Broadband active sonar classification

Charles F. Gaumond[1], Derek Brock[1], Paul Hines[2], Stefan Murphy[2], Colin Jemmott[3], Christina Wasylyshyn[1]

1 Naval Research Laboratory, 4555 Overlook Avenue SW, Washington DC 20375-5000, USA, c**harles.gaumond@nrl.navy.mil**
2 Defence Research and Development Canada, Dartmouth, NS, Canada
3 Pennsylvania State University, State College, PA, USA

Active sonar performance is limited by noise, reverberation or false alarm rate (FAR). High FAR is overcome through the use of signal classification, which depends on the information content of the signals. Information content and classification are addressed using a set of broadband sonar echoes. The set of broadband echoes generated by an explosive source scattered from 30 different geographic locations is described. A continuous wavelet transform is shown that simulates the audible content of the echoes. The root-mean-square log scalogram distance, an audio distance defined with wavelet scalograms, between echoes is shown to decrease with decreasing sonar-source bandwidth. Results from multidimensional scaling (MDS), additive tree analysis, and exit interviews from a previous human subject listening study are shown. The results suggest that subjects use a process other than one defined by a single feature space. A human-subjects-research experimental protocol is described that will elucidate an alternative approach to signal classification. Funding was provided by the Office of Naval Research of the United States Navy.

1 Introduction

This paper presents continuing work on the problem of active sonar classification using broadband signals. The long term goal of the project is the development of numerical methods that emulate the processes that people use to categorize or classify sonar echoes.

In previous work, impulsive, broadband sonar signals were used in a psychoacoustic studies [1,2,3,4,5]. The psychoacoustic listening studies measured listeners' assessments of dissimilarity between pairs of echoes, sequentially presented diotically, *viz*., the same signal presented identically to both ears. Freely elicited, verbal descriptions of each signal were also recorded from each listener during individual exit interviews.

The two kinds of data collected support different types of perceptual models. The collected dissimilarities form a dissimilarity matrix of numbers, with larger numbers corresponding to greater dissimilarities between two signals. A value of zero signifies identical sounds; the main diagonal of the dissimilarity matrix should contain zeros.

The dissimilarity matrix was analyzed with multidimensional scaling (MDS) that assumes a model of perception based on a linear vector space. MDS generates a linear vector space with each signal represented as a point in that space. Distances between point pairs represent the quantitative dissimilarities. The space is generated through a numerical fit of the distances to the dissimilarities in the matrix. This space with echoes as points is suggestive of a feature space found in classification where each signal is described with a feature vector [6].

However, other psychoacoustic models of dissimilarity exist, *e.g.*, the contrast model of Tversky [7]. In this model dissimilarities are modeled with a metric based on set theoretic descriptions of the signals. For example, one signal may have a set of qualities like {*medium frequency*, *muffled*}, whereas another may have a set of qualites like {*medium frequency*, *cannon*}. The contrast distance is defined as a linear function that is a weighted sum of the common and distinct features. It is not a linear vector space. This approach does not appear to have been followed in the sonar classification literature.

This paper describes the early stages of research using another impulsive sonar echo data set. This set was generated in the Boundary 04 experiment. The first section of this paper has a short description of the Boundary 04 data set [4,8], followed by two sections describing psychoacoustic inspired definitions of distances, with a description of the protocol for human subjects research that will be used to distinguish different types of perception and guide development of new classification methodologies.

2 Data Description

The data used are described in [4,8]. There are 28 data classes each caused by reflections from a distinct geographic location in the experiment. Each class contains example signals, or elements. Figure 1 below shows that classes 28 and 29 contains 62 and 36 elements, respectively. Classes 3, 5, 6, 13, 14 and 23 contain fewer than three elements.

Each example signal is a noisy, broadband echo. Each signal contains different background noise because the signals were recorded from different beams at different times. Therefore, there is a wide variety of background noise sources, *e.g.*, from surface ships. Human listeners are able to perceive the impulsive echo as an aural object that is distinct from the background noise. In the first experiment, listeners were instructed to ignore the background noise and judge the dissimilarities of the echoes only. Likewise, human mimetic signal processing algorithms ideally would ignore the background noise, too.

The next section discusses the signal processing steps taken to address those two concerns.

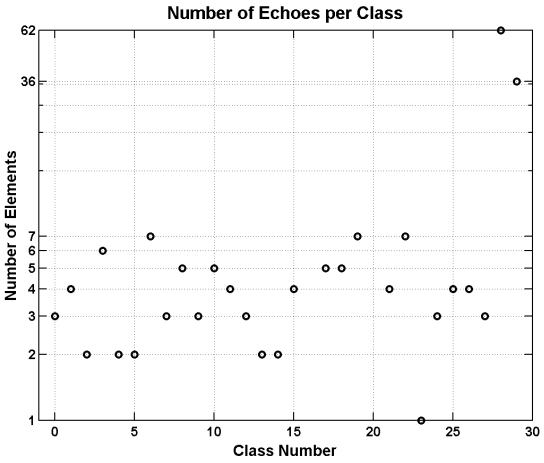

Figure 1 Histogram of Boundary 04 echoes in classes 0 through 29.

3 Signal Processing

3.1 Preliminary processing

The mitigation of the effects of background noise is accomplished in two steps. A wavelet domain is defined in a way that emulates aural perception. The signals are first transformed into the scale- time domain using the continuous wavelet transform [9,10] using the wavelet kernel

$$W(s,t)=\cos(\pi st/2)\,i\,e^{(i\pi Qst)},-1<st<1 \qquad (1)$$

with scale $s=2f/Q$ and with Q , the number of cycles in the wavelet, being set to six to simulate the the aural bandwidth of a gammatone filter, forming

$$\tilde{X}_{c,m}(s,t)=\int_0^T W(s,t)\,x_{c,m}(t)\,dt \quad . \qquad (2)$$

The subscript c denotes the class of clutter and m the index of the particular element of that class.

The integral is evaluated numerically using a summation over values of scale chosen to cover the frequency band from 200 to 1800 Hz with constant Mel spacing determined with the formula

$$f=1000\,(2^{Mel/1000}-1) \quad . \qquad (3)$$

The fineness of the Mel interval is adjusted so that the impulse response function of the wavelet transform smoothly covers the frequency interval of interest. This ensures that signals – that are transformed into the wavelet domain, filtered and then transformed back into the time domain – are not distorted by the transform itself.

This signal domain produces a two-dimensional visual, and also quantitative, representation of the signal that crudely mimics the audibility of the signal. The cognitive process of separating the echo, or foreground, from the noise, or background, is modeled by equalizing the noise level over the frequency band. An example noise-equalized signal is shown in Figure 2 where $\left|\tilde{X}_{28,1}(2f/Q,t)\right|$ is displayed.

The noise equalized signals have background noise that sounds very uniform over all the elements of all the classes. The echo components have different amplitudes, but the background noise levels are all equal. This should decrease the effects of background noise differences on the listeners' judgments of dissimilarity.

Although the noise-equalized signals have uniform background noise, the amount of background signal energy may not be negligible when signal processing algorithms are used to compute the distances between signal pairs. This problem arises for algorithms that mimic human perceptions of signal dissimilarity. In this paper, we use the term *dissimilarity* to refer to a human perception and the term *distance* to refer to the output of an algorithm.

With the goal of eliminating as much background noise as possible, the noise-equalized signals are denoised using a binary mask applied in the wavelet domain [11,12]. The binary filter equals unity at a value of scale and time if the value of the wavelet transform there exceeds a threshold; otherwise the binary filter has a value of zero. The binary-masked signals are perceived by the lead author as abrupt compared with the noise-equalized signals. The perceived differences are probably due to the inability of a human listener to accommodate to the background noise over the short duration of the binary-masked signal. These binary-masked signals are used for the estimation of information content because they have the highest signal energy to noise energy ratio possible from filtering in the wavelet domain. The higher SNR improves the meaningfulness of the information estimates. In other words, the estimates of information content of the signals are contaminated by the least amount of noise. The denoised signals are denoted as $\check{X}(f,t)$ in the scalogram domain and $\check{x}(t)$ in the time domain.

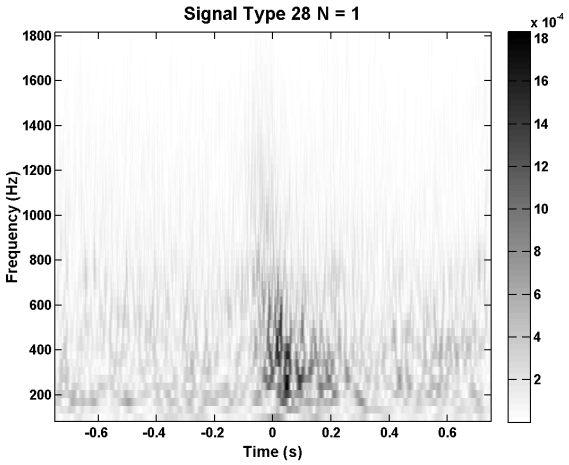

Figure 2 Scalogram $\left|\tilde{X}_{28,1}(f,t)\right|$ of the normalized signal class 28 number 1

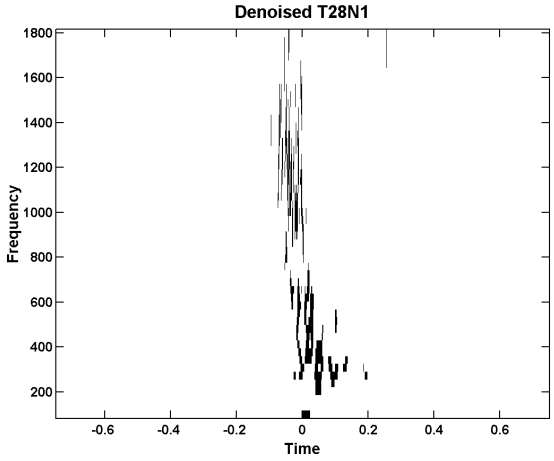

Figure 3 Scalogram $|\check{X}_{28,1}(f,t)|$ of the denoised signal class 28 number 1.

3.2 Linear vector space distances

In previous work, eleven human subjects were presented with pairs of impulsive sonar signals, similar to those from Boundary 04. Each subject judged the dissimilarity between the signals on a scale of zero to ten, with zero denoting the perception of two identical signals. The dissimilarity matrices of subjects who generated similar results were averaged and analyzed using MDS. The resulting three dimensional space is shown below in Fig. 4.

In Fig. 4, the signals are classed as either target T or clutter C followed by an integer element number. In this representation several clusters appear, indicating that there is not a single clutter region, but rather that there are two clutter groups that sound as distinct from each other as they do from the cluster of target echoes.

The existence of more than one clutter group indicates that experimenting with a data set containing more tags than T and C could confirm the existence of different kinds of clutter sources that are aurally distinct.

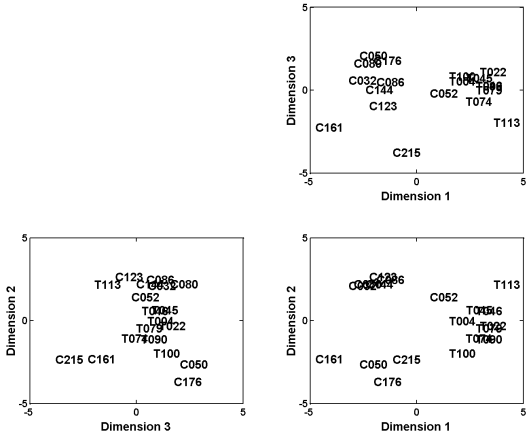

Figure 4 Resulting MDS representation of human-subjects data from a previous listening study

A natural interpretation of the configurations produced in the MDS solution is the assignment of aural features to the three axes of MDS space, or to a set of trajectories in MDS space [13,14]. The ultimate purpose is the replacement of human listeners with an algorithm. The algorithm would determine the aural features from a signal and then transform the aural feature vector onto the axes of the space generated from MDS listening studies. Decision boundaries that demarcate class boundaries in MDS space would then, in principle, mimic the behavior of a human listener.

An early step in the understanding of the relationship between signal distance and human-subject dissimilarity, is the development and study of algorithms that compute distances between signal pairs. A distance that is based on a simple signal-representation is the RMS log spectral distance[15]. This distance is modified for use with the scalogram representation of two signals $\check{X}_{c,m}(f,t)$ and $\check{X}_{d,n}(f,t)$

$$D_{RMS}(\check{X}_{c,m},\check{X}_{d,n};B)=(C_1\sum_{t=0}^{T}\sum_{f\in B}C_2(f,t))^{\frac{1}{2}}$$
$$C_1=\frac{1}{N_t N_f} \quad . (4)$$
$$C_2(f,t)=(\log(|\check{X}_{c,m}(f,t)|)-\log(|\check{X}_{d,n}(f,t)|))^2$$

Bandwidth is denoted as B. Note that the dependence on frequency and time is suppressed for simplicity in Eq. 4. Equation 4 is used to find distances between a selection of individual echoes.

The set of nine signals [{21,2} {2,14} {22,2} {24,1} {26,2} {1,1} {3,1} {4,1} {10,2}] were chosen because they have similar signal to noise ratio (SNR). The distances between all signal pairs is computed using Eqn. 4 and the results are displayed as distance matrices in Figs. 4 and 5 and histograms in Figs. 6 and 7. Figures 4 and 6 show the dissimilarity matrices for the broad-band case [200 Hz < f < 1800 Hz]. Figures 5 and 7 show the narrow-band case [900 Hz < f < 1000 Hz].

In Figs. 4 and 5, the horizontal and vertical axes are labeled with the class number c and the element number m. The signals are ordered in terms of decreasing SNR. The distance between a signal on the horizontal axis and a signal on the vertical axis is displayed using a gray scale. The gray scale is scaled to the maximum distance in each figure and the units are arbitrary but consistent between the two figures. The main diagonal values of zero show that each signal has a zero distance from itself, which is a requirement for the definition of distance. In Fig. 4, several signals have a significant distance from the others. Signals T26N2 and T4N1 are quite distinct from most other signals, including each other. Contrast this situation to that seen in Fig. 5, where only one signal, T26N2, is very distinct from the others. The distinctiveness of T4N1 has been removed through the selection of the narrower frequency band.

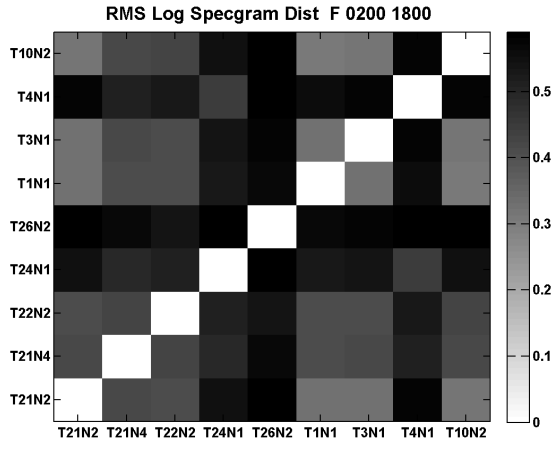

Figure 4 RMS log spectrogram distance matrix of broader bandwidth signals with 200 Hz $< f <$ 1800 Hz

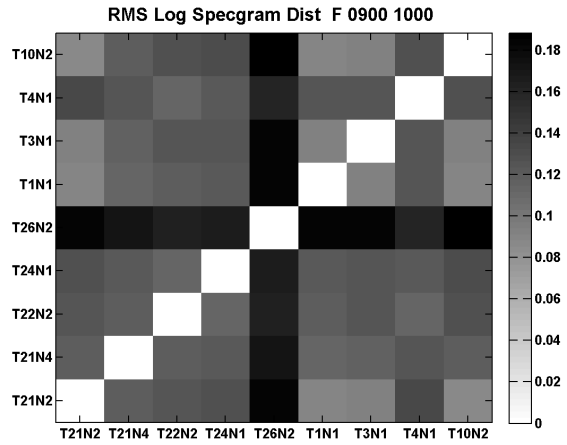

Figure 5 RMS log spectrum distance matrix with narrower bandwidth signals with 900 Hz $< f <$ 1000 Hz

This loss of distinctiveness can be seen in Figs. 6 and 7 that depict the distribution of distances as histograms. Figures 6 and 7 are plotted with the same horizontal scale that shows the distances, with the vertical scale showing the total number of signals falling within the interval. Figure 6 shows the broader and more evenly distributed distances between the broader bandwidth signals. The broad and even distribution shows that the signals in general have an appreciable distance between each other. If this distance model mimics human perception, then human subjects likewise would perceive appreciable dissimilarities between pairs of signals.

In contrast, the distribution found with narrower band signals shown in Fig. 7 is bunched to the left portion of the horizontal axis, showing that the distances are small. Not only are the distances small, but the distribution is not evenly spread over those small values. This is indicative of a clustering of the signals. In this case, one of the clusters would contain only one signal, T26N2, which is distinct from all other signals.

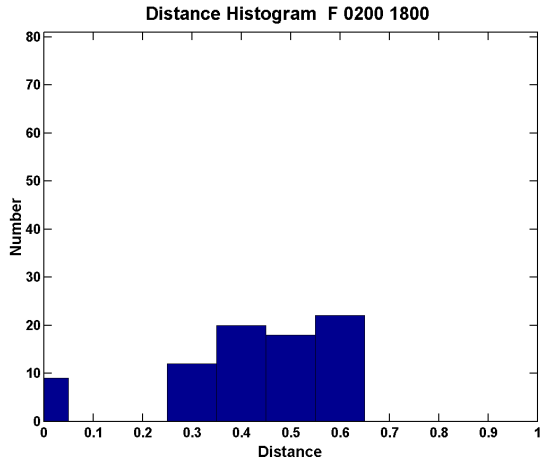

Figure 6 Histogram of RMS log spectrogram distances with broader band signals

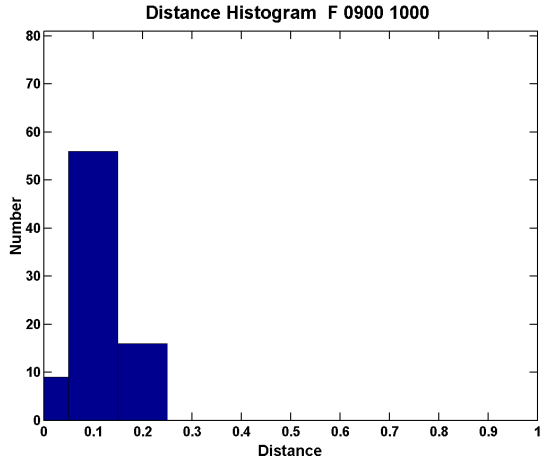

Figure 7 Histogram of RMS log spectrogram distances with narrower band signals.

These results qualitatively follow human experience where perception of signals is limited by the bandwidth that is presented to listeners. These distances above will be compared with the dissimilarities determined through human listening studies. Through this quantitative comparison to human perception, the quality of alternative signal distances can be assessed.

3.3 Information based class distances

A method of assessing the distance between groups or classes of signals also may be useful. Computing interclass distances could be based on the RMS log spectral distances defined with Eqn. 4 when combined with a method of clustering that defines distances between groups of signals as well as individual signals [16, 17]. An alternative to this is a method based on information theory that naturally applies to groups. The *a priori* grouping of the impulsive sonar signals in the Boundary 04 data set into classes enables the use of this information based technique. The

Shannon distance is based on the distinctiveness between the information found between classes [18]. The amount of information in a class c of denoised signals $\check{x}_c(t)$ is determined with the formula

$$H(\check{x}_c)=\sum_{x\in\check{x}_c(t)} P(x)\log_2(P(x)) \quad , \qquad (5)$$

where $P(x)$ is the probability of x that is a component of class c and where the summation is over statistically independent components. In this paper, the statistically independent components are generated from a finite set of signals using the singular value decomposition (SVD). This formula is based on the assumption that within each class each signal is equally likely. The signals in the class $\check{x}_c$ are first peak aligned. This alignment creates the largest common signal and therefore predicts the lowest information content for the collection of impulsive signals. Using SVD, the signals in class c are decomposed into expansion functions $\psi_l(t)$ that are orthogonal over summation of m in class c, namely

$$\check{x}_{c,m}(t)=\sum_{l=0}^{L} \upsilon_{c,m,l}\psi_l(t) \quad . \qquad (6)$$

For each l, the set of expansion coefficients $\upsilon_{c,m,l}$ are then converted into histograms using a common amplitude bin width. This bin width corresponds to the noise level used in signal information theory. The probabilities of each quantized coefficient value are then used in Eqn. (5).

The information within a signal class is representative of the variability within that signal class. In the context considered here, if a class contained only one noise-free signal, then it would be described with only one bit because there is only one expansion coefficient and the signal is either present or absent. If two signals were present and statistically orthogonal using the heuristic method above, then there would be two expansion coefficients and the the class would carry two bits of information.

The Shannon distance is a measure of the amount of distinct information between two classes of signals, c and d, and is defined as

$$D_S(c,d)=2H(\check{x}_c,\check{x}_d)-H(\check{x}_c)-H(\check{x}_d) \quad , \qquad (7)$$

using the definition of information in Eq. 5 and where the summation is over the probabilities that the signal lies in quantized, statistically independent intervals x of the whole signal space $\check{X}_c$. The determination of these quantities is performed on the denoised, frequency-normalized signals in the time domain, as opposed to the scale-time domain, due to their smaller dimension and limited computer memory.

The change of information content and the Shannon distance between classes 24 and 26 is shown in the Fig. 8 below. Figure 8 shows the amount of information in class 24 with the + symbol and class 26 with the x symbol. The Shannon distance is displayed as circles. The horizontal axis is bandwidth centered at 1000 Hz. Note the clear trend of decreasing information content and distance with decreasing bandwidth. There is, however, a lack of smoothness in the frequency dependence.

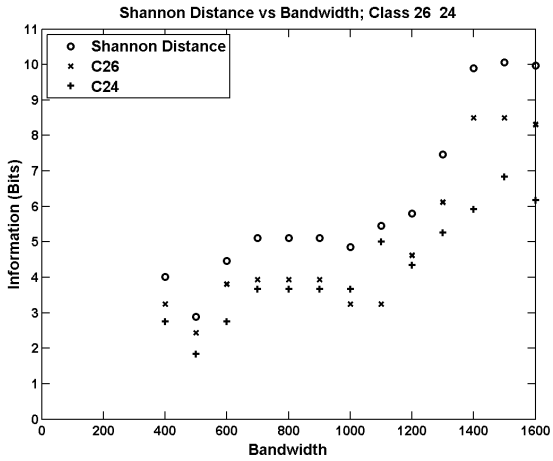

Figure 8 Shannon distance between classes 6 and 21 as a function of bandwidth centered at 1000 Hz

This lack of smoothness is the result of the discrete disappearances of contributing expansion coefficients in the Eqn. 5. Indeed, there are only three and four elements in the classes 24 and 26, respectively. Classes with more elements show smoother behavior due to the greater number of m values for each l in Eqn. 5.

Another observation from Fig. 8 is that the signal information and Shannon distances are not uniformly dependent on bandwidth. This is caused by the nonuniform distribution of signal energy over frequency. The signal classes are not uniformly broadband. This also causes abrupt changes in the amount of information in the groups of signals.

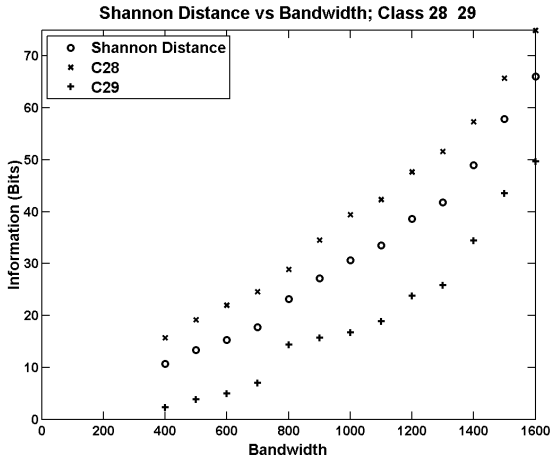

Figure 9 Shannon distance between classes 28 and 29 that contain many elements

The smoother behavior resulting from more class elements in shown in Fig. 9 where there is smooth and monotonic

dependence of information with bandwidth. These two signal classes have signal content from 200 to 1800 Hz, as is seen in the scalogram of a class 28 signal shown in Fig. 2. There is uniform frequency dependence of class 28 information content. The frequency dependence of class 29 is not as uniform, with a steeper dependence between 1300 to 1600 Hz bandwidths. This is due to less uniform frequency distribution of signal energy in class 29.

3.4 Future Work

Future work has two general components, signal processing and human subjects research. Some of the preliminary signal processing is presented here.

A human subjects experiment will be conducted to study more carefully how people compare one signal to another in this class of active sonar echoes. Presently, our plans are to select ten stimuli from archived data collected on the Malta Plateau. This will give us a 10x10 comparison matrix, which results in 100 possible ordered pairings, including those on the principal diagonal. Participants will be asked to aurally discriminate between each pair twice, for a total of 200 judgments. Next, we will ask participants to freely group these ten sounds into clusters of their choosing. After they have done this, they will be asked to explain why they clustered the sounds in the way they did. We anticipate participants' explanations will lead to a set of freely elicited aural descriptors for each sound. Finally, participants will be presented with a set of predetermined descriptors, such as *loud* and *high frequency*. They then will be asked to rate the appropriateness of each descriptor to each sound. The psychoacoustic data from the full experiment will be used to explore both continuous (e.g., multidimensional scaling) and non-continuous (e.g., set-theoretical constructs, such as additive trees) psychological representations of the aural stimuli. Additionally, the descriptor data will be used to explore the correspondence between physical features and how people verbally characterize sounds.

The dissimilarity data from the human subjects research will be compared with the RMS log spectral distances, information based distances, and other distances that may be implied by the behavior of the dissimilarities.

Likewise the relationship between the aural descriptors and the information content of the clutter groups will also be investigated. Some aural descriptors, such as *low frequency*, clearly imply some straightforward timbrel features. However other descriptors, such as *cannon*, seem to signify membership is a set of signals, that is defined by a physical response of some mechanical arrangement. These information based methods do not generate aural features from a signal, but instead generate a set of signals, membership of which can then be tested using signal processing algorithms.

In summary, the future work entails measuring new human responses to sonar sounds and then developing methods to simulate them with the hope that the ability to classify sonar echoes will be improved.

References

[1] J.E. Summers, *et al*., "Perceptual dimensions of impulsive-source active sonar echoes (A)." *J. Acoust. Soc. Amer.* 120, 3125 (2006).

[2] J.E. Summers, *et al*., "Modeling the mechanism and neural substrate for aural categorization of sonar echoes (A)," *J. Acoust. Soc. Am.* 123, 3345 (2008)

[3] C.F. Gaumond, *et al*., "Contrast models and classification of active sonar signals. (A)," *J. Acoust. Soc. Am*. 124, 2597 (2008)

[4] V. Young and P. Hines, "Perception-based automatic classification of impulsive-source active sonar echoes," *J. Acoust. Soc. Amer.* 122, 1502-1517 (2007)

[5] S. Philips, J. Pitton and L. Atlas, "Perceptual feature identification for active sonar echoes," *IEEE OCEANS 2006*, 1-6 (2006)

[6] R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification 2nd Edition, Wiley-Interscience: New York (2001)

[7] A. Tversky, "Features of similarity," *Psych. Rev.* 89, 327-352 (1977)

[8] M. Prior, "A scatter map for the Malta Plateau," *IEEE . Ocean. Eng.* 30, 696-690 (2005)

[9] C. Gaumond, "Application of wavelets to scattering for wavepacket synthesis," *Acoust. Res. Lett. Online* 1, 19-25 (2000)

[10] S. Mallat, A Wavelet Tour: The Sparse Way, Elsevier:Amsterdam, 102-115, (2009)

[11] N. Roman, D.L. Wang and G.J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.* 114, pp. 2236-2252, (2003)

[12] D.L. Wang and G.J.Brown, "Fundamentals of Computational Auditory Scene Analysis," in Computational Auditory Scene Analysis, D. L. Wang and G. J. Brown (eds.), Wiley-Interscience:Hoboken, 22-25, (2006)

[13] J. B. Kruskal and M. Wish, Multidimensional Scaling, p 30 *et seq*., Sage Press: Newbury Park, CA (1978)

[14] I. Borg and P.J.F. Groenen, Modern Multidimensional Scaling, Springer: New York (2005)

[15] E. Deza and M. M. Deza, Dictionary of Distances, p. 275, Elsevier: Amsterdam, (2006)

[16] J.E. Corter, Tree Models of Similarity and Association, Sage Publications: Thousand Oaks (1996)

[17] M.S. Aldenderfer and R.K.Blashfield, Cluster Analysis, Sage Publications: Thousand Oaks (1984)

[18] E. Deza and M. M. Deza, Dictionary of Distances, p. 187, Elsevier: Amsterdam, (2006)