REMOTE VOICE ACQUISITION IN MULTIMODAL SURVEILLANCE

Weihong Li[1], Zhigang Zhu[1,2], George Wolberg[1,2]

[1]Department of Computer Science, Graduate Center, The City University of New York, NY 10016
[2]Department of Computer Science, City College, The City University of New York, NY 10031
{wli, zhu, wolberg }@cs.ccny.cuny.edu

ABSTRACT

Multimodal surveillance systems using visible/IR cameras and other sensors are widely deployed today for security purpose, particularly when subjects are at a large distance. However, audio information as an important data source has not been well explored. One of the reasons is because audio detection using microphones needs installation close to the subjects in monitoring. In this paper, we investigate a novel "optical" sensor, called Laser Doppler Vibrometer (LDV), for capturing voice signals in a very large range to realize a truly remote and multimodal surveillance system. Speech enhancement approaches are studied based on the characteristics of LDV Audio. Experimental results show that remote voice detection via an LDV is promising when choosing appropriate targets close to human subjects in the environment.

1. INTRODUCTION

Multimodal/multi-sensor surveillance systems are widely deployed today for security purpose. Although a lot of progress has been made, particularly with the rapid improvements of color and infrared (IR) cameras and the corresponding algorithms for monitoring subjects at a large distance, audio information, as an important data source, has not yet been well explored. A few systems [1, 2] have been reported to integrate visual and acoustic sensors. But in these systems, the acoustic sensors need to be close to the subjects in monitoring. Parabolic microphones can be used for remote hearing and surveillance, which can capture voice at a fairly large distance in the direction pointed by the microphone. But it is very sensitive to noise caused by wind or sensor motion, and all the signals on the way get captured. Recently, Laser Doppler vibrometers have been widely used in the inspection industry. Laser vibrometers such as those manufactured by Polytec [4] and B&K Ometron [5] 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) [6-9]. 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, literature on remote voice detection using LDVs is rare. Therefore, the study of the novel application of an LDV for remote voice detection will be the main focus of this paper. A system with a color camera, an IR camera and an LDV has been described in our technical report [13].

The performance of the laser Doppler vibrometer strongly depends on the reflectance properties of the surfaces of the target that the laser beam is directed to. Important issues such as target surface properties and distances from the sensors are studied through several sets of indoor and outdoor experiments. The detected speech signals by the LDV may be corrupted by more than one noise source, such as laser photon noises, target movements, and background acoustic noises (wind, engine sound, etc.). 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 [11, 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 treatments.

This paper is organized as follows. Section 2 introduces the LDV principle and its use for voice detection at a large distance. Section 3 describes the acoustic signal enhancement for increasing intelligibility of the voice signals to human ears. Section 4 discusses experimental system design issues and presents some experimental results. Finally, we provide brief concluding remarks and some discussions in Section 5.

2. LDV PRINCIPLE AND AUDIO CAPTURE

Laser Doppler vibrometers 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 1), 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, l/2, which is 0.3169 μm (or 12.46 micro inches) in the case of helium-neon (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 following formula:

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

Objects vibrate while wave energy (including voice waves) is applied to them. Although 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. 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.

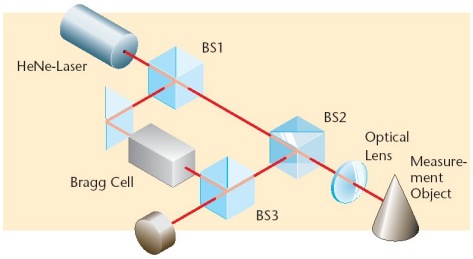

Figure 1: The modules of the Laser Doppler Vibrometer

We use a laser Doppler vibrometer from Polytec [4] that includes a controller OFV-5000 with a digital velocity decode card VD-6 and a sensor head OFV-505 (Figure 2). 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/V range, according to the manufacture's specification (with retro-tape treatment). Without retro-tape treatment, we have found that the LDV still has sensitivity on the order of 1 μm/s, i.e. one-thousandth of the full range.

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

The sensor head uses a particular HeNe red laser with wavelength of 633.8 nm and is equipped with a super long-range lens for long range listening. 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. We use the S/P-DIF output for obtaining signals of as highest quality as possible. We have also acquired a telescope VIB-A-P05 for accurate targeting of the laser beam at large distances.

3. LDV AUDIO SIGNAL ENHANCEMENT

The frequency range of human voice 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 (figure 3a). This renders the acoustic signals unintelligible to human ears. On the other hand, the inherent "speckle pattern" problem [3, 10] on a normal "rough" surface and the occlusion of the LDV laser beam by passing-by objects both introduce noises 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 volumes of the voice signals may change dramatically with changes in the vibration magnitudes of the target due to the changes of speech loudness (shouting, normal speaking, whispering), and also the distances of the human speakers to the target. Therefore, we have also designed an adaptive volume function to cope with this problem.

3.1 The Gaussian bandpass filter

To reduce the noise with frequency outside of normal speech frequency bandwidth, we produce a Gaussian bandpass transfer function by using the difference of two Gaussians of different widths, that is,

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

The impulse response of this filter is given by

$$h(t) = \frac{B}{\sqrt{2\pi\sigma_2^2}} e^{-t^2/2\sigma_2^2} - \frac{A}{\sqrt{2\pi\sigma_1^2}} e^{-t^2/2\sigma_1^2}, \quad \sigma_i = \frac{1}{2\pi\alpha_i} \qquad (3)$$

Notice that the broader Gaussian in the frequency domain creates a narrower Gaussian in the time domain, 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. (2), i.e., α_2, and the low frequency reduction is mainly controlled by the width of the second Gaussian function, i.e., α_1.

3.2. Volume selection and adaptation

The useful original signal obtained from the S/P-DIF output of the controller is a velocity signal. When taken as voice signal, the volume is too small to be heard by human ears. Furthermore, 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:

$$v = \frac{C_{\max}}{\left|\max(x_1, x_2 \ldots x_n)\right|} \qquad (4)$$

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 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.

The adaptive method will always give a suitable volume for any kind of the sampled data stream. However, because of

its non-continuous, it also introduces some artificial noise. In our LDV software system, both adaptive and fixed scaling methods are implemented, and the user can choose either method on the fly. Figure 3 shows a real example of filtering and scaling (corresponding audio clips can be found at [13]).

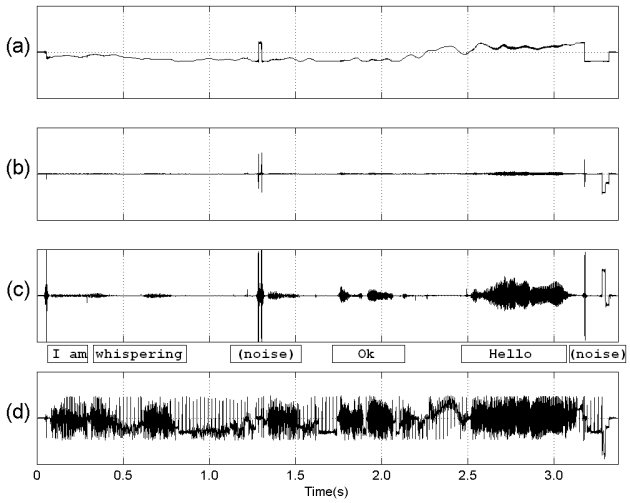

Figure 3: The waveform of (a) Original signal (b) x1 volume signal after band-pass filtering. (c) x8 volume signal after band-pass filtering. (d) Adaptive scaling after band-pass filtering.

4. SOFTWARE SYSTEM AND RESULTS

We have developed an integrated software surveillance system for remote multimodal surveillance. Based on the above algorithms, a speech enhancement sub-module is also incorporated into this software system for analyzing the detected LDV audio signals. We will introduce this system as well as the experiments results in the following two sub-sections.

4.1 System Design and Integration

We have developed the multimedia surveillance system in Java, which includes a video monitor module, an audio enhancement and analysis module and a configuration module. The video monitor module is used to control and display regular video (including both color and IR) information for detecting human subjects and for finding suitable reflective targets for LDV voice measurements at large distances. The configuration module is used to control the video sensors, as well as the LDV device through the RS232 serial port. Here, we will mainly focus on the real-time audio enhancement and analysis module, which is depicted in Figure 4.

There are a control panel and three graphical panels in this module. The control panel consists of all control components that are used for speech enhancement and analysis. For example, we can enable or disable low/high pass filter, and choose frequency thresholds for both filters by a drop-down menu. The selective volume and the alternative adaptive volume options are grouped in a radio box group. If the selective volume mode is chosen, the drop-down menu of the scale factors is enabled and an appropriate scale factor can be chosen interactively. All changes are available on the fly. The three graphical panels under the control panel are used for signal analysis, which display a waveform, its short-time Fourier transform (STFT) and spectrogram, respectively. The current processed frame is illustrated in red in the waveform, and its spectrum is shown in STFT graphical panel.

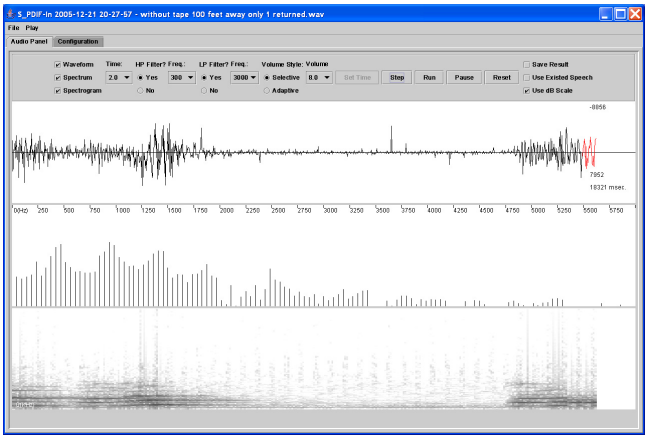

Figure 4. Audio enhancement and analysis module

4.2 Experiments Results 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 nearby the human speaker. Even though the vibration of most objects in man-made environments caused by waves of voices can be readily detected by the LDV, the LDV has to get signal returns from the laser reflection. The degrees of signal returns depend 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) we have used are capable of diffuse reflection 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 [13]. In all experiments, the LDV velocity range is 1 mm/s, and a person’s speech describes the experiment configurations. The same configurations (band-pass 300 – 3000 Hz, adaptive volume) are used in processing the data for all the experiments. In this paper we will only provide the results of one set of experiments: long range listening. More data collections and results can be found in our technical report [13]. Links to the audio files for both the original LDV audio clips and the processed audio clips (with one fixed configuration of filtering) are also included in the report. Most of 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.

We tested the long range LDV listening in an open space with various distances from about 30 to 300 meters [13]. 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, thanks 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.

5. CONCLUSIONS AND DISCUSSIONS

We investigate a novel sensor and signal enhancement techniques for voice acquisition at a large distance for remote and multimodal surveillance. We have found that the vibration of the objects caused by the voice energy reflects the voice itself. After the 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.

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. Besides, both objective and subjective performance evaluation based on these enhancement approaches should be conducted. The objective evaluation can include spectrogram comparison and segmental signal-noise ration (SNR), which is well known as correlated with the subjective perception of speech quality [14]. These are our ongoing work.

In the current implementation, the received signal is designed to bandpass in the 300Hz-3kHz range based on the assumption that speech is bandlimited to that range. However, speech can have significant spectral components up to as much as 8kHz. Speech bandpassed at the 300Hz-3kHz range is generally still intelligible but many sounds with higher frequency components, such as fricatives, lose much of their discriminative information. In Section 3, we have discussed that there is low frequency noise from the LDV targets, and high frequency noise from occlusion of the LDV laser beam. The spectral characteristics of these noise sources need to be investigated further. Therefore, the bandpass range of the filter might be automatically determined rather than simply choosing the 300Hz-3kHz range.

6. ACKNOWLEDGEMENTS

This work is supported by AFRL under Grant No F33615-03-1-6383, and partially supported by NSF under Grant No. CNS-0551598 and Grant No. CNS-0424539. We would like to thank Dr. Esther Levin and Dr. Robert Haralick for many fruitful discussions, and for Mr. Hao Tang for the help in collecting the data. We are grateful to anonymous reviewers for their insightful comments in improving the paper and constructive suggestions for some future work.

7. REFERENCES

[1].D. Zotkin, R. Duraiswami, H. Nanda, L. Davis, "Multimodal tracking for smart videoconferencing," Second International Conference on Multimedia and Expo, Tokyo, Japan, 2001.

[2]. X. Zou and B. Bhanu, Tracking humans using multimodal fusion, The 2nd Joint IEEE International Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum (OTCBVS'05), San Diego, CA, US, June 20, 2005

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

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

[5].Ometron Systems. http://www.imageautomation.com/

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

[7].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, pp. 773-783, 2003

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

[9]. 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, 2001, pp. 617-626

[10].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

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

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

[13].Z. Zhu, W. Li, Integration of laser vibrometer and infrared video for multimedia surveillance display, TR-2005006, CUNY Graduate Center, 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.

[14].I. Cohen, Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging, *IEEE Trans. Speech and Audio Processing*, vol. 11,pp. 466-475, Sep. 2003.