Integrating LDV Audio and IR Video for Remote Multimodal Surveillance

Zhigang Zhu, Weihong Li and George Wolberg
Department of Computer Science, City College and Graduate Center
The City University of New York, New York, NY 10031
{zhu, wli, wolberg}@cs.ccny.cuny.edu

Abstract

This paper describes a multimodal surveillance system for human signature detection. The system consists of three types of sensors: infrared (IR) cameras, pan/tilt/zoom (PTZ) color cameras and laser Doppler vibrometers (LDVs). The LDV is explored as a new non-contact remote voice detector. We have found that voice energy vibrates most objects and the vibrations can be detected by an LDV. Since signals captured by the LDV are very noisy, we have designed algorithms with Gaussian bandpass filtering and adaptive volume scaling to enhance the LDV voice signals. The enhanced voice signals are intelligible from targets without retro-reflective finishes at short or medium distances (<100m). By using retro-reflective tapes, the distance could be as far as 300 meters. However, the manual operation to search and focus the laser beam on a target with both vibration and reflection is very difficult at medium and large distances. Therefore, infrared (IR) imaging for target selection and localization is also discussed. Future work remains in automatic LDV targeting and intelligent refocusing for long range LDV listening.

Keywords: *laser vibrometry, clandestine listening, multimodal integration, audio signal enhancement, infrared video surveillance*

1. Introduction

Recent improvements in laser vibrometry [1-6] and day/night IR imaging technology [15, 16] have created the opportunity to create a long-range multimodal surveillance system for human signature detection. Such a system would have day and night operation [16]. The IR video system would provide the video surveillance necessary to permit the operator to select the best target for picking up acoustic signals (e.g. human speech). The LDV is then focused upon that target to recover the acoustic signals. This multimodal capability would greatly improve security force performance through clandestine listening of targets that are probing or penetrating a perimeter defense. The targets may be aware that they are observed but most likely would not infer that they could be heard. Unlike microphones, an LDV using the principle of laser interferometry is a non-contact, remote voice detector, working in a similar way as an IR or visible camera. In this sense, the LDV extends the spectrum of long-range sensing beyond visible and infrared ranges. This integrated system could also provide the feeds for advanced face and voice recognition systems.

Laser vibrometers such as those manufactured by Polytec™ [2] and B&K Ometron [3] can effectively detect vibration within two hundred meters with sensitivity on the order of 1μm/s. These instruments are designed for use in laboratories (0-5 m working distance) and field work (5-200 m) [2-7]. For example, these instruments have been used to measure the vibrations of civil structures like high-rise buildings, bridges, towers, etc. at distances of up to 200m. However, for distances above 200 meters, it will be necessary to treat the target surface with retro-reflective tape or paint to ensure sufficient retro-reflectivity. Another difficulty is that such an instrument uses a front lens to focus the laser beam on the target surface in order to minimize the size of the measuring point. At a distance above 200 m, the speckle pattern of the laser beam induces noise and signal dropout will be substantial [8].

The overall goal of this project is to create an advanced multimodal interface for human signature extraction (including audio, visible and thermal) using the state-of-the-art sensing technologies for perimeter surveillance. Meanwhile, the capabilities of sensors - infrared (IR) cameras, visible (EO) cameras, and the laser vibrometers (LDVs) in our study - are critical to surveillance tasks. Recently, IR and EO cameras have been widely used in human and vehicle detection in traffic and surveillance applications [16]. However, literature on remote voice detection using LDVs is rare. Therefore, the study of the novel LDV-based voice detection will be the main focus of this paper.

The performance of the laser Doppler vibrometer strongly depends on the reflectance properties of the target surface. Important issues such as target surface properties, size and shape, distance from the sensor, and sensor targeting and focusing are studied through

several sets of indoor and outdoor experiments. Furthermore, the LDV signal may be corrupted by laser photon noise, target movements, and background acoustic noise such as wind and engine sounds. Therefore, speech enhancement algorithms are applied to improve the performance of recognizing a noisy voice detected by the LDV system. Many speech enhancement algorithms have been proposed [9-12], but they have been mainly used for improving the performance of speech communication systems in noisy environments. Acoustic signals captured by laser vibrometers need special treatment.

This paper is organized as follows. In Section 2, we give an overall picture of our technical approach: the human-centered paradigm for the integration of laser Doppler vibrometry and IR imaging for multimodal surveillance. The two main sensor modules will be briefly described in Section 3. Then, we discuss various aspects of LDVs for voice detection: basic principles and problems in Section 4, signal enhancement algorithms in Section 5, and experimental designs in Section 6. We have designed a graphic human computer interface for signal analysis, signal filtering, and signal synthesis. In Section 7 we discuss how to use IR/EO imaging for target selection and localization for LDV listening. Finally we conclude our work in Section 8.

2. A Multimodal Integration Approach

There are three main components in our approach to multimodal human signature detection (Figure 1): the IR/EO video surveillance component, the LDV audio surveillance component, and the human-computer interaction components. Both the IR/EO and LDV sensing components can support day and night operation even though it will be better to use a standard EO camera (coupled with the IR camera) to perform the surveillance task during daytime. The overall approach is the integration of the IR/EO imaging and LDV audio detection for a long-range surveillance task. The integration has the following three steps.

Step 1. Target detection, tracking, and selection via the IR/EO imaging module. The targets of interest could be humans or vehicles (driven by humans). This will be performed by motion detection and human/ vehicle segmentation methods.

Step 2. Audio targeting and detection by the LDV audio module. The audio signals could be human voices or vehicle engine sounds. We mainly consider human voice detection in our work. The main issue is to select the LDV targeting points provided by the IR/EO imaging module to detect the vibration caused by human voices.

Step3. Optimal viewpoint selection from audio detection. By using audio feedback, the IR/EO imaging module can verify the existence of humans and capture the best face images for face recognition. Together with the voice recognition module, the surveillance system could further perform human identification and event understanding.

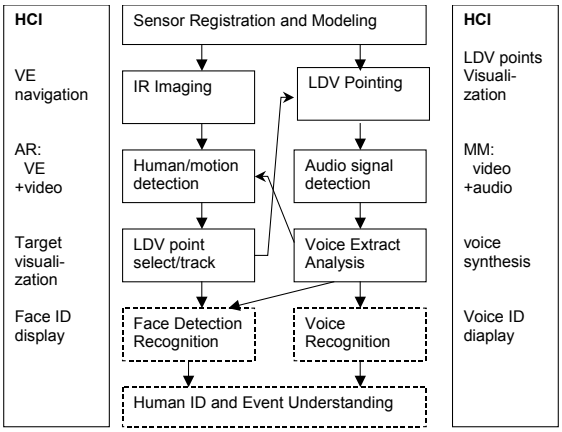


Figure 1. System components of a multimodal human signature detection system.

An important concept is to design a human-computer interface for human-centered multimodal (MM) surveillance. The basic idea is to provide an advanced virtual-environment (VE) based interface of the site (e.g., air base) to give the operator the best cognitive understanding of the environment, the sensors, and the events. One of the important issues is how to use IR imaging to help the laser Doppler vibrometer to select the appropriate targets. Figure 1 shows the human-computer interaction (HCI) synopsis for human-in-the-loop surveillance operation with augmented reality (AR) visualization, target selection, signal extraction and enhancement, and human identification.

3. Multimodal Sensors

To enable the study of multimodal sensor integration for human signature detection, we have acquired the following sensors: a Laser Doppler Vibrometer (LDV) OFV-505 from Polytec, a ThermoVision A40M infrared camera from FLIR, and a Canon color/near IR pan/tilt/zoom (PTZ) camera.

The Laser Doppler Vibrometer from Polytec [2] includes a controller OFV-5000 with a digital velocity decode card VD-6 and a sensor head OFV-505 (Figure 2). We also acquired a telescope VIB-A-P05 for accurate targeting at large distances. The sensor head uses a helium-neon (HeNe) red laser with a wavelength of 633.8 nm and is equipped with a super long-range lens. It sends the interferometry signals to the controller, which is connected to the computer via an RS-232 port. The controller box includes a velocity

decoder VD-06, which processes signals received from the sensor head. There are three types of output signal formats from the controller, including an S/P-DIF output, and digital and analogue velocity signal outputs.

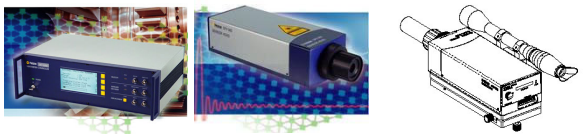


Figure 2 The Polytec™ LDV (a) Controller OFV-5000 (b) Sensor head OFV-505 (c) Telescope VIB-A-P05

Figure 3. A person sitting in darkness can be clearly seen in the IR image, and the temperature be accurately measured. The reading at the cross (Sp1) on the face is 33.1°C.

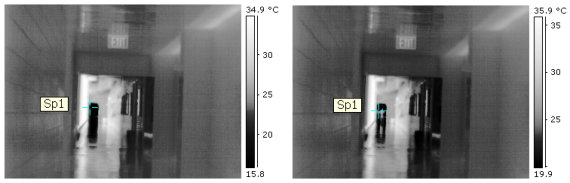


Figure 4. Two IR images before and after a person standing at a distance of about 200 feet. The reading of the temperature at the cross (Sp1) changes from 11°C to 27 °C.

The FLIR ThermoVision A40M IR camera has a 320x240 focal plane array with uncooled microbolometer detector. The spectral range is 7.5 to 13 μm. Its ability to accurately measure temperature makes it suitable for human and vehicle detection. The measurable temperature range is -40° to 500°C with an accuracy ± 2°C (or ± 2%). Figure 3 shows an example where a person sitting in a dark room can be clearly detected by the far-infrared camera. Furthermore, the accurate temperature measurements provide important information for discriminating human bodies from other hot/warm objects. After successful human detection, objects in the environment (such as the doors or walls in this example) can be searched whose vibration with audio waves could reveal what is being spoken. Since the FILR ThermoVision camera is a far-infrared thermal camera, it does not need to have active IR illumination, and it is suitable for detecting humans and vehicles at a distance (Figure 4).

4. LDV Long-Range Audio Capture

Laser Doppler vibrometers (LDVs) work according to the principle of laser interferometry. Measurements are made at the point where the laser beam strikes the structure under vibration. In the Heterodyning interferometer (Figure 5), a coherent laser beam is divided into object and reference beams by a beam splitter BS1. The object beam strikes a point on the moving (vibrating) object and light reflected from that point travels back to beam splitter BS2 and mixes (interferes) with the reference beam at beam splitter BS3. If the object is moving (vibrating), this mixing process produces an intensity fluctuation in the light. Whenever the object has moved by half the wavelength, λ/2, which is 0.3169 μm (or 12.46 micro inches) in the case of HeNe laser, the intensity has gone through a complete dark-bright-dark cycle. A detector converts this signal to a voltage fluctuation. The Doppler frequency f_D of this sinusoidal cycle is proportional to the velocity v of the object according to the formula

$$f_D = 2 \cdot v / \lambda \qquad (1)$$

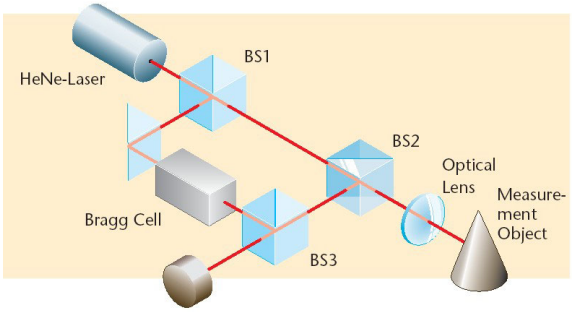


Figure 5 .The modules of the Laser Doppler Vibrometer

Instead of detecting the Doppler frequency, the velocity is directly obtained by a digital quadrature demodulation method [1, 2]. The Bragg cell, which is an acousto-optic modulator to shift the light frequency by 40 MHz, is used for identifying the sign of the velocity.

Objects vibrate while wave energy (including voice waves) is applied to them. Though the vibration caused by the voice energy is very small compared with other vibration, this tiny vibration can be detected by the LDV. Voice frequency f ranges from about 300 Hz to 3000 Hz. Velocity demodulation is better for detecting vibration with higher frequencies because of the following relationship between vibration velocity, frequency, and magnitude:

$$v = 2\pi f m \qquad (2)$$

Note that the velocity v will be large with a large frequency f, even under a small magnitude m. The Polytec LDV sensor OFV-505 and the controller OFV-5000 can be configured to detect vibrations under

several different velocity ranges: 1 mm/s, 2 mm/s, 10 mm/s, and 50 mm/s. For voice vibration, we usually use the 1mm/s velocity range. The best resolution is 0.02 μm/s under 1mm/s range, according to the manufacture's specification (with retro-tape treatment). Without retro-tape treatment, the LDV still has sensitivity on the order of 1 μm/s, i.e. one-thousandth of the full range. This indicates that the LDV can detect vibration (due to voice waves) at a magnitude as low as $m = v/2\pi f = 1/(2*3.14*300) = 0.5\ pm$. Note that voice waves are in a relative low frequency range. The Polytec OFV-505 LDV sensor that we have is capable of detecting vibration with a much higher frequency (up to 350K Hz).

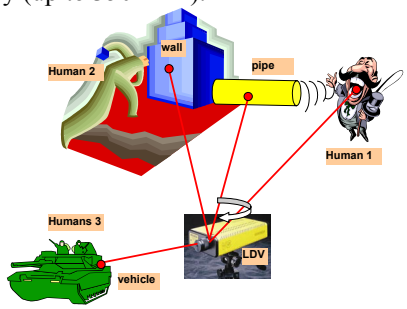


Figure 6. Target selection and multimodal display. The LDV can measure audio signals from tiny vibrations of the *LDV points* (indicated by the read beams and dots onto the objects) that couple with the audio sources

There are two important issues to consider in order to use an LDV to detect the vibration of a target caused by human voices. First, the target vibrates with the voices. Second, points on the surface of the target where the laser beam hits reflect the laser beam back to the LDV. We call such points *LDV targeting points*, or simply *LDV points*. Therefore, the LDV points selected for audio detection could be the following three types of targets (Figure 6).

(1) *Points on a human body*. For example, the throat of a human will be one of the most obvious parts where the vibration with the speech could be detected by the LDV. However, we have found that it is very challenging since it is "uncooperative": (a) it is not easily targeted since the human is hard to keep still; (b) it does not have a good reflective surface for the laser beam, and therefore a retro-reflective tape has to be used; (c) the vibration of the throat only includes the low frequency parts of the voice. For these reasons, our experiments will mainly focus on the remaining two types of targets.

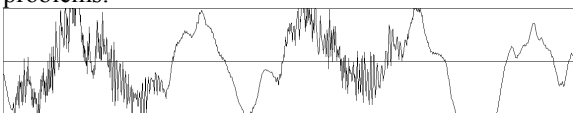
(2) *Points on a vehicle with humans within*. Human voice signals vibrate the body of a vehicle, which could be readily detected by the LDV. Even if the engine is on and the volume of the speech is low (e.g., in cases of whispering), we could still extract the human voice by signal decomposition since the human voice and engine noise have different frequency ranges. However, without applying retro-reflective tape, we have found that the body of the vehicle basically does not reflect the HeNe laser suitably for our purposes, even if the vehicle is stationary. With retro-tape, the signal returns with LDV are excellent when the targets (cars) are at various distances (10 to 50 meters in our experiments) and also with a large range of incident angles of the laser beam. This indicates that remotely detecting voices inside a vehicle will be possible if, for example, small retro-vibration "bullets" could be shot onto the body of the vehicle.

(3) *Points in the environment.* For perimeter surveillance, we can use existing facilities or install special facilities for human audio signal detection. We have found that most objects vibrate with voices, and many types of surfaces reflect the LDV laser beam within some distance (about 10 meters). Response is even better if we can paint or paste certain points of the facilities with retro-reflective tapes or paints; operating distances can increase to 300 meters (1000 feet) or more. Facilities like walls, pillars, lamp posts, large bulletin boards, and traffic signs vibrate very well with human voices, particularly during the relative silence of night. Note that an LDV has sensitivity on the order of 1 μm/s, and can therefore pick up very small vibrations.

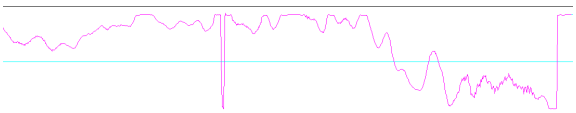
5. LDV Audio Signal Enhancement

For the human voice, the frequency range is about 300 Hz to 3 KHz. However, the frequency response range of the LDV is much wider than that. Even if we have used the on-board digital filters, we still get signals that are subject to large, slowly varying components corresponding to the slow but significant background vibrations of the targets. The magnitudes of the meaningful acoustic signals are relatively small, adding on top of the low frequency vibration signals. This renders the acoustic signals to be unintelligible to the human ear. On the other hand, the inherent "speckle pattern" problem on a normal "rough" surface and the occlusion of the LDV laser beam by passing objects introduce noise with large and high-frequency components. This creates undesirably loud noise when we directly listen to the acoustic signal. Therefore, we have applied a Gaussian bandpass filter to process the vibration signals captured by the LDV. In addition, the volume of the voice signal may change dramatically with changes in the vibration magnitudes of the target due to variability in shouting, normal speaking and whispering, and the distances of the human speakers to the target. Therefore, we have also designed an adaptive volume function to cope with this problem.

Figure 7 shows two real examples of these two types of problems.



(a) "Hello…Hello"



(b) "I am whispering…(high frequency noise)… OK … Hello (high frequency noise)"

Figure 7. Two real examples of LDV acoustic signals with both low and high frequency noises.

5.1. The Gaussian bandpass filter

It is well-known in image processing that the Gaussian bandpass transfer function can be expressed as the difference of two Gaussians of different widths [13]:

$$H(s) = Be^{-s^2/2\alpha_2^2} - Ae^{-s^2/2\alpha_1^2}, \quad B \geq A,\ \alpha_2 > \alpha_1 \qquad (3)$$

Figure 8 shows the function. The impulse response of this filter is given by

$$h(t) = \frac{B}{\sqrt{2\pi\sigma_2^2}} e^{-t^2/2\sigma_2^2} - \frac{A}{\sqrt{2\pi\sigma_1^2}} e^{-t^2/2\sigma_1^2}, \quad \sigma_i = \frac{1}{2\pi\alpha_i} \qquad (4)$$

Notice that the broader Gaussian in the frequency domain (Figure 8) creates a narrower Gaussian in the time domain (Figure 9), and vice versa. We want to reduce the signal magnitude outside the frequency range of human voices, i.e., below s_1= 300 Hz and above s_2 = 3K Hz. The high frequency reduction is mainly controlled by the width of the first (the broader) Gaussian function in Eq. (3), i.e., α_2, and the low frequency reduction is mainly controlled by the width of the second Gaussian function, i.e., α_1. Since the Gaussian function drops significantly when $|s_i| > 2\alpha_i$, (i=1, 2), as shown by a pair of '*'s and a pair of '+'s in Figure 8, respectively, we obtain the widths of the two Gaussian functions in the frequency domain as

$$\alpha_i = s_i / 2 \ \ (\text{Hz}), \ \ i = 1, 2 \qquad (5)$$

In practice, we process the waveform directly in the time domain, i.e., by convolving the waveform with the impulse response in Eq. (4). This leads to a real-time algorithm for LDV voice signal enhancement (with a slight delay). For doing this, we need to calculate the variances of the two Gaussian functions in the time domain. Combining Eq. (4) and Eq. (5) we have

$$\sigma_i = \frac{1}{\pi s_i} \ (\text{seconds}), \ \ i = 1, 2 \qquad (6)$$

For digital signals, we need to determine the size of the convolution kernel. Since the narrower Gaussian (with width α_1) in the frequency domain creates a broader Gaussian (with width σ_1) in the time domain, we use σ_1 to estimate the appropriate window size of the convolution. Again, we truncate the impulse function when $t > 2\sigma_1$. Therefore, the size of the Gaussian bandpass filter is calculated as

$$W_1 = 2m(2\sigma_1) + 1 = \frac{4m}{\pi s_1} + 1 \qquad (7)$$

where m is the sampling rate of the digital signal. Typically, we use m = 48 K samples/second with the Sony/Philips Digital Interconnect (S/P DIF) format. Therefore, the size of the window will be W_1 = 210. The size of the convolution kernel is marked by a pair of '*'s in Figure 9.

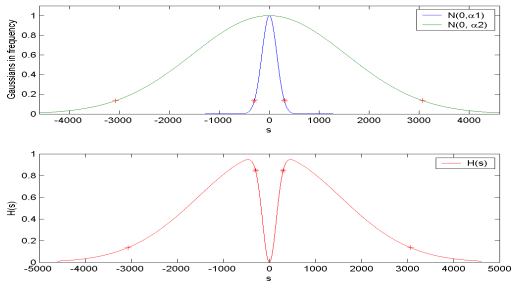


Figure 8. The Gaussian bandpass filter transfer function

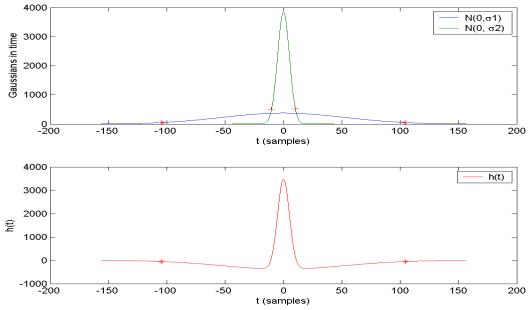


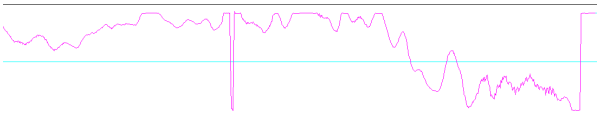
Figure 9. The Gaussian bandpass filter impulse response

5.2. Volume selection and adaptation

The useful original signal obtained from the S/P-DIF output of the controller is a velocity signal. When treated as the voice signal, the volume is too small to be heard by human ears. When volumes of the voice signals change dramatically within an audio clip, a fixed volume increase cannot lead to clearly audible playback. Therefore, we have designed an adaptive volume algorithm. For each audio frame, for example of 1024 samples, the volumes are scaled by a scale v that is determined by the following equation:

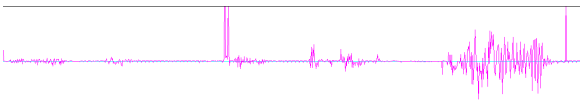
$$v = \frac{C_{\max}}{\left|\max(x_1, x_2 \ldots x_n)\right|} \qquad (10)$$

where $C_{\max}$ is the maximum constant value of the volume (defined as the largest short integer, i.e., 32767), and $x_1, x_2, ..., x_n$ are sample data in one speech frame (e.g. n = 1024 samples). The scaled sample data stream, $vx_1, vx_2, ..., vx_n$, will then be played via a speaker so that a suitable level of voice will be heard.

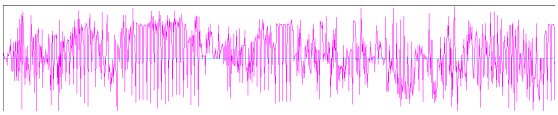


(a) Original signal

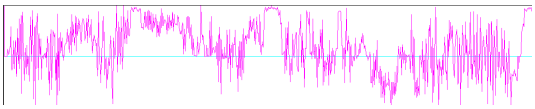
(b) ×1 after band-pass filtering



(c) ×8 after band-pass filtering



(d) adaptive scaling after band-pass filtering



(e) adaptive scaling after low-reduction filtering

Figure 10. The waveform of the original signal and the results of fixed scaling and adaptive scaling, after using suitable Gaussian filtering. The short audio clip reads "I am whispering…(noise)… OK … Hello (noise)", which was captured by the LDV OFV-505 from a metal cake-box carried by a person at a distance of about 30 meters from the LDV.

The adaptive method will always give a suitable volume for any kind of the sampled data stream. In our LDV software system, both adaptive and fixed scaling methods are implemented, and the user can choose either method on the fly. Figure 10 shows a real example of filtering and scaling. In this example, the best performance of the filtering is obtained with only the low-reduction filter (Figure 10e).

6. Experiment Designs and Analysis

In order to use an LDV to detect audio signals from a target, the target needs to meet two conditions: reflection to HeNe laser and vibration with voices. Due to the difficulty in detecting voice vibration directly from the body of a human speaker, we mainly focus on the use of targets in the environments near the human speaker. We have found that the vibration of most objects in man-made environments caused by waves of voices can be readily detected by the LDV. However, the LDV must get signal returns from the laser reflection. The degree of signal returns depends on the following conditions: (1) surface normal vs. laser beam direction; (2) surface color with spectral response to 632.8 nm; (3) surface roughness; and (4) the distance from the sensor head to the target. Retro-reflective traffic tapes or paints are a perfect solution to the above reflection problems *if* the targets are "cooperative", i.e., the surfaces of targets can be treated by such tapes or paints. The traffic retro-reflective tapes (retro-tapes) are capable of diffuse reflection in that they reflect the laser beam back in all directions within a rather large angular range.

We have performed experiments with the following settings: types of surface, surface directions, long-range listening, through-wall listening, and talking inside of cars. In all experiments, the LDV velocity range is 1 mm/s, and a person's speech describes the experiment configurations. A walkie-talkie is used for remote communication only. The same configurations (band-pass 300 – 3000 Hz, adaptive volume) are used in processing the data for all the experiments. Each audio clip should tell you most of the information for the experiment if it is intelligible. In this paper we will only provide the results of two sets of experiments: long range listening and through-window listening. More data collections and results can be found in our technical report [17], where we provide audio files for both the original LDV audio clips and the processed audio clips (with one fixed configuration of filtering). The original clips have very low volume so it is difficult to hear anything meaningful. On the other hand, the processed audio clips are not optimal at all for intelligibility. Our LDV program allows a user to interactively tune the filtering and scaling parameters in real-time, view the waveforms and the spectrograms, and hear the audio clips in order to get the best intelligibility of the enhanced LDV audio signals.

We tested the long range LDV listening in an open space with various distances from about 30 to 300 meters (100 ft to 1000 ft, Figure 11). A small metal cake box with retro-tape finish was fixed in front of the speaker's waist. The signal return of the LDV is insensitive to the incident angles of the laser beam, due

to the retro-tape finish. Both normal speech volumes and whispers have been successfully detected. The size of the laser spot changed from less than 1 mm to about 5-10 mm when the range changed from 30 to 300 meters. The noise levels also increased from 2 mV to 10 mV out of the total range of 20 V analogous LDV signals. The 260-meter measurement was obtained when the target was behind trees and bushes. With longer ranges, the laser is more difficult to localize and focus, and the signal return becomes weaker. Therefore, the noise levels become larger. Within 120 meters, the LDV voice is obviously intelligible; at 260-meter distance, many parts of the speech could be identified, even with some difficulty. For all the distances, the signal processing plays a significant role in making the speech intelligible. Without processing, the audio signal is buried in the low-frequency large-amplitude vibration and high-frequency speckle noises. It is important to emphasize that automatic targeting and intelligent refocusing is one of the important technical issues that deserve attention for long range LDV listening since it is extremely difficult to aim the laser beam at the target and keep it focused.

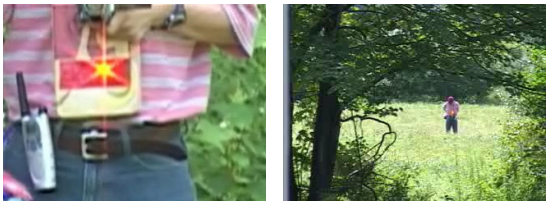


Figure 11. Long range LDV listening experiment. A metal cake box (left) is used, with a piece of 3M traffic retro-tape pasted. The laser spot can be clearly seen.

Figure 12. Listening through windows – a person was speaking outside the house, close to a window, while the LDV was listening inside the room via the window frame. Left: without retro-tape; right: with retro-tape.

In the experiments of LDV listening through windows, we used the window frames as vibration targets while a person was speaking outside the house (Figure 12). The LDV was inside the second floor of the house, several meters away from the window. The person spoke outside the house, two to three meters away from the window. Since the window frames are treated with paints, the reflection is good, even though the signal return strength is less than half of that with tape (see the bars in back of the LDV sensor Figure 12). We have also tested listening via the window frame and a closed door when the distance between the sensor head and the target was more than 20 meters (or 64 ft) away, while the distance between the speaker and the target was more than 5 meters. The LDV voice detection almost has the same performance as this short-range example.

7. Intelligent Targeting and Focusing

When using an LDV for voice detection, we need to find and localize the target that vibrates with voice waves and reflects the laser beam of the LDV, and then aim the laser beam of the LDV at the target. Multimodal integration of IR/EO imaging and LDV listening provides a solution for this problem. Ultimately this will lead to a fully automated system for clandestine listening for perimeter protection. Even when the LDV is used by a soldier in the field, automatic target detection, localization and LDV focusing will help the soldier to find and aim the LDV at the target for voice detection.

We have found that it is extremely difficult for a human operator to aim the laser beam of the LDV at a distant target and keep it focused. In the current experiments, the human operator turns the LDV sensor head in order to aim the laser beam at the target. The laser beam needs to be re-focused when the distance of the target is changed. Otherwise, the laser spot is out of focus. Consequently, it is very hard for the human to see the laser spot at a distance above 10 meters, and it is impossible to detect vibration when the laser spot is out of focus. Even with the automatic focus function of the Polytec OFV-505 sensor head, it usually takes more than 10 seconds for the LDV to search the full range of the focus parameter (0 – 3000) in order to bring the laser spot into focus. Therefore, automatic targeting and intelligent refocusing is one of the important technical challenges that require further attention for long-range LDV listening. Future research issues include the following three aspects. (1) *Target detection and localization via IR/EO imaging.* Techniques for detecting humans and their surroundings need to be developed for finding vibration targets for LDV listening. We have set up an IR/EO imaging system with an IR camera and a PTZ camera for this purpose. (2) *Registration between the IR/EO imaging system and the LDV system.* Two types of sensors need to be precisely aligned so that we can point the laser beam of the LDV to the target that the IR/EO imaging system has detected. (3) *Automated targeting and focusing.* Our current LDV system has real-time signal return strength measurements as well

as the real-time vibration signals. The search range of the focus function can also be controlled by program. Algorithms are in development to incrementally perform real-time laser focus *updating* by using the feedback of the LDV signal return strengths and the actual vibration signals.

8. Concluding Remarks

The LDV is a non-contact, remote voice detector with high-resolution in both space and time. In this paper, we have mainly focused on the experimental study in LDV-based voice detection. We also briefly discussed how we can use IR/EO imaging for target selection and localization for LDV listening. We have developed a multimodal system with three types of sensors (IR cameras, PTZ color cameras and LDVs) for human signature detection. We investigated the quality of voice captured by LDV devices that point to the objects nearby the voice sources. We have found that the vibration of the objects caused by the voice energy reflects the voice itself. After enhancement with Gaussian bandpass filtering and adaptive volume scaling, the LDV voice signals are mostly intelligible from targets without retro-reflective tapes at short distances (<100m). By using retro-reflective tapes, the distance could be as far as 300 meters.

However, without retro-reflective tape treatment, the LDV voice signals are very noisy from targets at medium and large distances. Therefore, further LDV sensor improvement is required. With current state-of-the-art sensor technology, we realize that more advanced signal enhancement techniques need to be developed than the simple band-pass filtering and adaptive volume scaling. For example, model-based voice signal enhancement could be a solution in that background noises might be captured and analyzed, and models could be developed from the resulting data.

We also want to emphasize that automatic targeting and intelligent refocusing is one of the important technical issues that deserve attention for long-range LDV listening. We believe that LDV voice detection techniques combined with the IR/EO video processing techniques can provide a more useful and powerful surveillance technology for both military and civilian applications.

9. Acknowledgements

This work is supported by the Air Force Research Laboratory (AFRL) under Grant No F33615-03-1-6383, by PSC-CUNY, and a CUNY Equipment Grant. We are grateful to Lt. Jonathan Lee and Mr. Robert Lee at AFRL for their guidance and valuable discussions on many technical issues during the course of this work. Prof. Ning Xiang at RPI provided his consulting services on LDVs that led us to a better understanding of the new sensor. Prof. Esther Levin at the City College has provided valuable discussions on speech signal processing. We also thank Mr. Robert T. Hill at the City College for proofreading the document.

10. References

[1] C.B. Scruby and L. E. Drain, Laser Ultrasonics Technologies and Applications, Bristol/Philadelphia/New York, Adam Hilger, 1990

[2] Polytec Laser Vibrometer, http://www.polytec.com/

[3] Ometron Vibration Measurement Systems. http://www.imageautomation.com/

[4] MetroLaser Laser Vibrometer, http://www.metrolaserinc.com/vibrometer.htm

[5] B.J. Halkon, S.R. Frizzel and S.J. Rothberg, "Vibration Measurements using Continuous Scanning Laser Vibrometry: Velocity Sensitivity Model Experimental Validation.", Measurement Science and Technology, 14(6), pp. 773-783, 2003

[6] Laser Radar Remote Sensing Vibrometer, http://sbir.gsfc.nasa.gov/SBIR/successes/ss/4-006text.html

[7] D. Costley, J. M. Sabatier and N. Xiang, " Forward-looking acoustic mine detection system", Proc. SPIE 15th Conference on Detection and Remediation Technologies for Mines and Minelike Targets IV, ed. by Dubey, A.C. et al. 2001, pp. 617-626

[8] J.W. Goodman, "Laser speckle and related phenomena" in Topics on Applied Physics, V. 9, Ed. J.C. Dainty, Springer-Verlag, Berlin, New York 1984

[9] I. Cohen, "On speech enhancement under signal presence uncertainty", ICASSP-2001, May 2001, pp.167-170

[10] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A Brief Survey of Speech Enhancement", the electronic handbook, CRC Press, (2003).

[11] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise", ICASSP-2002, May 2002, pp.573-576

[12] R. Vetter, "Single channel speech enhancement using MDL-based subspace approach in bark domain", ICASSP-2002, May 2001, pp.641-644

[13] K. R. Castleman, Digital Image Processing, Prentice-Hall, 1979

[14] L. Rabiner and B. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[15] FLIR Systems Security ThermoVision Cameras. http://www.flir.com/

[16] Joint IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS' 04), Washington DC, USA July 02, 2004

[17]. Z. Zhu, W. Li, Integration of laser vibrometer and infrared video for multimedia surveillance display, TR-2005006, CS Dept, CUNY Graduate Center, April 2005. An html version with links to LDV audio clips may be found at http://www-cs.ccny.cuny.edu/~zhu/LDV/FinalReportsHTML/CCNY-LDV-Tech-Report-html.htm