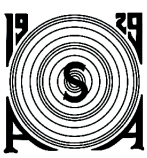


euronoise

Similarity-Based Perceptual Feature Identification for Active Sonar Signal Classification

Scott Philips[a] and James Pitton[b]

[a]MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420-9108, USA
[b]Office of Naval Research Global, Blenheim Crescent, West Ruislip, Middlesex, HA4 7HL London, UK
jpitton@onrglobal.navy.mil

In many aural (acoustic) signal processing tasks, humans are known to perform better than automated classification systems. For such applications, it may be beneficial to identify how humans relate different sounds to one another and incorporate that information into an automatic classification system. This paper presents a method for using psychoacoustic information from a human listening experiment to generate a novel kernel function that can be used to improve automated signal classification. We have conducted a similarity-based listening experiment on a series of impulsive-source sonar echoes. In this experiment, humans were asked to rate perceived similarity between pairings of target and clutter echoes. These ratings combine to form a similarity matrix that reflects the underlying distance measure humans use when judging these echoes. This similarity matrix is a perceptual equivalent to the similarity matrix used in modern kernel methods used in automatic classification systems (e.g. Support Vector Machine). By fitting an appropriate distance metric to the results of the perceptual experiment we can identify novel, perceptually-inspired, kernel functions. This paper presents a series of new approaches for the identification of a perceptual kernel function. We then compare the classification performance between these perceptual kernels and more standard kernel functions.

1 Introduction

For many aural (acoustic) signal processing tasks, humans are known to perform better than automated classification systems. For such applications, it may be beneficial to identify aspects of the approach used by humans and integrate those aspects into an automatic classification system. In applications such as speech recognition, superior human performance could be the result of high level processing (e.g. language models); in other applications, such as identification of transient signals, the key to human performance may lie closer to the periphery, perhaps in the identification and extraction of useful acoustic signal features for classification. Current features used in transient signal classification do not always provide acceptable performance; accordingly, new features are desired that yield the superior classification performance observed in humans. This research focuses on the acoustic feature problem as it relates to distance metrics between signals. Specifically, we utilize the results of a similarity-based listening experiment to examine short duration transient signals from active sonar systems with the goal of learning new, potentially useful, distance measures to aid automatic classification.

In order to understand how humans perform an aural classification task, detailed information regarding a sound's perceptual quality is needed. Most quantitative approaches describing perception of sounds use some measure for perceptual distance between sounds [1]. These perceptual distance measures are usually gathered directly from a similarity experiment in which subjects are asked to rate the similarity between pairs of sounds. This type of experiment provides a metric describing a subject's underlying perceptual feature space. Typically, this perceptual space is estimated using a technique called multidimensional scaling (MDS) [2]. This type of analysis allows the researcher to compare the estimated perceptual space with hypothesized signal features. Any features that correlate well with this space are said to be perceptually significant. In practice though, no test feature ever fully correlates to this space and the researcher is forced to choose the "best" perceptual feature. This problem is due to the limited set of commonly known features at the researcher's disposal. In addition, this approach does not provide an avenue to identify new features derived solely from the listening experiment data.

Recently, a more systematic approach to identify perceptually relevant features was introduced [3]. This approach provides an avenue to identify new perceptual signal features drawn from a general class of signal representations. This framework broadens the choice made by the researcher to an appropriate class of signal representations as opposed to a specific list of features. By following this method, new features can be identified that play an important role in perception. This paper expands upon this approach by removing the requirement of MDS to identify the perceptual space by directly estimating the distance metric used by subjects in a similarity-based listening experiment. Directly estimation a perceptual distance metric instead of the underlying features space allows errors introduced by multidimensional scaling to be avoided.

The rest of the paper is organized as follows. The next section describes two similarity-based listening experiments that were conducted using acoustic signals from impulsive-source active sonar systems. In section III two approaches are introduced for identifying a novel kernel function from perceptual similarity data. Section IV presents the application of these approaches on the results of the sonar listening experiment. Finally, in section V conclusions and future work are discussed.

2 Listening Experiment

The purpose of these experiments are to collect aural similarity measures from human responses to active sonar target and clutter echoes in order to explore the perception of these types of underwater sounds.

2.1 Data

The signals used in the following experiments are short duration sonar echoes from two impulsive-source active sonar systems (System A and system B). Detections from both systems were recorded and labeled as either a true target or false target (clutter) based on the known location of all the targets. This information provides ground truth for ongoing research in automatic classification. The "hard-case clutter" subset of each dataset is a collection of false target detections that represents all false targets that were misclassified by an automatic classifier. This clutter subset, along with detected true targets were collected and used to perform two listening experiments, one using data from system A and one from system B.

2.2 Experimental Setup

Task: Each subject was presented two sounds in succession. The subject was then asked to rate how similar the sounds were on a scale of one to five in experiment one, and one to ten in experiment two. A rating of one indicated the sounds were very similar while a rating of five/ten indicated the sounds were very different. After receiving instructions the subjects were given five "practice" trials in order to get a sense of the task. These practice trials were not included in the results or analysis.

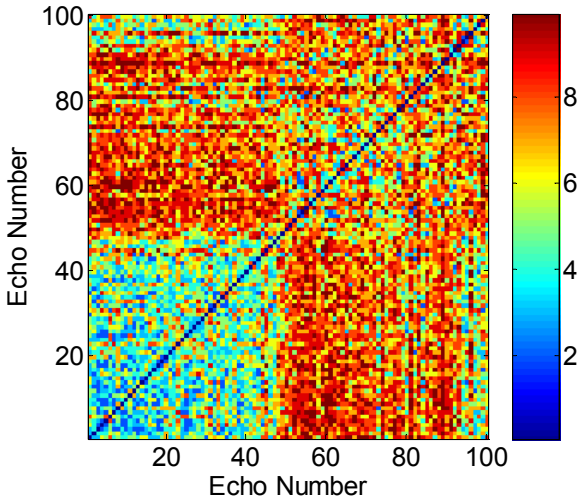
Subjects: Sixteen subjects were recruited from the staff of the Applied Physics Lab at the University of Washington. None of the subjects had operational experience in sonar systems. Their initial understanding of the sonar classification problem ranged from truly naive to detailed technical understanding of sonar signal processing. Previous experiments showed that all subjects were able to perform target/clutter discrimination significantly above chance [4].

Stimuli: In experiment one, a set of 50 targets and 50 clutter echoes from system A were randomly chosen from a database of greater than 200 echoes. Even with only 100 stimuli, the number of pair-wise combinations to be judged by each subject would have been 4950. In order to accommodate this large number, the set of pairings was randomly divided into four subsets with each subset rated by two subjects. This division of large datasets into smaller pair-wise subsets has been shown to provide good results in previous MDS similarity experiments [5]. In experiment two, a set of 25 targets and 25 clutter echoes from system B were used.

Presentation: Subjects were presented .wav files directly from a computer via a Matlab script. The stimulus pairs were presented in random order through an M-audio A/D board over Sennheiser HD 280 Pro headphones. Subjects were allowed to replay stimulus pairs as desired.

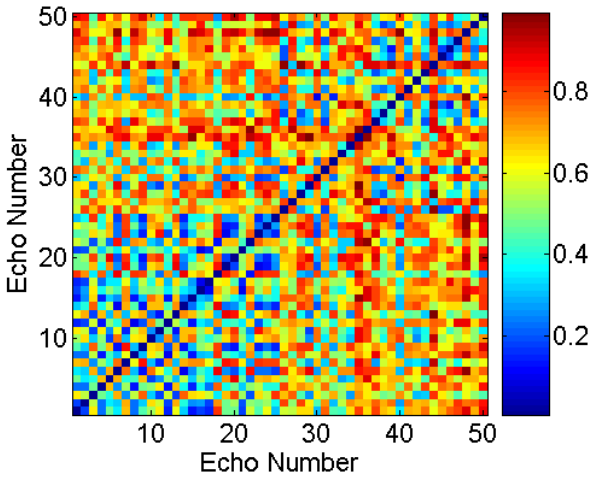
2.3 Results

Figure 1 shows the perceptual similarity matrix δ that results from the subjects' similarity responses to signals from system A. Each entry in the matrix represents one stimulus pair. The matrix is symmetric as only one ordering of the stimuli was used, and the resulting data were reflected about the diagonal. Each stimulus pair was presented to two subjects, each of whom rated the similarity between 1 and 5. The two subjects' responses were added and entered into the matrix; thus the range of values is between 2 and 10. Note that the lower left quadrant of the matrix, corresponding to target-target stimulus pairs, has substantially lower values than the rest of the matrix, indicating more similar sounds. The values in the lower right quadrant (target-clutter pairs) are much higher, indicating that targets and clutter were usually judged more dissimilar. The upper right quadrant (clutter-clutter) has a wide range of values compared to the other two quadrants, which suggests that the clutter examples are not a single coherent class, but rather span a wide range of stimulus types that differ from the target class.



.Fig.1 Aural similarity matrix for both targets (echoes 1-50) and clutter (echoes 51-100) from experiment 1. Note that only upper right quadrant corresponding to clutter-clutter pairings will be used in the analysis shown in this paper.

The results from system B are shown in Figure 2. Each stimulus pair was presented to four subjects, each of whom rated the similarity between 1 and 10. The four subjects' responses were then averaged, normalized, and entered into the matrix. As before, the similarity matrix is exactly symmetric as the results were reflected about the diagonal. The lower left quadrant (target-target pairings) again has relatively lower values while the the lower right quadrant (target-clutter pairings) has higher values. While this distinction is not as noticeable as in the previous experiment, there still seems to be some class separation. The increased variability seen in the target-target pairings could be due to the fact that the type of targets used in experiment two are not as uniform as in the previous experiment. With these similarity matrices, we can explore the attributes that are used by the subject in the perception of these sounds. Note that for system A only the clutter-clutter similarity pairings will be used in the analysis shown in this paper.



.Fig.2 Aural similarity matrix for both targets (echoes 1-25) and clutter (echoes 26-50) from experiment 2.

3 Perceptual Kernel Identification

Pervious methods for identifying perceptual signal information have required the use of multidimensional scaling. This technique projects a similarity matrix onto a low dimensional Euclidian space. Signal features are then correlated to this space to assess perceptual relevance. Any feature that correlates well to this space is said to be perceptually relevant. The requirement of first identifying a Euclidian space before features can be found adds an unnecessary layer of estimation. This section introduces an alternate approach whereby perceptual distance metrics are found by fitting directly to the similarity measures gathered from the listening experiments. We observe that the listener similarity matrix is a perceptual equivalent to matrices used in kernel methods for regression and classification such as kernel-based PCA [6] and Support Vector Machines [7]. We propose two methods for learning novel kernel function from perceptual similarity data.

3.1 Kernel Feature Space

Kernel methods use relational measures between data points for regression and classification instead of directly using signal features. These methods compute "similarity" between signals via a predefined kernel. These kernels are commonly defined as an inner product of the form

$$K\left(x_i, x_j\right) = \left\langle \phi(x_i), \phi(x_j) \right\rangle * \tag{1}$$

where $\phi(\cdot)$ is a mapping from input space to a feature space. The power of this approach is that the underlying feature space does not need to be defined explicitly, only the function that measures the relation between signals $K\left(x_i, x_j\right)$ is required.

Identifying an appropriate kernel function for a given problem is often a difficult task. There are many bivariate functions to choose from and there is no clear rule as to which will work best for a specific application. Three examples of commonly used kernels are polynomial kernels

$$K\left(x_i, x_j\right) = \left(\left\langle x_i, x_j \right\rangle + c\right)^p \tag{2}$$

hyperbolic tangent kernels

$$K\left(x_i, x_j\right) = \tanh\left(\kappa \left\langle x_i, x_j \right\rangle + \Theta\right) \tag{3}$$

and the radial basis kernel

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{4}$$

These kernels have all been shown to perform well in various studies, but the ultimate choice on which to apply remains trial and error.

Our perceptual model we used for feature identification fits directly into this framework of kernel based classification. We assume that a human listener extracts perceptually relevant acoustic cues (or features) from a signal whenever they are asked to make judgments in an experiment. While we do not have direct access to these features, we can record the similarity measures they provide in a listening experiment. By finding a functional form for these similarity judgments we can learn a perceptually appropriate kernel function without needing to know the exact acoustic cues that are being used. This kernel approach allows us to bypass estimating the perceptual space via MDS all together. In addition to this computational nicety, previous studies have indicated that humans may use relational measures rather than specific features when aurally classifying sounds [8].

3.2 Similarity Fitting

To learn a numeric function representing perceptual similarity, we follow our previous approach [3] by regressing over a functional model for similarity. We employ the following model

$$\hat{h} = \underset{h}{argmin} \left\| \delta\left(\mathbf{x}, \mathbf{x}'\right) - d_h\left(\mathbf{x}, \mathbf{x}'\right) \right\|^2 \tag{5}$$

where δ is a perceptual similarity matrix and d_h is a numeric similarity matrix. With this approach, our goal is to learn a numeric similarity measure between signals that is as close as possible to the similarity measures provided by human listeners. One way to view this is as a warping of a numeric feature space to a new space in which signals are arranged according to their perceptual similarity.

3.2.1 Linear Similarity Fitting

The goal of similarity fitting is to learn a distance metric d_h in a numeric feature space $\phi(\cdot)$ in which signals are separated according to perception. A linear model for d_h is

$$d_h\left(x_i, x_j\right) = \sum_{k=1}^{p} h_k \left| \phi_k\left(x_i\right) - \phi_k\left(x_j\right) \right|^2 \tag{6}$$

where h_k is a scaling factor for the kth dimension of the numeric feature space and p is the total number of dimensions. This model is a weighted Euclidian distance measure in which each feature dimension can be weighted according to their fit to perceptual similarity.

While this weighted Euclidian distance allows each feature dimension to be scaled, it does not allow for cross-terms between dimensions. Another L_2 model for similarity with freedom to scale cross-terms is

$$d_H\left(x_i, x_j\right) = \phi\left(x_i\right)^T \cdot H \cdot \phi\left(x_j\right) \tag{7}$$

This similarity model is based on the Mahalanobis distance, where H is a $p \times p$ scaling matrix. This model for similarity is a weighted inner product between feature vectors. If H is set to the identity matrix then this measure is a standard inner product. In contrast to Eq. (6), this model assumes that perceptual similarity is numerically assigned such that larger numbers equate to greater similarity.

Using the Mahalanobis distance model, we now wish to learn the weights H that provides the best fit to perception, according to Eq. (5). In order to learn these weights, we first define a data matrix of training data as

$$\mathbf{\Phi} = \begin{bmatrix} \phi^T\left(x_1\right) \\ \phi^T\left(x_2\right) \\ \vdots \\ \phi^T\left(x_N\right) \end{bmatrix}_{N \times p} \tag{8}$$

With this data matrix, we construct an $N \times N$ numeric distance matrix $d_H = \Phi H \Phi^T$. Substituting this distance matrix into Eq. (5) we get

$$\hat{H} = \underset{H}{argmin} \left\| \delta - \Phi H \Phi^T \right\|^2 \quad (9)$$

Using the Frobenius norm this minimization can be rewritten as

$$\hat{H} = \underset{H}{argmin}\ \mathrm{tr}\left\{ \left(\delta - \Phi H \Phi^T \right)^T \left(\delta - \Phi H \Phi^T \right) \right\} \quad (10)$$

where tr is the trace of the matrix. We can now solving for the weights H by taking the derivative of Eq. (10) with respect to H

$$\frac{d}{dH} = 2\Phi^T \Phi H \Phi^T \Phi - 2\Phi^T \delta^T \Phi \quad (11)$$

Setting Eq. (11) equal to zero and solving for H, we find

$$\hat{H} = \left(\Phi^T \Phi \right)^{-1} \Phi^T \delta \Phi \left(\Phi^T \Phi \right)^{-1} \quad (12)$$

The above derivation provides a nice closed form solution for learning a perceptually-driven metric space. The derived distance measure d_H can now be used as a kernel function in kernel based signal classification and regression.

3.2.2 Nonlinear Similarity Fitting

A linear warping of signals in a feature space is not always enough to accurately describe perceptual similarity. In this case a more complex nonlinear approach is required. As we do not have insight into all of the processes used by humans when judging perceptual similarity, it would be difficult to identify a specific nonlinear model to use. Therefore, we will take a nonparametric approach to nonlinear similarity fitting.

Assume we wish to measure the similarity between a new test signal $x*$ and training signal x_i. To measure this similarity, we first identify a set of training signals G that are the k-nearest-neighbors of $x*$. We will estimate the similarity between $x*$ and x_i using a weighted sum of nearest-neighbors

$$\hat{\delta}\left(x*, x_i \right) = \sum_{j \in G} \beta_j \cdot \delta\left(x*, x_j \right) \quad (13)$$

where β_j is a weighting coefficient for the j[th] nearest-neighbor. The estimate of similarity between $x*$ and x_j is therefore a weighted average of similarities between x_i and the nearest-neighbors to $x*$.

We choose the weighting coefficients such that the weights sum to one, making Eq. (13) a true average. We also choose the weights such that the closer the nearest-neighbors the greater the weight, according to the rule

$$\beta_j = \frac{\exp\left\{ -\alpha \cdot d^2\left(x*, x_j \right) \right\}}{\sum_{m \in G} \exp\left\{ -\alpha \cdot d^2\left(x*, x_m \right) \right\}} \quad (14)$$

where $d(\cdot,\cdot)$ is a numeric distance measure that is used to identify nearest neighbors and α is a scaling factor.

The advantage of this technique is that it does not impose any structure on the estimated similarity. It does not even require a numeric feature space ϕ. The technique only requires a distance measure between signals and it identifies which training signals are most like the test signal. It then leverages their known perceptual similarity ratings. The disadvantage of this approach is that it is dependent upon the training signal spanning the space of possible signals.

4 Results

In order to identify a functional distance metric that relates to perceptual similarity we first calculate the standard Euclidian distance between signals. This distance measure not only quantifies a baseline to compare any further results, but it also provides a starting point for our regression techniques. To measure the distances between signals, we first identify a simple yet descriptive feature set.

For this feature set we choose to extract local time moments over each subband of a spectrogram, $S(t,\omega)$. That is,

$$\left\langle t^n \right\rangle_\omega = \frac{\sum_t t^n S\left(t,\omega\right)}{\sum_t S\left(t,\omega\right)} \quad (15)$$

These moments describe the average time, duration, skew, etc of the echo at various frequency regions. We extract the first three moments, n = {1, 2, 3}, from a spectrogram with nine subbands. This results in a 27 dimensional feature space. The standard Euclidian distance is calculated from these features and used as a point of comparison to the metric spaces found using similarity fitting.

Mahalanobis Distance

To improve upon the Euclidian distance matrix, we seek to identify a perceptually appropriate Mahalanobis distance measure. Our Mahalanobis regression technique requires that our similarity matrix normalized and transformed such that larger numbers equate to a greater degree of similarity. To do this, we simply take one minus the values of a normalized similarity matrix derived from those shown in Figure 1 and 2. Next we construct a data matrix Φ using the local moments calculated from the clutter echoes. With this data matrix, a scaling matrix $\hat{H}$ can be found via Eq. (12).

Nearest-Neighbor (NN)

Another approach to improving upon the Euclidian distance metric is Nearest-Neighbor distance regression. To identify a similarity matrix for the clutter signals using this approach we employ a leave-one-out cross validation strategy. First we designate a testing signal from the total set of signals. Nearest-neighbors are then calculated from the feature space of local moments. Next, the test echo's similarity to all other echoes is calculated using Eq. (13). This process is repeated for every clutter echo in the dataset.

To compare each of these approaches we calculate the correlation between perceptual similarity and distances calculated using the Euclidian, Mahalanobis and Nearest-Neighbor metrics. This comparison between distance measures is referred to as the alignment between matrices [9].

Table 1 summarizes the results from both experiment 1 and 2. The table shows the alignment between the perceptual similarity matrices and the three distance measures. In both

experiments the Mahalanobis and Nearest-Neighbor distance measures improve the alignment over a standard Euclidian distance. These new distance measures provide a better functional representation of perception. In both cases the Mahalanobis approach achieved the best result.

	Euclidian	Mahalanobis	NN
Experiment 1	0.34	0.84	0.54
Experiment 2	0.61	0.93	0.81

Table 1 Alignment between the perceptual similarity and numeric distance matrices.

5 Conclusions

Kernel methods for regression and classification describe signals not based on a set of features, but on relational measures between signals. In this paper, we demonstrated how this description of signals is analogous to the results of perceptual similarity experiments that are commonly used to identify signal features. We then proposed two methods for learning a perceptual kernel (distance metric) that depict how humans relate signals to one another. Mahalanobis distance regression identifies this kernel function using a linear parametric regression while Nearest-Neighbor distance regression identifies the kernel using a completely nonlinear nonparametric approach. These methods transform a standard feature space such that signals are arranged according to perceptual distances. This approach provides a framework for uncovering new perceptually-inspired kernel functions from listening experiment data.

References

[1] H. Terasawa, M. Slaney, and J. Berger, “Perceptual distance in timbre space,”in International Conference on Auditory Display, 2005.

[2] J. M. Grey, “Multidimensional perceptual scaling of musical timbres,” Journal of the Acoustical Society of America, vol. 61, 1977.

[3] •S. Philips and J. Pitton, “Perceptual feature identification for active sonar echoes,” IEEE Oceans, 2006.

[4] J. Pitton, J. Ballas, S. Philips, L. Atlas, D. Brock, M. Miller, and B. McClimens, “Aural classification of impulsive-source active sonar echoes,” Journal of the Acoustical Society of America, vol. 119, p3394, 2006.

[5] F. W. Young, C. H. Null, W. Sarle, and D. L. Hoffman, Proximity and preference: Problems in the multi-dimensional analysis of large data sets. Minneapolis, MN: University of Minnesota Press., 1981, ch. Interactively ordering the similarities among a large set of stimuli.

[6] B. Schölkopf, A. Smola, and K.-R. Muller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Computation, vol. 10, pp. 1299–1319, 1998.

[7] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[8] R. L. Goldstone and A. W. Kersten, Comprehensive Handbook of Psychology. New Jersey: Wiley, 2003, vol. 4, ch. Concepts and Categorization, pp. 599–621.

[9] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, “On kernel-target alignment,” in Neural Information Processing Systems, 2001.