Accepted to Proceedings of International Conference on Pattern Recognition *(ICPR'06)*, Hong Kong, China, 20-24 August 2006

# LDV Remote Voice Acquisition and Enhancement

Weihong Li[1*], Ming Liu[2+], Zhigang Zhu[3, 1*], Thomas S. Huang[2+]
*[1]Department of Computer Science, Graduate Center, CUNY, NY 10016*
*[2]Beckman Institute for Advanced Science and Technology, UIUC, Urbana, IL 61801*
*[3]Department of Computer Science, City College, CUNY, NY 10031*
**{wli, zhu}@cs.ccny.cuny.edu, +{mingliu1, huang}@ifp.uiuc.edu*

## Abstract

*Laser Doppler vibrometers (LDVs) have been widely used in industry inspection. One of the superior characteristics of an LDV is that it can detect and measure extremely tiny vibration of a target at a large distance, with sensitivity on the order of 1µm/s. On the other hand, we have found that most objects nearby audio sources can be vibrated by the audio waves. These two aspects motivate our research in a new application of the LDVs, namely remote voice detection from surrounding vibrated objects. However, the detected speech signals may be corrupted by many noise sources, such as laser photon noises, target movements, and background acoustic noises (wind, engine sound, etc.). Therefore, speech enhancement algorithms based on Gaussian bandpass and Wiener filters are designed to effectively improve the intelligibility of the noisy voice signals detected by the LDV system Experimental results show that remote voice detection via an LDV is very promising, when choosing appropriate targets close to human subjects and using the proposed enhancement techniques.*

## 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 fully 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 could 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 the noise caused by wind or sensor motion, and all the signals on the way get captured. Recently, laser Doppler vibrometers (LDVs) have been widely used in industry inspection. 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. For example, they 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 [10].

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 have been studied through several sets of indoor and outdoor experiments in our previous research [10]. As one of the most important research problems, 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 needed in order to improve the performance of recognizing a noisy voice detected by the LDV system. Many speech enhancement algorithms have been proposed [7, 8], but they have been mainly used for improving the performance of speech communication systems in noisy environments. Acoustic signals captured by laser vibrometers need some special treatments.

This paper is organized as follows. Section 2 briefly introduces the LDV principle and its use for voice detection at a large distance. Section 3 describes our acoustic signal enhancement techniques effective 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 in Section 5.

## 2. LDV Principle for Audio Capture

A laser Doppler vibrometer (LDV) works 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 [4] (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 mm (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 voice waves can be readily detected by the LDV.

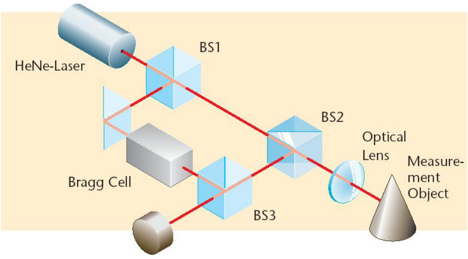


**Figure 1: The modules of the Laser Doppler Vibrometer**

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 manufacturer'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. 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. We use the S/P-DIF output of the controller 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.

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

## 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. On the other hand, the inherent "speckle pattern" problem [3, 6] on a normal "rough" surface and the occlusion of the LDV laser beam by passing-by objects both introduce noise spreading across the voice frequency range. This creates undesirably loud noise when we directly listen to the acoustic signal. 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, and also the distances of the human speakers to the target. Therefore, we apply some particular speech enhancement techniques to cope with these problems: bandpass filtering, Wiener filtering and volume adaptation.

### 3.1 Gaussian bandpass filtering

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, as has been widely used in image processing [9], i.e.

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

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. (3), i.e., $\alpha_2$, and the low frequency reduction is mainly controlled by the width of the second Gaussian function, i.e., $\alpha_1$.

### 3.2 Wiener filtering

After applied the bandpass filter, the noise outside of the voice frequency is attenuated, but the noise energy falling inside the voice frequency range still exists. This problem is handled using Wiener filtering, which is one of the most effective speech enhancement approaches in literature [11].

Wiener filtering is an optimal filtering technique in the sense of minimum mean squared error (MMSE) criteria. The degraded speech can be modeled as the summation of a clean speech signal and additive noise which is illustrated as

$$y[n] = s[n] + d[n] \qquad (4)$$

where $d[n]$ is additive noise signal, and the clean signal $s[n]$ is independent of the noise $d[n]$. With this assumption, their autocorrelation signals satisfies

$$R_{yy}[n] = R_{ss}[n] + R_{dd}[n] \qquad (5)$$

Since the power density functions are the Fourier transforms of the corresponding autocorrelation signals, so

$$S_{yy}(\omega) = S_{ss}(\omega) + S_{dd}(\omega) \qquad (6)$$

We know that $S_{yy}(\omega) = |Y(\omega)|^2$, where $Y(\omega)$ is the Fourier transform of $y[n]$, this leads to

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \qquad (7)$$

Under MMSE criteria, the optimal linear filter for speech enhancement is the Wiener filter, and its frequency response is

$$H(\omega) = \frac{S_{ss}(\omega)}{S_{ss}(\omega) + S_{dd}(\omega)} \qquad (8)$$

Since we have already estimated the $\hat{S}_{dd}(\omega)$, the frequency response of the optimal filter is

$$H(\omega) = 1 - \frac{\hat{S}_{dd}(\omega)}{S_{yy}(\omega)} \qquad (9)$$

The estimate frequency response could have negative values, so in practice zeros are used to replace the negative values. In Wiener filtering, the output signal has the following formula,

$$\hat{s}[n] = IFFT\left\{\frac{|Y(\omega)|^2 - |\hat{D}(\omega)|^2}{|Y(\omega)|^2} \cdot |Y(\omega)| \, e^{j\theta_{sn}(\omega)}\right\} \qquad (10)$$

where $\theta_{sn}(\omega)$ is the phase information of the degraded signal $y[n]$. In practice, the noise is only stationary in a short period of time. In order to update the optimal filter, adaptive noise spectrum estimation is conducted using a moving window which contains one second of signal. The noise part is assumed to be the low-energy part (20%) of this signal. Then the noise power spectrum ($\hat{D}(\omega)$) is estimated using the noise part of this signal. The optimal filter is therefore updated based on this new noise estimation.

### 3.3 Volume 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:

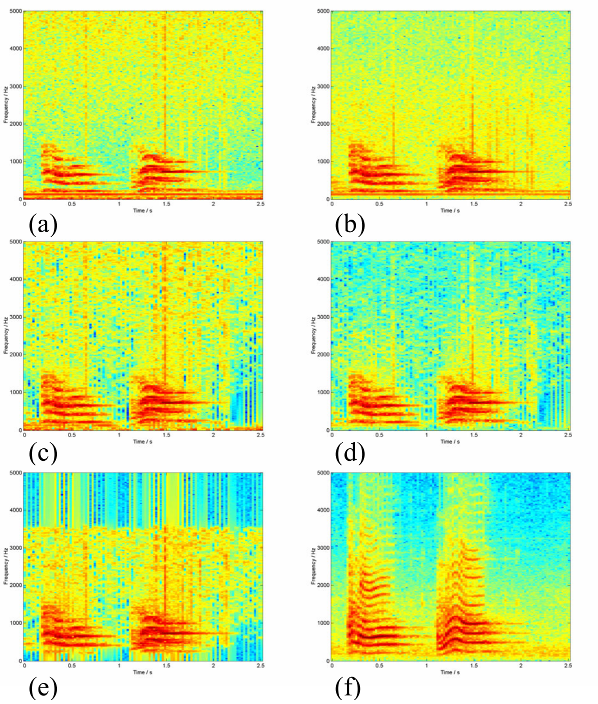
$$v = \frac{C_{\max}}{\left|\max(x_{1,}x_2 ... x_n)\right|} \qquad (11)$$

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.

## 4. Performance Evaluation and Analysis

To evaluate the performance of the speech enhancement by the proposed techniques, both subjective and objective evaluation can be conducted. We mainly focus on objective approach in this paper, which is based on two criteria, namely spectrogram comparison and segmental signal-noise ratio (SNR).

For the first criterion, we compare the spectrograms of LDV audio, its enhanced speech signals, and a corresponding clean signal captured at the same time using a wireless microphone. The LDV audio signal in Figure 3 was captured 100 feet away by aiming the laser beam at a metal cake box (without retro-reflective finish), and the clean signal was captured using the wireless microphone connected to a laptop placed next to the target (i.e., the metal box). The clean signal (Figure 3f) was aligned with LDV signal.



(a) (b) (c) (d) (e) (f)

**Figure 3: The spectrogram of (a) original LDV signal (b) Gaussian bandpass filtered signal (c) Wiener filtered signal (d) Wiener filtered + Gaussian bandpass filtered signal (e) Wiener filtered + Hann bandpass filtered signal (f) clean signal. All correspond to the speech of "Hello, Hello".**

As can be seen in the spectrograms, most of high frequency energy in clean signal (Figure 3f) disappeared in LDV audio signal (Figure 3a). This is because the measured object vibrated by voice energy could not vibrate in such high frequency as air or the eardrum does. In Figure 3a, we can also see strong noise in the low frequency part (red part at the bottom), as well as relative weak noise in high frequency (yellow part at the top). Figure 3b shows the enhanced signal spectrogram by Gaussian bandpass filter, which reduces both the low and high frequency noise. Figure 3c shows the enhanced signal by the Wiener filter only. Figure 3d shows that the noise is largely attenuated by the combined approach, which is the application of the wiener filter followed by the Gaussian bandpass filter. As a matter of fact, we have found that the order of the combination does not matter much. Figure 3e shows Hann bandpass filter, which almost removes the noise with frequency lower than 300 Hz or higher than 3 KHz, but introduces a lot of artificial noise in the middle of the speech frequency. In comparision, the Gaussian plus Wiener filtering gives the best result (3d), which is the closest to the clean signal (3f).

Another criterion we used for objective evaluation is to compare the SNR values of LDV audio signals enhanced by various methods. We use the segmental SNR as our distortion measurement, which is well known as correlated with the subjective perception of speech quality [12]. The segmental SNR is defined as

$$SegSNR = \frac{10}{|L|}\sum_{l\in L}\log\frac{\sum_{k}|S(k,l)|^2}{\sum_{k}|D(k,l)|^2} \qquad (12)$$

where $k$ represents the frequency bin index, and $l$ the frame index. $L$ represents the set of frames that contain speech, and $|L|$ its cardinality. Entities $|S(k, l)|^2$ and $|D(k, l)|^2$ represent the Fourier transforms in $k$th frequency bin of the $l$th frame of clean speech and background noise (i.e. the frame without speech), respectively.

The results are shown in the Table 1. The segmental SNR values are computed based on Eq. (12) for original signals and enhanced signals by the corresponding methods, namely Gaussian bandpass only, Wiener filter only, and the combined approach. Two possible combination strategies, i.e., Bandpass filter followed by Wiener filter (BW) and Wiener filter followed by Bandpass (WB), are conducted and the results are shown in the last column of Table 1. Three different types of reflecting surfaces are tested: the small empty mental cake box (with retro-tape), the mental box surface itself (without retro-tape), and the wood hose box surface. They are all 100 feet away from the sensor head. The signal captured from the retro-tape surface contains strong low frequency noise but the voice is much clearer to percept than other two signals captured on the surfaces that produce strong high frequency noise. Several audio clips before and after processing are provided in the supplemental material (ICPR06-LDV-Suppl.rar). From the results, we can see that the combined approach largely improves the original signals and also outperforms the methods using only one of the filters. In addition, both BW and WB perform almost equally well for cases with high frequency noise (from nature surfaces as shown in the bottom two rows). For the case with obvious low frequency noise (reflected from the retro-tape), Gaussian bandpass filtering is more significant than the Wiener filtering. Therefore, for this case, bandpass filter followed by Wiener filter outperforms the other one. The reason is that the Wiener filtering has the best effect after the high-energy low-frequency part has been removed by the Gaussian bandpass filtering.

**Table 1: The segmental SNRs (dB).**

| | Original | Bandpass | Wiener | BW/WB * |
|---|---|---|---|---|
| Box w/ tape | 66.5 | 87.8 | 68.4 | 92.1/83.4 |
| Box w/o tape | 66.1 | 75.6 | 71.3 | 85.7/85.2 |
| Hose box | 64.4 | 72.5 | 66.8 | 80.6/78.4 |

*BW: Gaussian Bandpass followed by Wiener filter
WB: Wiener filter followed by Gaussian Bandpass

## 5. Conclusions

We have investigated a novel sensor (i.e., the LDV) and a set of signal enhancement techniques particularly effective for voice acquisition at a large distance for remote surveillance. We have found that the vibration of the objects caused by the voice energy reflects the voice itself. We have proposed a hybrid enhancement approach, which combines the Gaussian bandpass and Wiener filtering, followed by the adaptive volume scaling, to improve the intelligible of the detected LDV audio signals. Objective evaluation shows that this hybrid approach outperforms other approaches. As a next step, we are preparing for a subjective evaluation to verify the proposed approach.

With the combined filtering and adaptive volume scaling, the LDV voice signals are mostly intelligible from targets without retro-reflective tapes at a relative large distance (100 m). 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 very large distances. With current state-of-the-art sensor technology, we realize that more advanced signal enhancement techniques need to be developed. For example, model-based voice signal enhancement could be a solution in that background noises might be captured and analyzed, and better noise modeling in Eq. (10) could be developed from the resulting data. This is our ongoing work.

## Acknowledgements

This work is supported by AFRL under Grant No F33615-03-1-6383, and by a CUNY Equipment Grant Competition Award. We would like to thank Lt. Jonathan Lee, Dr. George Wolberg, Dr. Esther Levin and Dr. Robert Haralick for many fruitful discussions, and Mr. Hao Tang for the help in collecting the data.

## 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, Adam Hilger, 1990

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

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

[6]. 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

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

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

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

[10]. Z. Zhu, W. Li, Integration of laser vibrometer and infrared video for multimedia surveillance display, TR-2005006, CS Dept, CUNY Graduate Center, April 2005.

[11].Y. Ephraim, H. Lev-Ari and W.J.J. Roberts, A Brief Survey of Speech Enhancement, in *CRC Electronic Handbook*, 2nd edition, CRC Press, Feb. 2005.

[12]. 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.