Multi-sensory Features for Personnel Detection at Border Crossing

Po-Sen Huang[1]*, Thyagaraju Damarla*[2]*, Mark Hasegawa-Johnson*[1]

[1]Beckman Institute, ECE Department, University of Illinois at Urbana-Champaign, U.S.A.
[2]US Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783, U.S.A.

huang146@illinois.edu, rdamarla@arl.army.mil, jhasegaw@ad.uiuc.edu

Abstract

Personnel detection at border crossing becomes an important issue recently. To reduce the false alarm caused by nonhuman animals or the existence of multiple objects, it is important to discriminate between humans and nonhuman animals. In this paper, based on phenomenology of the differences between humans and four-legged animals, we propose using enhanced autocorrelation pattern for feature extraction for seismic sensors, a multi-stage feature selection framework for acoustic sensors. Along with ultrasonic sensors, we use decision fusion for multi-modality fusion and use Gaussian Mixture Models for classification. From experimental results, we show that our proposed methods improve the robustness of the system.

Index Terms: Gaussian Mixture Models, sensor fusion, footstep detection, personnel detection

1. Introduction

Personnel detection is an important task for Intelligence, Surveillance, and Reconnaissance (ISR) requirements [1, 2]. One might like to detect intruders to a certain area during day and night so that the proper authorities can be alerted. For example, Homeland Security often requires detection of illegal aliens crossing the border. There are numerous other applications where personnel detection is important.

However, personnel detection is a challenging problem. Video sensors consume high amounts of power and require a large volume for storage. Hence, the emphasis is on non-imaging sensors, since they tend to use low amounts of power and are long-lasting, which are suitable for border crossing scenario. Moreover, the false-alarm caused by nonhuman objects or the existence of multiple objects makes personnel detection more challenging.

Traditionally, personnel detection research concentrated on using seismic sensors. When a person walks, his/her impact on the ground causes seismic vibrations, which are captured by the seismic sensors. Previous works have relied on the fundamental gait frequency gait frequency estimation [3, 4]. Hyung et al. proposed the method of extracting temporal gait patterns to provide information on temporal distribution of the gait beats [5].

At border crossing, animals such as mules, horses, or donkeys are often known to carry loads. Animal hoof sounds make them distinct from human footstep sounds. In particular, when humans and four-legged animals walk together, the sounds they make are still distinguishable from their combination. Similarly, in acoustic event detection, Zhuang et al. utilized the distinct characteristic of each event, using Perceptual Linear Predictive (PLP) as features, for detection [6, 7, 8].

Passive and active ultrasonic methods were proposed for the detection of walking personnel for ultrasound signals [9]. The passive method utilizes the footsteps' ultrasonic signals generated by friction forces while the active method used the human Doppler ultrasonic signature. In an outdoor scene, the passive ultrasound signals are limited in distance and are noisy. For the active ultrasound method, when a person walks, each limb is a compound pendulum and has distinct oscillatory characteristics which in turn results in the micro Doppler effect. Similarly, the torso also oscillates at a particular frequency. The ultrasonic sensors can detect the ultrasonic signature generated by footsteps and movements of the torso. Zhang et. al. reported the different micro-Doppler gait signatures between human and four-legged animals [10]. These arise from the different physical mechanisms found in the two different species. Kalgaonkar et al. analyzed spectral patterns to classify human walking (walker identification, approach v.s. away, male v.s. female) [11].

As shown in the above literature review, existing research only use a single sensor recored in clean environments with a single object (a person or a four-legged animal) walking. However, in reality, when there are many objects such as people or four-legged animals walking or running in noisy environments, it is difficult to distinguish from people and/or four-legged animals using a single sensor alone with previous approach.

In this paper, we propose a multi-stage acoustic features selection method and enhanced autocorrelation pattern for seismic feature extraction. Along with ultrasonic feature, we use the decision fusion to examine the robustness of our methods.

The organization of this paper is as follows: Section 2 introduces the multi-sensor multi-modality data and events. Section 3 discusses the feature extraction from acoustic, seismic, and ultrasonic sensors. Section 4 discusses Gaussian mixture models classifiers, and decision fusion. Section 5 shows the experiments on multi-sensor multi-modality dataset, followed by Section 6 discussion. We conclude this paper with future work in Section 7.

2. Data

In this paper, we use a multi-sensor multi-modality realistic dataset collected by U.S. Army Research Lab and University of Mississippi in Arizona. The data is collected in a realistic environment in an open field. There are three selected vantage points in the area. These three points are known to be used by the illegal aliens crossing the border. These places where the data is collected are: (a) wash (a flash flood river bed with fine grain of sand), (b) trail (a path through the shrubs and bushes of wild and (c) choke point (a valley between two hills.) The data is recorded using several sensor modalities, namely, acoustic, seismic, passive infra red (PIR), magnetic, E-field, passive ultrasonic, sonar, both infra red and visi-

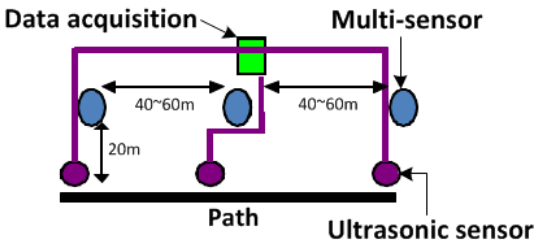

Figure 1: Sensor layout, where a multi-sensor multi-modality system has acoustic, seismic, passive Infra-Red (PIR), radar, magnetic, and electric field sensors

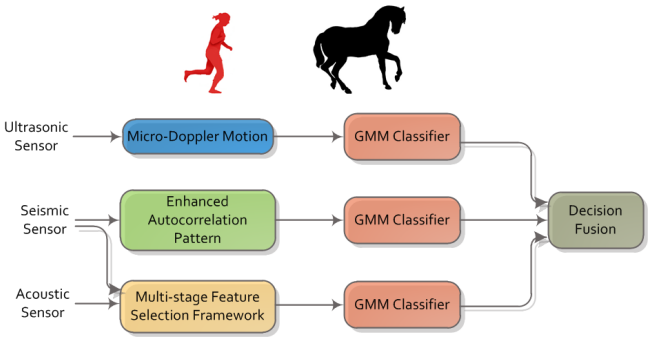

Figure 2: The overall flow: feature extraction based on phenomenology, GMM classifier, and decision fusion

ble video sensors. Each sensor suite is placed along the path with a spacing of 40 to 60 meters apart. The detailed layout of the sensors are shown in Figure 1. Test subjects walked or ran along the path and returned back along the same path.

A total of 26 scenarios with various combinations of people, animals and payload are enacted and collected the data at those three sites. We can categorize them as: *single person (11.6%), two people (13%), three people (21.7%), one person with one animal (14.5%), two people with two animals (15.9%), three people with three animals (17.4%), and seven people with a dog (5.9%)*, where the animals can be a mule, a donkey, a horse, or a dog, and number is the parentheses represents the percentage of the data. The data is collected over a period of four days; each day at a different site and different environment. Also, there is a variable wind in the recording environment.

2.1. Detection and classification

The time duration for subjects passing by is short (about ten to twenty seconds at a time) compared to the whole recording time (five to six minutes recording). Therefore, without the ground truth of the footsteps time, we would like to extract the time duration when test subjects passing through, similar to voice activity detector in speech processing. For acoustic sensors, in an outdoor scene, the signals are contaminated by wind sounds, human voices, or unexpected airplane engine sounds. Seismic and PIR sensors, on the other hand, are relatively clean. Hence, we use either seismic or PIR sensors by energy detection to determine the time duration when test subjects pass by (we use seismic sensors with ten seconds durations). For ultrasonic sensors, which are collected separately, we use energy threshold to determine the time duration when test subjects pass by (ten seconds duration). For each recording, there are two segment of signals (walked or ran along the path and returned back along the same path). In this paper, we emphasize on the classification between humans and humans with four-legged animals.

3. Features Extraction

Based on phenomenology of the differences (**micro-Doppler motion**, **enhanced autocorrelation pattern**, **footstep sound**) between humans and four-legged animals, we discuss the features from ultrasonic, seismic, and acoustic sensors. The overall flow is shown in Figure 2.

3.1. Seismic

Seismic sensors capture the vibrations in the ground caused by the motion of the targets or ground coupling of acoustic waves. There is a distinct feature between humans and four-legged animals - gait pattern. Previous approach does not consider the case for multiple human and/or four-legged animals [5]. When there are multiple human and/or four-legged animals, it is not reliable for the estimation of gait period based on single pitch (fundamental frequency) detection method [12, 13]. Inspired by Hyung's temporal gait pattern approach [5] and the progress in multipitch analysis [14], we propose a gait pattern based on enhanced autocorrelation, as shown in Figure 3.

We form analytic signals by Hilbert transform and then use full wave rectification followed by low-pass filtering and downsampling for envelope detection. Then, we use **enhanced autocorrelation** to estimate the gait pattern and generate 12 features using triangular window. The idea of enhanced autocorrelation is to prune the periodicity of autocorrelation function. In a typical case this representation helps in finding the fundamental periodicities of harmonic complex tones in a mixture of such tones. It removes the common periodicities such as the root tone of musical chords [14].

3.2. Acoustic

In acoustic signals, for footsteps, the hoof sounds of animals such as horses, donkeys, or mules are distinct from human footstep sounds. Based on this phenomena, we use Perceptual Linear Predictive (PLP) feature [15], which is a common feature in speech recognition. There are several strategies due to the property of the realistic dataset:

- Noisy
  As mentioned in Section 2, the data is recorded in a open field. There are noisy wind sounds blowing in the recordings. We use spectral subtraction method to reduce the effect by noise [16, 17].
- No label for exact footsteps time
  In the dataset, since there is no label for the exact time of footsteps sounds, we have to use the seismic sensor information, assuming that the peaks in the seismic signals are corresponding to footsteps. Suppose there are n peaks in seismic signals at time t_i, for $i = 1, ..., n$, we choose a small time δ around the peaks and extract PLP features within the time

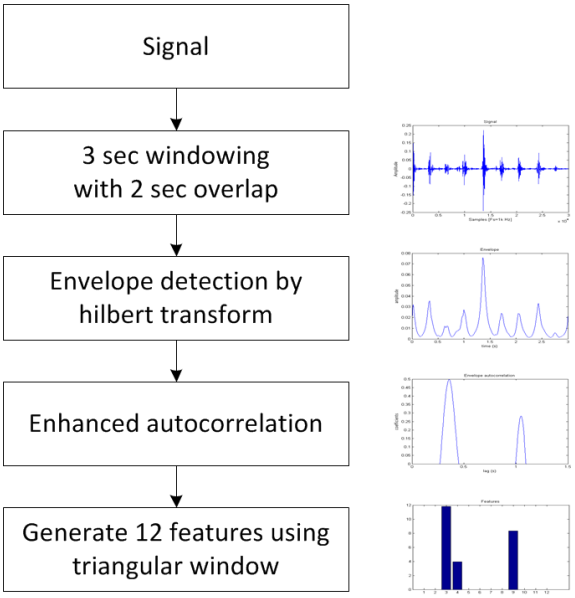

Figure 3: Seismic feature extraction algorithm

periods $(t_i - \delta, t_i + \delta)$, for $i = 1, ..., n$. In each time period, we extract 13-dimensional PLP features using 186ms Hamming windows with 75% overlap, where 186 ms is approximately equal to the time duration of a single footstep (from heel strike to toe slap). Along with the delta and delta-delta coefficients of the PLP features, there are totally 39-dimensional features.

- Overlapping class
  Our goal is to classify *humans only* and *humans with four-legged animals* (or abbreviated as *humans with animals*). In the *humans with animals* class, there are instances that the footstep sounds are from humans. Therefore, there will be some overlaps between the two classes in the feature space. Discriminative methods such as support vector machine (SVM) will not work properly under this condition. It is also not suitable for training the generative models for the two classes directly. Hence, we propose a multi-stage framework for feature selection as shown in Figure 4. The idea of the framework is to drop the features in *humans with animals* class which are similar to the features in *human only* class.

  The algorithm of the framework is as follows:

  1. Train two models for *humans only* and *humans with animals* using training data as shown in the left block.
  2. Predict the training data of *humans with animals* class using trained models as shown in the middle block. Each frame in the training data is predicted as either *humans* class or *humans with animals* class.
  3. Keep the frames which predicted as *humans with animals* class and train a new model as *estimated animals only* class.
  4. Use the new model of *estimated animals only* class and the original *humans only* model for classification as shown in the right block.

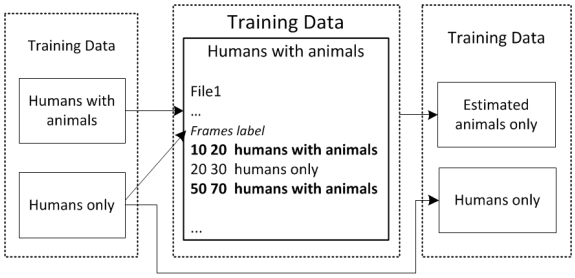

Figure 4: Multi-stage framework for acoustic feature selection

Note that the acoustic features belong to short-time features which extracts information within a short time duration, while seismic, ultrasonic features belong to long-time features. Therefore, the multi-stage feature selection framework applies for acoustic features only.

3.3. Ultrasound

Ultrasonic sensor, also known as acoustic Doppler sensor [9], emits acoustic waves toward the objects and receives reflected response from objects. Benefits of using ultrasonic sensors include low-cost ($5 USD in 2011) and low-power. The limitation is that, by being acoustic in nature, ultrasonic sensor has a limited range of the order of ten meters.

The velocity of a moving object relative to an observer can be estimated by measuring the frequency shift of a wave radiated or scattered by the object, that is known as the Doppler effect. If the object itself contains moving parts, each moving part will result in a modulation of the base Doppler frequency shift, known as the micro-Doppler effect. Given an acoustic wave transmitted by an observer, the frequency of the received wave by a single point scatterer is

$$f = f_0 \left(1 + \frac{2v}{c}\right) \quad (1)$$

where f_0 is the frequency of the transmitted acoustic wave, v is the velocity of the scattered wave relative to the observer and c is the speed of sound. The Doppler frequency shift $\Delta f = \frac{2v}{c}$ is proportional to the velocity of the scattered wave relative to the observer.

A human body is an articulated object, comprising a number of rigid bones connected by joints. When a continuous tone is incident on a walking person/an animal, the reflected signal contains a spectrum of frequencies arising from the Doppler shifts of the carrier tone by the velocities of various moving body parts.

As reported in [10], based on different physical walking mechanisms, the micro-Doppler gait signatures between a person and an four-legged animal are different. We use this concept to extract feature in order to distinguish between humans and four-legged animals.

For ultrasound signal processing, given the data with two channels, 25 kHz and 40 kHz, first, we use a band-pass filter with stop-band at 20 kHz and 30 kHz; 35 kHz and 45 kHz, and passband at 22.5 kHz and 27.5 kHz; 37.5 kHz and 42.5 kHz, respectively. Then, we use Hilbert transform demodulating the captured Doppler signals to emphasize the contribution of various velocities. Then, following [11], we use cepstral coefficients for representing the patterns in the spectrogram, which is also a common technique in audio processing. We use 62ms with 75% overlap Hamming window and

use the first 40 coefficients of the cepstral vectors and delta-ceptral coefficients, for modeling the differential spectrum. Note that the frames are relatively wide because of the slow varying nature of the signal. The resulting 80-dimensional vector is the feature to represent ultrasonic signals.

4. Methods

4.1. Gaussian Mixture Models Classifier

Given the above features, we use Gaussian Mixture Models (GMMs) for classification. The motivation for using Gaussian mixture densities is that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. GMM has the ability to form smooth approximations to arbitrarily-shaped densities. GMM is often used in speaker verification/identification [18].

A Gaussian mixture density is a weighted sum of M component densities, as shown in the following equation,

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{2}$$

where $\vec{x}$ is a D-dimension random vector, $b_i(\vec{x})$, $i = 1, \ldots, M$, are the component densities and p_i, $i = 1, \ldots, M$, are the mixture weights. Each component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\Sigma_i^{-1}(\vec{x} - \vec{\mu}_i)\} \tag{3}$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices (using diagonal covariance matrix here) and mixture weights from all component densities. These parameters are collectively represented by the notation $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}, i = 1, \ldots, M$. For classification, each class is represented by a GMM and is referred to with its model λ.

Given training data from each class, the goal of model training is to estimate the parameters of the GMM. Maximum likelihood model parameters are estimated using the iterative Expectation-Maximization (EM) algorithm. Generally, ten iterations are sufficient for parameters convergence.

The objective is to find the class model which has the maximum *a posteriori* probability for a given observation sequence X.

$$\hat{N} = \underset{1 \le k \le N}{\operatorname{argmax}}\ p(\lambda_k|X) = \underset{1 \le k \le N}{\operatorname{argmax}} \frac{p(X|\lambda_k)p(\lambda_k)}{p(X)} \tag{4}$$

where the second equation is due to Bayes' rule. Assuming equal likelihood for all classes (i.e., $p(\lambda_k) = 1/N$) and $p(X)$ is the same for all class models, the classification rule simplifies to

$$\hat{N} = \underset{1 \le k \le N}{\operatorname{argmax}}\ p(X|\lambda_k) = \underset{1 \le k \le N}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{5}$$

where the second equation uses logarithms and the independence between observations. T is the number of observations.

4.2. Decision Fusion

The features mentioned above, which come from different modalities, are concatenated. Further, we can expand the feature vector $\vec{x}$ formed from different modalities. For example,

$$p(\vec{x}|\lambda) = p(\vec{x}_{a,s,u}|\lambda) = \prod_{m \in M} p(\vec{x}_m|\lambda)^{w_m} \tag{6}$$

where $M = \{a, s, u\}$, a, s, u represents acoustic, seismic, and ultrasound modalities, respectively. w_m denotes the modality weights, that are nonnegative, and models the reliability for the modality M. In this paper, we constrain the weights to sum up to one. For simplicity reason, we choose weights by grid-search of global weights on validation sets [19], as shown in Section 5. Note that (6) corresponds to a linear combination in the log-likelihood domain; however, it does not represent a probability distribution in general, and will be referred to as a score.

5. Experiments

In this section, we describe three experiments in order to compare our proposed methods with previous approach in classifying *humans only* and *humans with four-legged animals*. Totally, there are 69 recordings in the dataset. We divide the recordings into four groups and choose two for training and two for testing at a time, totally six-fold cross-validations. We choose the best mixture number according to an additional validation set. The experimental results are represented by mean $\pm$ standard error.

5.1. Seismic features

As describe in Section 3.1, we compare our gait pattern features based on enhanced autocorrelation with the temporal gait pattern [5]. The experimental results are shown in Table 1.

Feature	Accuracy (%)
Temporal gait pattern [5]	71.883±4.607
Enhanced autocorrelation pattern	81.707±2.564

Table 1: Experimental Results of Seismic features

5.2. Acoustic features

As describe in Section 3.2, we want to examine the effect of (1) spectral subtraction, (2) using seismic peaks with different δ's, and (3) our proposed multi-stage feature selection framework. The experimental results are shown in Table 2.

5.3. Decision fusion with seismic, acoustic, and ultrasonic features

Using the feature extraction methods mentioned above, we then concatenate features for decision fusion. Note that, for ultrasonic data, within 186ms, there are eight moving windows resulting in 640-dimension features. We use principal component analysis (PCA) keeping 99% energy and reduce features to 7 dimensions. We compare the features our (1) enhanced autocorrelation pattern, PLP features with spectral subtraction, $\delta = 0.3$ seismic peaks, and

Feature	Accuracy (%)
PLP features without (1) (2) (3)	73.768±5.462
PLP features with (1)	76.105±4.098
PLP features with (1) (2), $\delta = 0.1$	74,975±5.079
PLP features with (1) (2) (3), $\delta = 0.1$	75.737±2.936
PLP features with (1) (2), $\delta = 0.3$	77.555±4.268
PLP features with (1) (2) (3), $\delta = 0.3$	79.015±3.799
PLP features with (1) (2), $\delta = 0.5$	75.392±3.376
PLP features with (1) (2) (3), $\delta = 0.5$	77.688±3.149

Table 2: Experimental Results of Acoustic features, where (1) represents spectral subtraction, (2) represents the usage of seismic peaks with different δ, and (3) represents the usage of our proposed multi-stage feature selection framework

multi-stage feature selection framework, and (2) temporal gait pattern [5], PLP features without spectral subtraction, using the whole segments, and without multi-stage feature selection, (3) ultrasonic feature. The experimental results are shown in Table 3.

Feature	Accuracy (%)
(1) our proposed method with ultrasonic feature	86.092±5.667
(2) previous method with ultrasonic feature	81.903±3.144
(3) ultrasonic features	79.338±5.574

Table 3: Experimental Results of Decision Fusion

6. Discussion

From the experimental results of Table 1, our proposed method using enhanced autocorrelation pattern outperforms previous methods [5], because previous method did not consider the case of multiple objects. For Table 2, we can observe that using spectral subtraction enhances the performance under noisy environment. We can also observe using seismic peaks with $\delta = 0.2$ achieves the best result. Possible explanation is that there is a little asynchrony between acoustic sounds and seismic peaks. Hence, the $\delta = 0.1$ cannot exactly capture the footstep sounds. Also, we can observe using multi-stage feature selection framework improves the accuracy under different δ's. In Table 3, we show that using our proposed method in seismic and acoustic features along with ultrasonic features greatly improves the robustness of the system.

7. Conclusion

In this paper, we use a challenging realistic multi-sensor multi-modality dataset for personnel detection. Based on phenomenology of the differences (micro-Doppler motion, enhanced autocorrelation pattern, footstep sound) between humans and four-legged animals, we propose a new seismic feature extraction method based on enhanced autocorrelation, and a multi-stage acoustic feature selection framework. From experimental results, we show that the combination of multi-modality sensors improves the robustness of the system over previous approach. It is possible to further extend the current fusion system for sensor network fusion. It is inexpensive to deploy unattended ground sensors such as acoustic, seismic, ultrasonic sensors in target areas.

8. Acknowledgments

This research is supported by ARO MURI 2009-31.

9. References

[1] T. Damarla, "Sensor fusion for isr assets," M. A. Kolodny, Ed., vol. 7694, no. 1. SPIE, 2010, p. 76941C.

[2] T. Damarla, L. Kaplan, and A. Chan, "Human infrastructure & human activity detection," in *Information Fusion, 2007 10th International Conference on*, 9-12 2007, pp. 1 –8.

[3] J. M. Sabatier and A. E. Ekimov, "Range limitation for seismic footstep detection," E. M. Carapezza, Ed., vol. 6963, no. 1. SPIE, 2008, p. 69630V.

[4] K. M. Houston and D. P. McGaffigan, "Spectrum analysis techniques for personnel detection using seismic sensors," E. M. Carapezza, Ed., vol. 5090, no. 1. SPIE, 2003, pp. 162–173.

[5] H. O. Park, A. A. Dibazar, and T. W. Berger, "Cadence analysis of temporal gait patterns for seismic discrimination between human and quadruped footsteps," in *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1749–1752.

[6] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using gaussian mixture models and gmm supervectors," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 69–72, 2009.

[7] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543 – 1551, 2010.

[8] P.-S. Huang, X. Zhuang, and M. A. Hasegawa-Johnson, "Improving acoustic event detection using generalizable visual features and multi-modality modeling," in *Acoustics, Speech and Signal Processing. ICASSP 2011. IEEE International Conference on*, 2011.

[9] A. Ekimov and J. M. Sabatier, "Human detection range by active doppler and passive ultrasonic methods," E. M. Carapezza, Ed., vol. 6943, no. 1. SPIE, 2008, p. 69430R.

[10] Z. Zhang, P. Pouliquen, A. Waxman, and A. Andreou, "Acoustic micro-doppler gait signatures of humans and animals," in *Information Sciences and Systems, 2007. CISS '07. 41st Annual Conference on*, 14-16 2007, pp. 627 –630.

[11] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recogination," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, 5-7 2007, pp. 27 –32.

[12] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 1, pp. 24 – 33, feb 1977.

[13] P. D. L. Cuadra and A. Master, "Efficient pitch detection techniques for interactive music," in *In Proceedings of the 2001 International Computer Music Conference, La Habana*, 2001.

[14] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 708 –716, Nov. 2000.

[15] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, vol. 4, Apr. 1979, pp. 208 – 211.

[17] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504 –512, Jul. 2001.

[18] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72 –83, jan 1995.

[19] Guillaume, G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and mce based hmm stream weight estimation for audio-visual asr," in *Proc. Int. Conf. Acous. Speech Sig. Process*, 2002, pp. 853–856.