Integrated Models of Signals and Background for an HMM/Neural Net Ocean Acoustic Events Classifier

William Y. Huang
Code 535
Marine Sciences and Technology
Naval Ocean Systems Center
San Diego, CA 92152-5000

Richard C. Rose
Speech Systems Technology
Lincoln Laboratory, M.I.T.
Lexington, MA 02173

Abstract

This paper investigates the use of Hidden Markov models (HMM's) for the classification and detection of ocean acoustic events in a nonstationary ocean background. A statistical formalism is described for integrating models for dynamic acoustic events and ocean background into a unified statistical framework. In this framework, both signal processes and background processes are modeled as HMM's, and signal classification is performed by obtaining the likelihood of a corrupted observation sequence through a combined state space of signal and background. Techniques are presented for estimating the acoustic event model parameters from training exemplars that are observed in these difficult background conditions. Finally, a novel neural network technique is proposed for the automatic learning of the nonlinear mechanism through which signal and background observations interact. Experimental results are presented.

1 Introduction

The ocean acoustic events that are of interest in this work can generally by characterized as short duration non-stationary events whose spectral energy evolve according to some characteristic temporal structure. The detection and classification of these acoustic events in an ocean environment is complicated by the presence of background signals that are not well modeled as traditional wideband or impulsive noise processes. In fact, the ocean background may itself contain acoustic events which are similar in nature to those events that we are trying to detect. Existing techniques that have been developed for ocean signal classification do not explicitly account for the desired signal having been observed in this difficult ocean background environment [1, 2, 3, 4, 5]. Failure to do so, however, can result in severe performance degradation, especially when a significant mismatch in the background characteristics exists between the training and testing of the classifier.

The principal motivation for applying HMM techniques to classification and detection of acoustic events is that they provide a means for temporal integration of short-time frame based spectral measurements. When temporal information is an important part of the signal representation, as is the case in ocean acoustic events [6], frame based static classifiers can provide poor event classification performance. This point was illustrated by a study comparing the performance of selected static pattern classifiers in classifying vowels sounds as spoken by a large population of speakers[7]. It was often the case in this study that a classifier that achieved a very low classification error rate when classifying independent speech frames, achieved a very high error rate when classifying the entire utterance. Hence, if the signal model suffers from an impoverished representation of temporal information in a signal, the performance of the resulting classifier is also likely to suffer. This issue has been addressed in previous work by applying heuristic rules [8] or neural networks with time delayed inputs [9, 2]. Recently, hidden Markov models have been applied with some success to the problem of ocean acoustic event classification [10, 11].

The principal contribution of this paper is to extend the definition of the HMM to incorporate the effects of ocean background. By modeling component sources, the system developed here provides a robust way to train and classify signals under differing background conditions. In Section 2, the form of the model is introduced. In Section 3 a maximum likelihood technique for estimating the acoustic event model parameters is introduced, along with a proposed hybrid neural network approach for estimating the process of signal corruption by background. Finally, in Section 4 a set of experiments is performed to evaluate the effectiveness of the approach in detecting an ocean acoustic event in the presence of actual ocean background.

2 Modeling Assumptions

We define an ocean acoustic event $a^i \in \{A\}$, $i = 1, \ldots, M$, taken from a set of possible events $\{A\}$. It is assumed that an acoustic event is produced with prior probability $P(a)$ and that there is an acoustic channel which produces D dimensional signal vectors, $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T)$ with probability $P(\vec{x} \mid a)$. We depart from the traditional HMM model development by assuming that the signal vectors are observed in the presence of of an ocean background process b which gives rise to D dimensional background observation vectors Z. The output sequence $Y = (\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_T)$ is then observed as a component-wise function of signal

1058-6393/91 $01.00 © 1991 IEEE

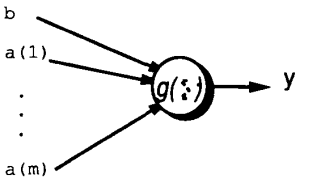

Figure 1: The observed signal, y, is a composite of signal from a background, b, and M signal generators, a^1, a^2, ..., a^M

and background, $y_t = g(z_t, x_t)$. We consider those functions $g(\cdot)$ for which the equation $y = g(z, x)$ define a one dimensional contour in the $x - z$ plane.

Both the signal and the background processes are represented by HMM models. The choice of the topology of the HMM models that are used for signal and background was made experimentally. The background HMM is a 4 state fully connected model, and the signal event model is a N state left–to–right model containing $N - 1$ non–null states and a "null state" that always emits a zero, to account for the "no signal" condition. The definition of this null signal state allows the background model parameters to be estimated from those observations that have been classified as representing background and not signal. The observation probabilities for all states, both signal and background, consist of single Gaussian densities.

The goal in acoustic event classification is to choose that event $\hat{a}$ by maximizing $P(a|Y)$, which from Bayes rule,

$$P(a \mid Y) = \frac{P(Y \mid a)P(a)}{P(Y)}, \qquad (1)$$

is equivalent to maximizing $P(Y \mid a)P(a)$. Estimating $P(Y \mid a)$ is accomplished using a probabilistic HMM to represent the acoustic event a. The prior probability $P(a)$ is estimated from higher level non–acoustic source of knowledge. These higher level hierarchical sources of knowledge have been shown in [6] to be critical in acoustic event classification by humans. The success of HMM's in continuous speech recognition is partly attributable to the ability of HMM's to combine these hierarchical sources of knowledge. It is expected that HMM's will provide similar benefits in the area of ocean acoustic event classification.

3 Estimating Model Parameters

3.1 Maximum Likelihood Formulation

The ML parameter estimation employed here is based on Rose et al. [12] and is similar to approaches taken in [13] and [14]. The noise corrupted observations Y arise from underlying state sequences $I = (i_1, i_2, \ldots, i_t)$, of the signal HMM, and $J = (j_1, j_2, \ldots, j_T)$ of the background HMM. The likelihood of the output sequence given signal model λ, which consists of the mean $\vec{\mu}_k$ and standard deviation $\vec{\sigma}_k$ of the Gaussian HMM observation probabilities $p_a(\vec{x}_t \mid k)$ and transition probabilities $p_a(k \mid l)$ is given as

$$P(Y \mid \lambda) = \sum_I \sum_J \oint_C P(X, Z, I, J \mid \lambda) dX dY \qquad (2)$$

where the summation is over all possible state sequences in the signal–background state space, and the notation $\oint_C$ refers to the integral along the contour C_t in the signal background observation space defined by $y_t = g(x_t, z_t)$. Expanding the observation probability in terms of the joint probability of the hidden state sequences and hidden data sequences allows us to isolate pertinent terms relating to the signal model parameters. The complete data likelihood in Eq. 2 is given as

$$P(X, Z, I, J \mid \lambda) = \prod_{t=1}^{T} p_a(i_{t+1} \mid i_t) p_a(\vec{x}_{t+1} \mid i_t) p_b(j_{t+1} \mid j_t) p_b(\vec{z}_{t+1} \mid j_t) \qquad (3)$$

Given an initial estimate of the acoustic signal model parameters λ, and following the method of Baum et al. [15] it is possible to find a new set of model parameters $\hat{\lambda}$ such that $P(Y \mid \hat{\lambda}) \geq P(Y \mid \lambda)$. This is done by maximizing the auxiliary Q function

$$Q(\lambda, \hat{\lambda}) = \sum_I \sum_J \oint_C P(X, Y, I, J|\lambda) \log P(X, Y, I, J|\hat{\lambda}) dX dY. \qquad (4)$$

In our simulations the probabilities obtained in the forward backward algorithm were replaced by a state sequence produced by the process of Viterbi training. Viterbi decoding in this context involves selecting the single path through the signal–background state space illustrated by the diagram in Figure 2 that maximizes $P(Y \mid \lambda)$. In this case, the summation over all possible state sequences in Eq. 2 is replaced by a max over all I and J.

Space does not permit a detailed description of the steps leading to the expressions for the ML parameter estimates. Taking the partial derivative of Equation 4 with respect to the signal component mean yields the estimate

$$\hat{\mu}_{a_i,ML} = \frac{\sum_t \sum_j P(i_t, j_t|Y, \lambda) E(\vec{x}_i|\vec{y}_t, i_t, j_t, \lambda)}{\sum_t \sum_j P(i_t, j_t|Y, \lambda)}, \qquad (5)$$

where

$$E(\vec{x}_t|\vec{y}_t, i_t, j_t, \lambda) = \frac{\oint_C \vec{x}_t p_a(\vec{x}_t \mid i_t) p_b(\vec{z}_t \mid j_t)\, d\vec{x}_t d\vec{y}_t}{\oint_C p_a(\vec{x}_t \mid i_t) p_b(\vec{z}_t \mid j_t)\, d\vec{x}_t d\vec{y}_t}, \qquad (6)$$

and $P(i_t = i, j_t = j \mid Y) = 1$ if the optimum Viterbi path passes through states i and j at time t, and equals 0 otherwise.

The contour integral depends on the definition of the noise corruption function $g()$. While the choice

this function can be very general, not all choices of $g()$ will lead to a contour integral that has a closed form solution. For instance if $g()$ represents an additive function of signal and background

$$g(x, z) = x + z, \tag{7}$$

then, using the notation x to refer to a single component of the vector $\vec{x}$, it can be shown that $E(x \mid y)$ of Eq. 5 is

$$E(x|y) = \frac{\sigma_a^2}{\sigma_b^2 + \sigma_a^2}\left[y + \left(\frac{\sigma_b^2}{\sigma_a^2}\mu_a - \mu_b\right)\right]. \tag{8}$$

3.2 A Neural Net Mapping Function

For spectral parameters, the additive assumption of Eq. 7 is inaccurate. If the observations are averages of N magnitude squared spectral frames, then the noise corrupted observation probability is non-central χ^2 distributed with $2N$ degrees of freedom. and parameters $\sigma^2 = \sigma_b^2 + \sigma_a^2$ and $\delta = \mu\sqrt{2N}$, where $\mu = \mu_b + \mu_a$ [16]. No closed form solution for $E(x|y)$ has been found for this case; nor have closed form solutions been found for other important noise corruption functions. Numerical evaluation of the integrals required for $E(x|y)$ can involve costly series expansions. Reasonably accurate approximations can be obtained using the trapezoidal rule and approximating the non-central χ^2 density with 50 summands [16]. However, it may be simpler and faster to approximate $E(x|y)$ by training the parameters of a general mapping function using actual observations. Neural nets can be trained using a minimum mean squared error criteria to approximate a mapping from clean to noise corrupted observations that converges with infinite data to $E(x|y)$ [17]. This is true for a broad class of networks including the Multi-Layered Perceptron (MLP) trained using back-propagation, and the Radial Basis Function (RBF). This suggests the use of an HMM-neural net hybrid system where the HMM provides the temporal decoding and the neural network performs the mapping from the space of uncorrupted signal vectors to the space of observable noise corrupted observation vectors.

3.3 Algorithm Summary

The Viterbi lattice that results from the composition of signal and background HMM's is multi-dimensional (Fig. 2). Viterbi decoding on a multi-dimensional lattice is a straight-forward extension of the standard 2-D case. The complete algorithm, summarized in Fig. 3, has one Expectation Maximization (EM) estimator embedded in another. Convergence is guaranteed since $P(Y)$ cannot decrease at any step in Fig. 3, by virtue of the properties of Viterbi decoding and ML estimation.

However, if neural nets are used to estimate $E(x|y)$, convergence in the inner EM estimator is not guaranteed. The system illustrated in Fig. 4, simulates the ocean environment internally during training. The internal simulation is necessary to generate training data for the neural net mapping function. Neural net training for estimating $E(x|y)$ has the problem shown

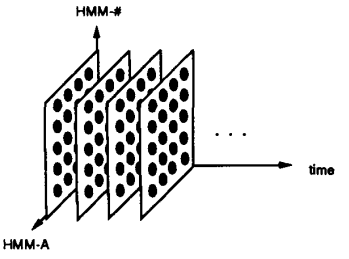

Figure 2: Multi-dimensional Viterbi lattice. In this example, an HMM for background (labeled b), containing 4 nodes, and an HMM for source a, with four nodes, formed a Cartesian product with 16 possible states (or state combinations) per time frame. Viterbi decoding on this lattice is equivalent to decoding on a standard 2-D lattice with 16 nodes.

1. Estimate $P(i_t, j_t \mid Y)$ via Viterbi decoding
2. Estimate the probability density functions for signal and background
   (i) Compute $\hat{\mu}_{b,ML}$, $\hat{\sigma}_{b,ML}$ for decoded "non-signal" observations
   (ii) Compute $\hat{\mu}_{a,ML}$, $\hat{\sigma}_{a^{(1)},ML}$
3. Obtain new mapping function based on $g()$ and new source densities
4. Go to 2 until convergence in source densities
5. Go to 1 until overall convergence is achieved

Figure 3: Training algorithm summary.

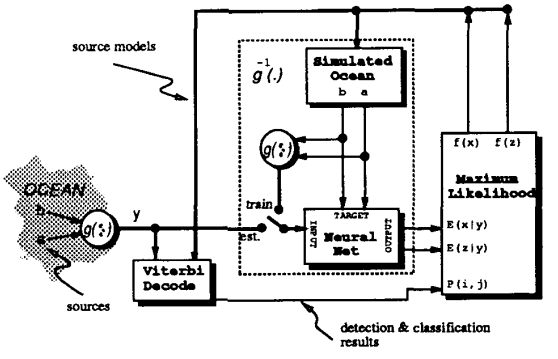

Figure 4: Algorithm of the HMM-neural net hybrid system. The input to the neural net switches to simulated ocean during neural net training, and to observed ocean when estimating source model parameters.

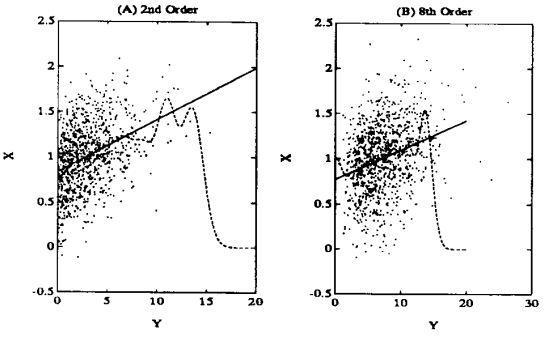

Figure 5: Scatterplot for 1000 samples of signal (a) vs. its (A) 2nd order non-central χ^2-distribution (a single frame of magnitude-squared spectra), and (B) 8th the order case (4 spectral frame averages). The corresponding estimate of $E(x|y)$ are plotted for the LMS (solid line) and RBF (dashed line) techniques. $\mu_a = 1$, $\sigma_a^2 = .14$, $\mu_b = .3$, $\sigma_b^2 = .5$

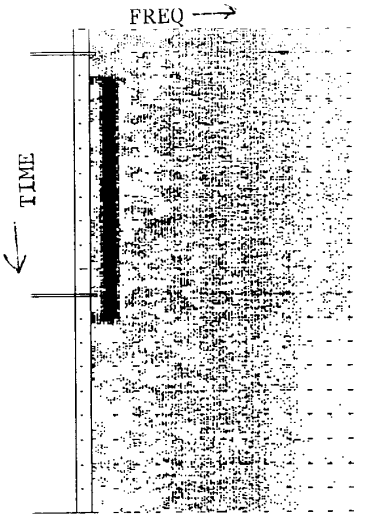

Figure 6: Time vs. frequency (via wavelet decomposition) for a sample in the test set. Horizontal ticks mark .1 sec intervals. Solid horizontal long-dashes on the left side are the Viterbi decoder's estimates of the starting and stopping boundaries for signal "A" (a 10 msec broadband event).

Table 1: Summary of experiments performed on additive normal components. Results labeled with "LMS" were obtained using neural net.

TRAINING SNR	*GAIN*	*FALSE ALARM RATE* (sec^{-1})	*MISS (% ERROR)*
high	0	0	0%
low	0	.041	0%
low (LMS)	0	0	0%
high	-6dB	.123	5%
low	-6dB	.165	20%
low (LMS)	-6dB	.082	20%

in Fig. 5. The neural net does not perform well in regions with limited training data. These regions from a mismatch between the real the simulated ocean. This is illustrated by the noise in the Radial Basis Function (RBF) output for large y's in Fig. 5. For large values of y, $E(x|y)$ is nearly linear. A linear mapping, implemented as using the LMS algorithm, does a good job in this case, and in the case of approximating Eq. 8, which is linear to begin with. However, the linear estimator is inaccurate for small values of y, especially in the second order case in Fig. 5, where the RBF correctly estimated a slight downward nonlinearity. The linear LMS estimator is used for the work presented here.

4 Experimental Results

The purpose of the preliminary synthetic data experiment described in this section is to validate the algorithms presented above. The DARPA Standard Phase I database is designed to test conventional classifiers. It does not have the adverse background conditions that this algorithm is designed to handle. Therefore, an adverse condition testing set was constructed by mixing different parts of the Phase I dataset. The Phase I signals used were signal "A", a 10msec wideband pulse, "E", a 1 second low frequency tonal, and the "quiet ocean" background. The testing set consists of 24.3 seconds of data, containing 10 samples of E, and 20 samples of A. The E samples overlapped A samples 9 times, to simulate an adverse, highly nonstationary background. The signals were mixed into the background with a gain of either 0 or -6dB. Fig. 6

shows a sample from this testing set. The solid horizontal long–dashes at the left are the Viterbi decoder estimates of the starting and ending times for signal A events that it detected. The Viterbi decoder used models that were trained either offline, in a high SNR and no interference condition, or in adverse conditions that are similar to the test set.

To evaluate the robustness of the neural net approximation of $E(x|y)$, relative to the closed form solution, this first set of experiments involve mixing signals in background *after* spectral decomposition. This was necessary since Eq. 8 works only for additive normal components, and it is found that spectral components are not well approximated by Eq. 7.

Tab. 1 summarizes the results on this database. Models were trained either under similar adverse conditions, or they were trained with no noise and with no other signal (high "training SNR" in Tab. 1). The results in Tab. 1 show that it is possible to (1) train models of signal A under adverse background conditions, and to (2) train models of A under high SNR conditions, and then detect and classify A under adverse background conditions.[1]

5 Summary

The techniques presented here uses Hidden Markov Models and the Maximum Likelihood (ML) formalism to address the issues of temporal structures and nonstationary backgrounds. Temporal structures provide an important cue for human acoustic events classifiers [6]; but is not well exploited by static, frame–based classifiers. HMM's can provide a succinct model for temporal structures. The ML technique integrates HMM models of component processes to provide a robust way to handle highly nonstationary background. In the experiments presented here, signal models were trained offline under high SNR conditions, and then used to detect and classify signal events under an adverse, highly nonstationary background. The experiments also demonstrates an ability to train for signal parameters when a training set of isolated events is not available.

References

[1] S. Beck, L. Deuser, R. Still, and J. Whiteley, "A hybrid neural network classifier of short duration acoustic signals," in *Internation Joint Conference on Neural Networks*, pp. I–119–124, IEEE, July 8–12 1991.

[2] L. I. Perlovsky, "Model based classification of transient signals using the MLANS neural network," in *IEEE Conference on Neural Networks for Ocean Engineering*, pp. 239–246, 1991.

[1] Tab. 1 shows that the use of neural net (LMS) to estimate $E(x|y)$ lead to better results than when the closed form solution is used. This is probably a procedural artifact: The closed form system was trained to five iterations, yielding the results reported in Tab. 1. Neural net training was initialized with the final parameters from the closed form training. $P(Y)$ increased during each of 5 neural net iterations. Therefore, the results reported for neural net training is slightly better. Tests using Radial Basis Function (RBF) and Multi–Layered Perceptron (MLP) had very high false alarm rates, for reasons discussed above.

[3] J. C. Solinsky and E. A. Nash, "Neural–network performance assessment in sonar applications," in *Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering*, pp. 1–12, IEEE, August 1991.

[4] D. Montana and K. Theriault, "Neural–network–based classification of acoustic transients," in *Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering*, pp. 247–254, IEEE, August 1991.

[5] F. L. Casselman, D. F. Freeman, D. A. Kerrigan, S. E. Lane, N. H. Millstrom, and W. G. Nichols, "A neural network–based passive sonar detection and classification design with a low false alarm rate," in *Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering*, pp. 49–55, IEEE, August 1991.

[6] J. H. Howard and J. J. O'Hare, "Human classification of complex sounds," *Naval Research Review*, vol. xxxvi, pp. 26–31, 1984.

[7] W. Huang and R. Lippmann, "HMM speech recognition with neural net discrimination," in *Advances in Neural Information Processing Systems 2* (D. S. Touretzky, ed.), pp. 194–202, San Mateo, CA: Morgan Kaufmann, 1990.

[8] D. W. Cottle and D. J. Hamilton, "All neural network sonar discrimination system," in *Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering*, pp. 13–19, IEEE, August 1991.

[9] Y.-H. Pao, T. L. Hemminger, D. J. Adams, and S. Clary, "An episodal neural–net computing approach to the detection and interpretation of underwater acoustic transients," in *Proceedings of the IEEE Conference on Neural Networks for Ocean Engineering*, pp. 21–28, IEEE, August 1991.

[10] M. K. Shields and C. W. Therrien, "A hidden Markov model approach to the classification of acoustic transients," in *Procedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 2731–2734, April 1990.

[11] J. P. Woodard, "Modeling and classification of acoustic transients by speech recognition techniques," *Journal of Underwater Acoustics*, Oct 1990.

[12] R. Rose, J. Fitzmaurice, E. Hofstetter, and D. Reynolds, "Robust speaker identification in noisy environments using noise adaptive speaker models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, May 1991.

[13] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions in ASSP*, vol. 37, pp. 1495–1503, 1989.

[14] A. P. Varga and R. E. Moore, "Hidden Markov model decomposition of speech and noise," in *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1990.

[15] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1972.