Signatures of Walking Humans from Passive and Active Acoustic Data using Time-Varying Vector Autoregressions

Melanie B. Rudoy, Charles E. Rohrs, Jingdong Chen
Massachusetts Institute of Technology
Digital Signal Processing Group
77 Massachusetts Avenue, Cambridge, MA 02139
{mbs, crohrs, jc455}@mit.edu

***Abstract*— A sensor fusion framework for characterizing the signature of walking targets using data collected from passive acoustic and active ultrasound sensors is investigated. We compute local estimates of the acoustic energy of the footsteps and the velocity of the torso and limbs. A time-varying vector autoregression (TV-VAR) is used to model the evolution of these signals, and captures the physical correlations between them, creating a natural data fusion across different sensor modalities. The signature is defined as a subset of the parameters from the TV-VAR model, and the quality of this feature set is evaluated using a support vector machine framework to classify multiple test subjects for both detection and discrimination applications.**

I. Introduction

The ability to detect moving humans is crucial in many applications, ranging from building surveillance and automotive safety systems, to situational awareness enhancement in military applications. Video monitoring and machine vision methods are among the best of current approaches, but require careful calibration, high power, high memory, high cost, and may require environments with adequate lighting for effective use. On the other hand, acoustic sensors are often cheaper, require less power and data storage, can be operated in total darkness, and are rapidly deployable, making them an appealing alternative. In this paper, we investigate the use of passive and active acoustic sensors to detect and identify walking subjects.

It has been previously demonstrated that it is possible to characterize the manifestation of human footsteps in data collected by passive acoustic (e.g., microphone) sensors. In [6], a set of features is given, including the mel-cepstral peak frequencies, walking interval period and the footstep power spectral envelope. This work uses a single microphone, and as a result the performance is sensitive to background noise that may mask the sound of the footsteps. In [1], data from both passive acoustic and seismic (e.g., accelerometer) sensors are considered. Seismic sensors are robust to many types of background interference, such as people talking or street noise. When the seismic data are modeled using a time-varying autoregressive (TVAR) model, features based on the TVAR parameters form clusters corresponding to different positions of the walker's foot (e.g., contact of heel to ground). The temporal trajectories of these features are periodic, with the period corresponding to the time between successive footsteps. While the work in [1] considers two different sensor modalities, the information from each sensor is combined at the inferential, not feature, level. The use of TVAR models to define acoustic signatures is also seen in [2], where a scalar AR process is used to analyze the signature of moving vehicles from passive acoustic sensors.

This paper addresses the limitations of these previous approaches by fusing data from passive acoustic and active ultrasound sensors, in order to find a robust characterization of the signature of a walking subject. As described in Section II, our approach is based on jointly modeling a set of signals derived from two different sensor modalities using a time-varying vector autoregressive (TV-VAR) process. Details relating to the data preprocessing and TV-VAR parameter estimation are given in Section III. A subset of the TV-VAR parameters are used to define the signature of a walker, and this signature is evaluated through the use of a support vector machine (SVM) based classification system, as described in Section IV. Section V presents experimental results from both detection and discrimination test scenarios, using two representative test subjects.

II. Problem Statement and Model

We consider the problem of defining the signature of a walking human using both passive and active acoustic sensors. The data used in this analysis were collected at the National Center for Physical Acoustics at the University of Mississippi [5]. During the test scenario, the target walks down a hallway containing co-located active ultrasound and passive acoustic sensor devices, as shown in Figure 1. While data are available for multiple human and animal test subjects, this paper focuses on only two representative data sets, corresponding to an adult male and a dog.

The passive sensor operates at a resonant frequency of 25 kHz, with a sampling frequency of 96 KHz. The signal

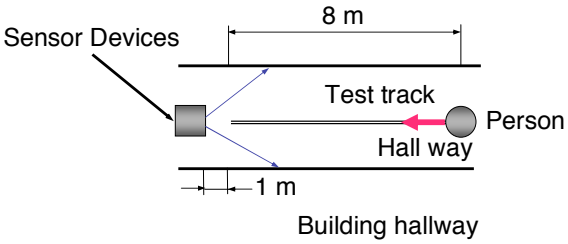

Fig. 1. Experimental Setup at the National Center for Physical Acoustics, University of Mississippi.

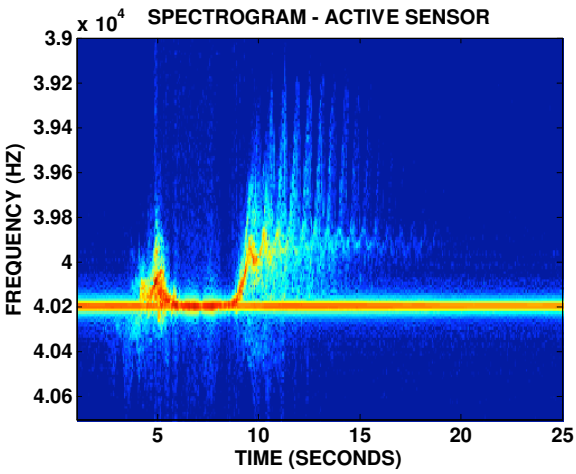

Fig. 2. Typical spectrogram of active sensor data for human test subject. A single, continuous tone is transmitted, and the Doppler shift induced by the walking target is measured.

of interest, corresponding to the sound made by a foot upon impact with the floor and subsequent friction created when lifting the foot for next step, is found in the $20-30$ kHz band. The active sensor operates at a resonant frequency of 40 kHz, with a sampling frequency of 96 KHz. A continuous tone is transmitted at $f_c = 40$ kHz, and the return signal containing a Doppler shift, Δf, induced by the walker is measured. Due to the natural swaying of the torso and swinging of the limbs while walking, coupled with the fact that all points within a body segment do not move at a constant velocity, one expects to see a band of Doppler shifts following a sinusoidal motion in the spectrogram, as confirmed by Figure 2.

We model the state of the system as consisting of two variables representing the velocity of the target, and one variable corresponding to the acoustic energy of the actual footstep. Let the average velocity of the torso be denoted as v_t, the average velocity of the limbs as v_l, and the total energy of the footstep sound as e, so that the state at each short-time segment is given by:

$$\mathbf{x}_n = \begin{pmatrix} v_t[n] & v_l[n] & e[n] \end{pmatrix}^T. \tag{1}$$

The state evolution is modeled as a second-order, time-varying vector autoregressive (TV-VAR) process, defined by:

$$\mathbf{x}_n = \mathbf{A}_{1,n}\mathbf{x}_{n-1} + \mathbf{A}_{2,n}\mathbf{x}_{n-2} + \boldsymbol{\epsilon}_n, \tag{2}$$

where the matrices $A_{j,n}$ encode the coupling between the j^{th} lag of $\mathbf{x}_n$ with the current state vector. The additive noise term, $\boldsymbol{\epsilon}_n$, is assumed to be a zero-mean, Gaussian random variable, with covariance $\boldsymbol{\Sigma}_n$. The TV-VAR process is chosen in order to capture both the autocorrelation within each signal, as well as the cross-correlations among them. These correlations exist due to the fact that when a person walks, the torso and limbs do not move independently of one another, and this motion is physically correlated with the rate and timing of the footsteps captured by the passive sensor.

III. Data Analysis and Model Estimation

Given $v_t[n]$, $v_l[n]$, and $e[n]$, the parameters of the TV-VAR model defined in Equation 2 are estimated on a segment by segment basis, using a sliding window of 50 samples of $\mathbf{x}_n$, with 50% overlap. For each segment, the parameters $\{\mathbf{A}_{1,n}, \mathbf{A}_{2,n}, \boldsymbol{\Sigma}_n\}$ are computed using the standard maximum likelihood estimation procedure, as described in [3]. Given these parameter estimates, the value of $\mathbf{x}_n$ at each short-time frame is computed as described below.

The passive and active data are non-stationary due to the motion of the subject and variable distance from the sensors, but were found to be locally stationary over short-time frames of 53 ms, motivating use of a short-time Fourier analysis. The data from both sensors are appropriately bandpass filtered, demodulated, and downsampled to 19.2 kHz. From the passive data, we compute $e[n]$, the log of the total energy, using 53 ms frames and 50% overlap, as:

$$e[n] = \log\left(\sum_k |P(n,k)|\right),$$

where $P(n,k)$ is the Fourier transform of the passive signal data over the n^{th} time window, in the k^{th} frequency bin. This operation preserves the periodic and impulsive nature of the passive acoustic data, as shown in Figure 3 for typical data.

Envelopes of the torso and limb Doppler shifts in the time-frequency plane are extracted from the data collected by the active sensor. First, the average torso Doppler shift, Δf_t, is computed using a combination of adaptive thresholding and linear regression on the spectrogram image. Specifically, a bandpass filter is applied in order to restrict attention to the area either strictly above or below the carrier, corresponding to times when the target is moving towards or away from the sensors. For each frame, a binary mask is created using a thresholding algorithm, so that only the top percentile of the spectrogram magnitude data per time frame are assigned a weight of unity. The time-frequency bins that pass the threshold test are used to fit a piecewise linear curve using ordinary least squares regression. The slope of each line segment is computed over a window of five short-time frames, and a continuity constraint is imposed so that the

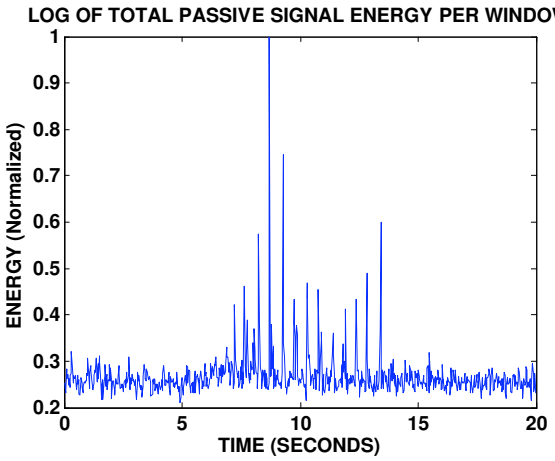

Fig. 3. Typical plot showing log of the total acoustic energy collected from the passive sensor in the band of interest ($20-30$ kHz), using 53 ms frames with 50% overlap.

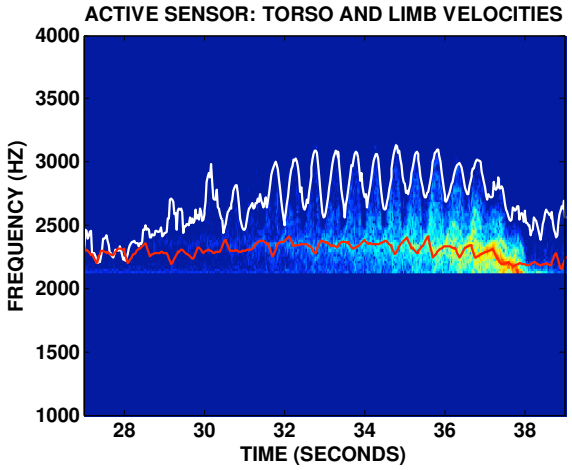

Fig. 4. Average torso (red) and limb (white) Doppler shifts, superimposed on spectrogram computed from representative data collected from active sensor.

ends of the line segments meet at the boundary between neighboring segments. The results of applying this algorithm on a representative data sample is shown in Figure 4. An estimate of the torso velocity, $v_t[n]$, is computed from the Doppler shift, $\Delta f[n]$, as

$$v_t[n] = \frac{c\Delta f_t[n]}{2f_c}, \tag{3}$$

where $c = 340$ m/s is the speed of sound.

Second, the average limb Doppler shift, Δf_l, is estimated by applying an edge detection algorithm to the spectrogram image. Again, attention is restricted to the area either strictly above or below the carrier. However, rather than using a relative threshold that changes frame to frame, a hard threshold of -70 dB is enforced. For each frame, a single boundary point is identified that represents either the smallest (when looking below the carrier) or the largest (above the carrier) frequency that meets the threshold requirement. The resulting curve is shown in Figure 4, and the average limb velocity, $v_l[n]$, is computed according to Equation 3.

IV. Classification

A. Support Vector Machine Based Classification

The quality of the TV-VAR based signature of a walking subject is investigated through the implementation of two classification systems. The first system detects the presence of a walking subject against the null hypothesis of no target present, while the second system discriminates among multiple targets. Both classifiers are implemented using an SVM. The SVM computes the optimal separating hyperplane to partition two classes of data, by maximizing the distance of the training points to the hyperplane. In the case where the data are not linearly separable, the SVM allows for some of the training points to be on the wrong side of the hyperplane through the inclusion of slack variables. For a detailed description of SVM-based classification, see [4]. Here, all training data are hand labeled as belonging to one of two classes and are assigned a label of ± 1, corresponding to the target being present or absent, respectively.

B. Feature Extraction

Once the TV-VAR parameters are estimated for each segment, a set of N representative features must be selected, which are used to train the classifier. Candidate features include the eigenvalues of the state transition matrix, or individual elements from the A_j matrices. The features are selected so that they result in classes that are linearly separable in the N-dimensional feature space. Here we consider the $N = 3$ features corresponding to $A_1(1,2)$, $A_1(2,2)$, and $A_1(3,3)$. These coefficients capture the autocorrelations within each signal and the cross-correlations across the time series.

V. Experimental Results

A. Detection

The effectiveness of the signature derived from the TV-VAR model parameters is demonstrated using one human and one non-human subject, corresponding to an adult male and a dog. For each test subject, the preprocessing and parameter estimation tasks described in Section III are performed, and the appropriate features are extracted, as detailed in Section IV-B. The labeled data are passed to the SVM algorithm, and the resulting hyperplanes are depicted in Figures 5 and 6. In both cases, the classes are linearly separable in the feature space. In order to validate the results, Leave One Out Cross Validation (LOOCV) is performed, and the results are given in Table I. In LOOCV, a single observation is withheld, and the system is trained using the remaining data [4]. The withheld sample is then used to test the classifier, and this process is repeated for each observation.

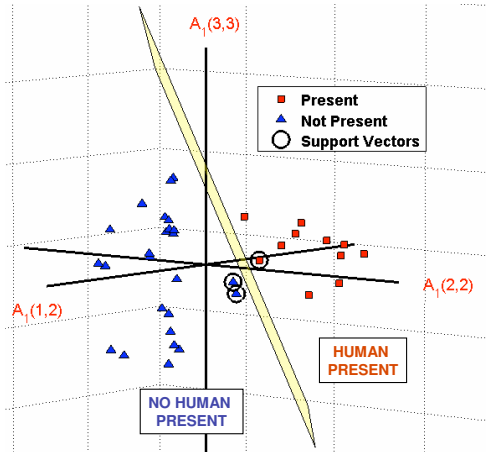

Fig. 5. SVM based classifier for detecting presence of a human test subject.

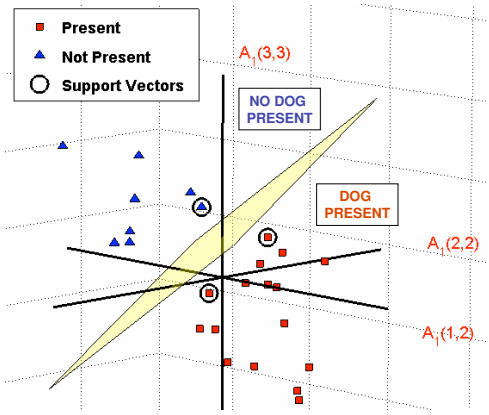

Fig. 6. SVM based classifier for detecting presence of a dog test subject.

B. Discrimination

The ability of the TV-VAR signature to discriminate among multiple subjects is demonstrated using the same data used in the detection experiments. The three decision regions are computed by first training the SVM to classify "Human Present" versus all other classes, followed by a second phase to classify "Dog Present" versus all other classes. The result is the creation of two hyperplanes that partition the data into four regions, as shown in Figure 7. The training points corresponding to times when the human is not present and times when the dog is not present tend to cluster spatially, forming a single class corresponding to "No Target Present". The fourth region corresponds to the overlap between the human and dog classes; test points that lie within this space are classified as belonging to the "Human Present" class if the distance from the test point to the "Human Present" hyperplane exceeds the distance from the test point to the "Dog Present" hyperplane. Cross-validation results are given in Table I.

Data Description	Percent Correct
Human - Detection	97.22%
Dog - Detection	100%
Human/Dog - Discrimination	95%%

TABLE I
LEAVE ONE OUT CROSS VALIDATION RESULTS.

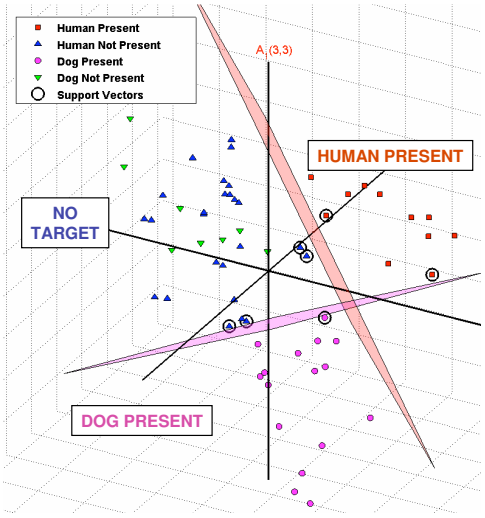

Fig. 7. SVM based classifier for human/dog discrimination.

C. Discussion

The results given in this paper represent preliminary attempts to characterize the signature of walking human and non-human test subjects, using both passive and active acoustic sensors. Clearly more work is needed to assess the usefulness of these techniques to differentiate among multiple humans or animals, or to detect a target from out-of-sample data. The main contributions of this work are the formulation of a TV-VAR model to capture the joint dynamics of the passive and active data, and the utilization of three specific features derived from this model for the design of classification systems for detection and discrimination applications.

ACKNOWLEDGMENTS

This research was supported by participation in the Georgia Institute of Technology MURI 2004 sponsored by the Army Research Office (ARO) under award W911NF-04-1-0190.

REFERENCES

[1] Bland, Ross. Acoustic and Seismic Signal Processing for Footstep Detection. Masters' thesis. Department of Electrical and Computer Engineering, Massachusetts Institute of Technology, June 2005.
[2] Eom, Kie B. Analysis of Acoustic Signatues from Moving Vehicles Using Time-Varying Autoregressive Models. Multidimensional Systems and Signal Processing, Vol. 10, 1999.
[3] Hamilton, James. Time Series Analysis. Princeton University Press, Princeton, NJ, 1994.
[4] Hastie, T., R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning. Springer, 2001.
[5] Sabatier, James M. Human Signatures for Personnel Detection. Multidisciplinary Research Program of the University Research Initiative (MURI). October, 2006.
[6] Shoji, Yasuhiro, Akitoshi Itai, and Hiroshi Yasukawa. Personal Identification Using Footstep Detection in In-Door Environment. IEICE Transaction Fundamentals, Vol. E88-A, No. 8, August 2005.