* which has been uniformly sampled at the rate T_s:

$$x_k^* = x(kT_s). \tag{6.1}$$

If T_s is at least as large as the Nyquist rate then an N-point discrete Fourier transform (DFT), $\{X_k\}$, on the sampled signal, x^*, will yield samples of X with resolution $\omega_\Delta \triangleq \frac{2\pi}{NT_s}$, i.e.

$$X_k = X(\frac{2\pi}{NT_s}k) = X(k\omega_\Delta). \tag{6.2}$$

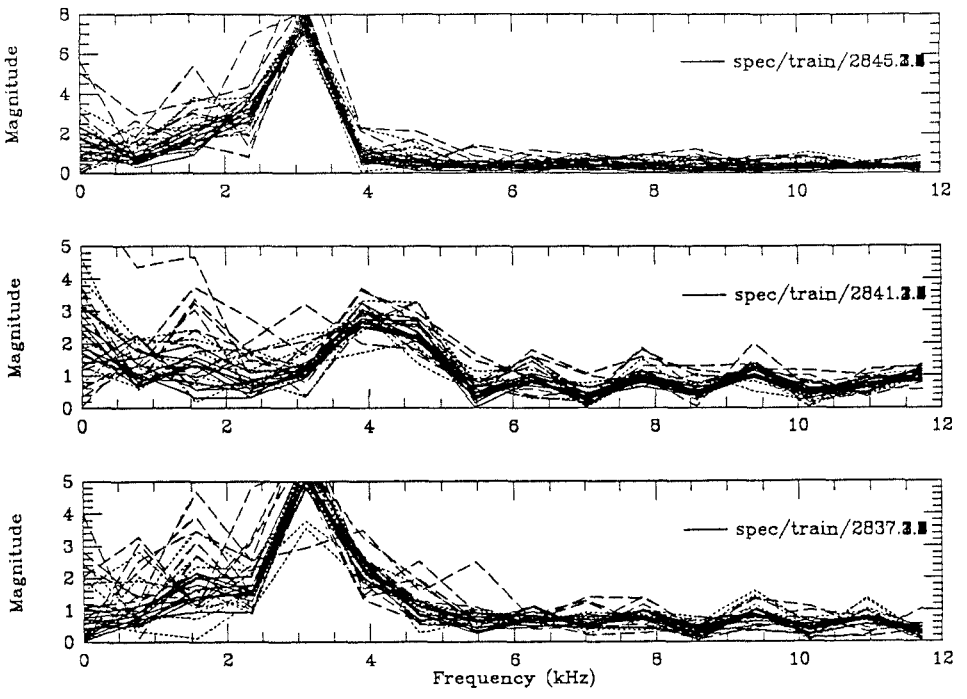


Figure 6: Example Patterns for Spectral Classification

Thus, an N-point DFT can be thought as a bank of linear filters, $\{H_k\}$, where each filter response, H_k, is an impulse at the kth harmonic of ω_Δ

$$H_k(\omega) = \delta(\omega - k\omega_\Delta). \tag{6.3}$$

We think of the ω_Δ as the frequency resolution of the filter bank.

6.2 Auditory Representation

Section 4 details a functional view of the auditory representation of signals. Of main concern here is the initial wavelet transform since it is this first linear stage which is responsible for segregation in frequency. Taking the view of the WT as a bank of linear filters with impulse responses generated by the mother wavelet given in equation 5.5 (repeated here for convenience)

$$h_0(t) = te^{-t^2/2}\cos(\omega_0 t), \tag{6.4}$$

we see that the entire filter bank is parameterized by ω_0. Thinking of the filter responses on the frequency axis it is easy to see that ω_0 controls the position of the entire filter bank.

6.3 Comparison: choosing ω_0

As mentioned earlier, in order to make a fair comparison between the spectral and auditory representations it is necessary to insure that the auditory filter at

the frequency of interest ω^* (=3kHz for the STDS) has its bandwidth equal to the resolution of the spectral filters, i.e. ω_Δ. By choosing ω_0 properly the bandwidth of the spectral filter at a specific frequency, ω^*, can be adjusted to a value within the precision allowed by the resolution of the dilation lattice $\left\{a_0^k\right\}$. In other words any filter in the auditory bank $\{H_k\}$ may be translated to the location ω^* by the choice of ω_0. This idea of matching bandwidth to resolution is illustrated in Figure 7.

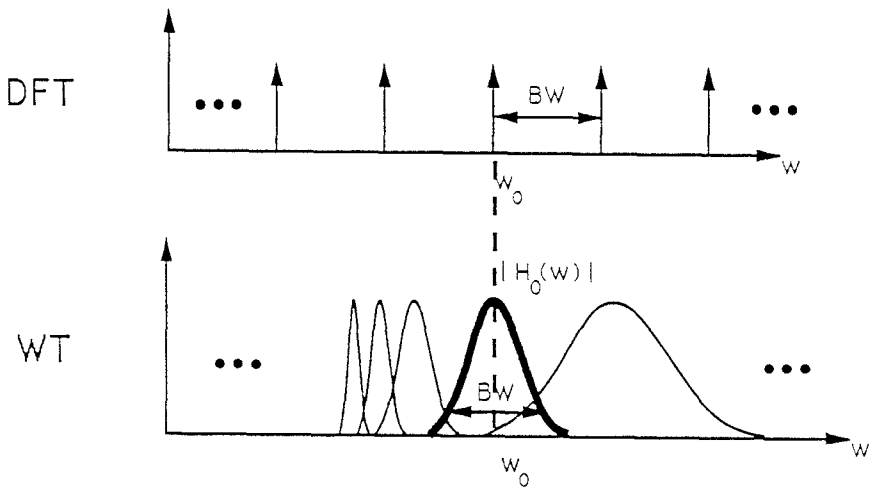


Figure 7: Spectral vs. Auditory Filters

7 Performance Evaluation

7.1 Training

Let $e_i, i = 1, \ldots 6$ denote the ith standard vector in $\mathbb{R}^6$, i.e. $e_i \in \mathbb{R}^6$ has a 1 at the ith position and zeros everywhere else. Given a windowed signal x of length T_w from one of the basic transient signals in the STDS, a corrupted version of x is formed as

$$\tilde{x} = x + \alpha n, \tag{7.1}$$

where n is a frame of noise chosen at random from the STDS signal 2836. The parameter α represents the noise level. Given a signal x and a noise level α a whole set of M signals may be generated via Equation 7.1. We call this random set $\mathcal{T}_\alpha^M(x)$.

A training set, $\mathcal{T}$, of data is formed as

$$\mathcal{T} = \bigcup_{i=1}^{3} \left\{(x, e_i) : x \in T_\alpha^{10}(s_i),\ \alpha = 1, 2, 3\right\} \tag{7.2}$$

where $s_1 = A, s_2 = B, s_3 = C$ are signals (respectively 2837, 2841, 2845) from the STDS. Similarly, a testing set $\mathcal{S}$ is formed as

$$\mathcal{S} = \bigcup_{i=1}^{3} \left\{ (x, e_i) : x \in T_\alpha^{25}(s_i),\ \alpha = 1 \ldots 9 \right\}. \tag{7.3}$$

Once these sets are formed they are fixed and used in both methods of detection.

After the network has learned the desired association the output to a member of class i would ideally be e_i. Generally speaking the network output will not be e_i but something close to it. For this reason it is necessary to introduce a network decision function, $d_\tau(\cdot) : \mathbb{R}^6 \rightarrow \{A, B, C, \ldots, NS\}$ (i.e. it takes the output of the network and assigns to it a label). Given an output pattern of the network $z \in \mathbb{R}^6$ the network decision rule which we have used here is given as

$$d_\tau(z) = \begin{cases} s_i, & z_i \geq \tau,\ \ z_j < \tau,\ \forall j \neq i \\ NS, & else \end{cases}. \tag{7.4}$$

where $0 < \tau < 1$ is a threshold.

Figures 5 and 6 show respectively the patterns generated by the two transformations (auditory and spectral) on the training set $\mathcal{T}$. In the Figures they are label by their STDS names 2837, 2841, 2845. Different line types are associated with different noise levels (solid being $\alpha = 1$).

7.2 Testing: Receiver Operating Characteristics

As a performance measure the receiver operating characteristics (ROC) are measured. Recall that the ROC is a collection of curves describing the probability of detection as a function of the probability of false alarm for a set of different noise levels. For the transient problem these probabilities [2] are given as

$$P_D = \text{Prob}\left\{ A|A \bigcup B|B \bigcup C|C \right\}, \tag{7.5}$$

and

$$P_{FA} = \text{Prob}\left\{ A \bigcup B \bigcup C \ \middle|\ NS \right\}, \tag{7.6}$$

where NS denotes a 'no signal present' condition.

For a given signal to noise ratio (SNR) the ROC curve is generated by stepping the classifier threshold parameter τ through the interval $(0, 1)$ and then estimating the detection and false alarm probabilities as sample averages. Thus for each value of $\tau \in (0, 1)$ a corresponding point $(P_{FA}(\tau), P_D(\tau))$ on the ROC curve is generated.

[2] Notation: Prob $\{X|Y\}$ is read 'the probability that the classifier labels the signal as 'X' given that the signal is 'Y'.

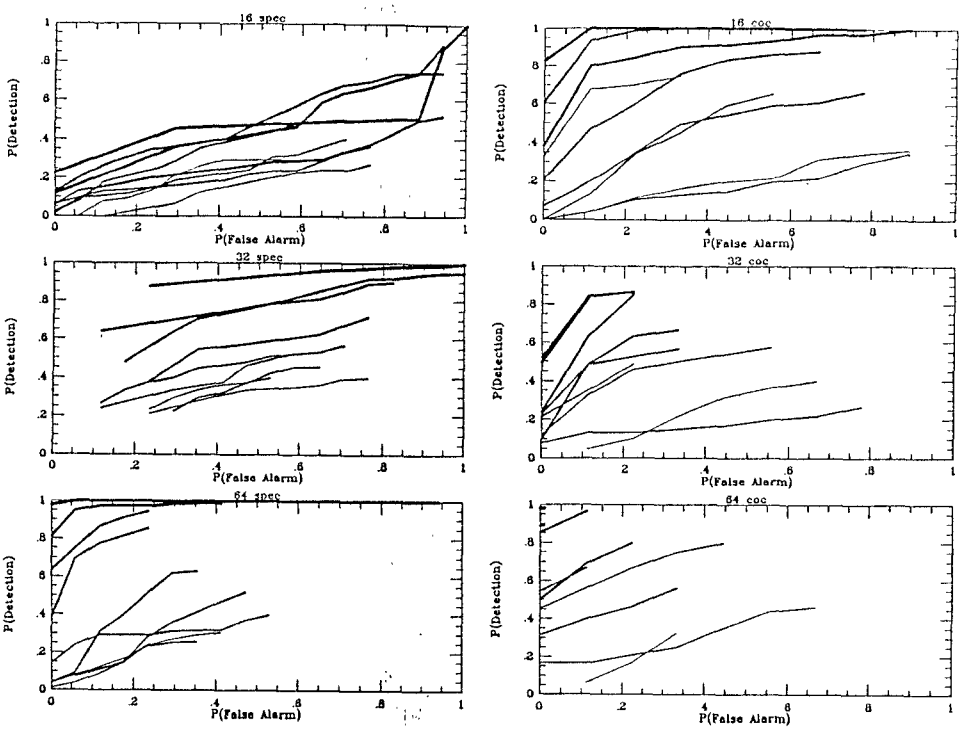


Figure 8: ROCs for Classifier Schemes

Figure 8 displays the ROC's generated for three different values of ω_Δ corresponding to N-point DFTs [3] with $N = 16,\ 32,\ 64$. Each ROC contains nine curves corresponding to the noise levels $\alpha = 1, 2, \ldots, 9$. To distinguish the curves line widths have been made inversely proportional to the noise level which they represent. So that the fattest line corresponds to a noise level of one and the thinnest to a noise level of nine. Table 3 gives these noise levels in terms of their decibel equivalents.

α	dB
1	0.0
2	-6.0
3	-9.6
4	-12.0
5	-14.0
6	-15.6
7	-17.0
8	-18.0
9	-19.0

Table 3: Noise level: decibel conversion

[3] of course the DFTs are computed via FFTs

7.3 Evaluation via ROCs

It is clear that the very best a detector could ever hope to achieve is a simultaneous probability of detection of 1.0 and probability of false alarm of 0.0 for all SNRs. This would correspond to a ROC which consists of a single point at $p^* \triangleq (0,1)$. With this in mind it seems that a plausible measure of the badness of a detector would be the extent to which its ROC contains curves far from the point p^*. In other words detectors whose ROCs are close to the point p^* can be considered better then those which are far from it.

Making this notion of badness via the ROC more precise we define a badness measure on a discrete ROC as

$$\text{ROC } \textit{badness} = \sum_{\sigma} \sum_{k} \|\gamma_\sigma(k) - p^*\|^2 . \tag{7.7}$$

Here σ indicates a noise level and γ_σ is the sampled (measured) ROC curve corresponding to that noise level. Obviously $\gamma_\sigma(k)$ indicates the kth sample of the curve. Certainly the ROC badness measure is only meaningful as a comparison among ROCs which have an identical sampling structure.

With the aid of this measure we can compare the ROCs in Figure 8 in a precise and quantitative way. Since this is a measure of badness then clearly we say that a detector whose ROC has a small measure compared to one which has a large measure is 'better'.

Table 4 lists the results of the computation of the measure 7.7 on the ROCs of Figure 8. The first column of the table displays the number of points in the DFT (which directly gives ω_Δ). The second and third columns give respectively the value of the badness measure for the spectral and auditory based classifiers.

N	*Spectral Badness*	*Auditory Badness*
16	7.78	4.54
32	5.86	4.11
64	3.66	2.29

Table 4: Comparison via badness measure

Clearly, the table indicates that the auditory based classifier outperforms the spectral based classifier in every case. Moreover, as the bandwidth of the filters is decreased (i.e. ω_Δ decreased) the performance for the auditory based filters increases (badness decreases) at a faster rate then for the spectral case. Both of these facts demonstrate the superiority of the auditory based classifier.

8 Conclusion

We have examined the problem of transient signal classification using two distinct pattern transformations. In the first we looked to a simple (magnitude) Fourier decomposition of the the transient signals. In the second we implemented a model based on the processing of acoustic signals in the mammalian auditory system.

We have empirically demonstrated the superiority of the auditory based classifier over the spectral based classifier.

References

[1] G. Cybenko. Approximation by superpositions of a sigmoidal function. Technical report, Tufts University, Department of Computer Science, October 1988.

[2] R. P. Gorman and T. J. Sejnowski. Learned classification of sonar targets using a massively parallel network. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7), July 1988.

[3] Daubechies Ingrid. The wavelet transform: A method for time-frequency localization. Technical report, AT&T Bell Laboratories, Murray Hill, New Jersey, 1990.

[4] R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine.*, pages 4–22, April 1987.

[5] H. Vincent Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, 1988.

[6] David Rumelhart and G. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, Ma., 1988.

[7] S. Shamma. Speech processing in the auditory system i: The representation of speech sounds in the responses of the auditory nerve. *Journal of The Acoustical Society of America*, 78(5):1612–1621, November 1985.

[8] S. Shamma. Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of The Acoustical Society of America*, 78(11):1622–1632, November 1985.

[9] X. Yang, K. Wang, and S. Shamma. Auditory representations of acoustic signals. Technical report, University of MD, Systems Research Center, 1991.