Long-distance audio event detection using a laser Doppler vibrometer

Tao Wang, Zhigang Zhu, Yufu Qu, and Ajay Divakaran

Long-range, non-contact acoustic measurement devices can be used to obtain the audio signals of a human, a vehicle, or another object.

Long-range audio signature acquisition and event detection are always a challenge using existing microphone arrays, which have a short detection range and therefore must be placed close to their targets. A parabolic microphone can capture voice signals at a fairly large distance, but it needs to point directly at the target, and all signals along the path are captured. A laser Doppler vibrometer (LDV), on the other hand, is a long-range, non-contact acoustic measurement device that can be used to obtain the audio signals of a target at a great distance. This is achieved by detecting surface vibrations caused by the sound of the target in a nearby reflecting surface toward which the laser points.

In the past few years, we have studied long-range voice detection using an LDV and developed a multimodal system that includes a pan-tilt-zoom (PTZ) camera to help refine the pointing direction. We have found that the quality of the signal acquired from the LDV is determined mainly by the reflective and vibrational properties of selected background surfaces near the target. Therefore, we have also built robust acoustic background models of various reflecting surfaces to detect acoustic events. As an example, we have applied these models to long-range audio and video event integration.

In our system setup, shown in Figure 1, a PTZ camera is mounted on top of the LDV to find the best reflection surface in the image. The LDV works according to the principle of laser interferometry. We use an LDV with a helium-neon red laser at a wavelength of 633.8nm and a velocity detection range of 1mm/s for sensing voice vibrations. A mirror mounted on top of a pan-tilt unit (PTU) reflects the laser beam to points at different locations. The PTU provides the pan and tilt angles needed to generate a laser point observed in the PTZ camera’s field of view. As a result, each laser point is associated with a 4D vector: the x- and y-pixel coordinates in the PTZ image and the pan and tilt angles of the PTU. Therefore, the 3D location of the laser point on the selected surface can be obtained using triangulation. This provides two advantages: fast focus of the laser beam using distance measurements and fast selection of a reflective surface according to the location of a moving target. Then, the audio signals of various surfaces at different locations near the target of interest are collected.

The LDV can detect acoustic signals from various vibrating surfaces, including window frames, concrete walls, and traffic signs. However, a reliable acoustic background modeling technique is needed to separate the outliers from the background sound, which includes both the real background sound and the signals created by the electronic-optical noises of the LDV. A Gaussian mixture model (GMM) is typically used to model the feature distributions of signals. Each component of the mixture in a surface acoustic model is represented by a unique Gaussian mean vector and a covariance matrix. However, the GMM does not build relationships among different mixture components in a surface model, and components in another surface model may...
Figure 2. Audio-video integration in a 420-foot corridor.

be very similar to them. To represent the temporal dependencies of components in a surface model, we use a window-based aggregation technique for a GMM with more than one component. The basic idea is to select a sequence of overlapping windows, each containing consecutive features in a time series. Then we construct a normalized histogram on the basis of how those features are identified. In general, a feature is determined to be either a background component or part of the foreground. The average of all constructed window-based histograms for the correct background model creates a temporal pattern that can be used to evaluate any input signals.

As an example, we constructed audio background models for various surfaces including a metal box, painted metal door, chalkboard, whiteboard, and wall. We tested the models in an indoor corridor about 420 feet long for long-range audio-visual event detection. Foreground audio events were extracted using the corresponding background surface model. The target person in the video was detected using a standard background subtraction technique in computer vision. Results from the audio and video were combined to demonstrate the final event determination. In Figure 2, the red box indicates a person behind the wall who cannot be observed in the image but whose speech is detected from the audio stream (in the shaded region). The blue box represents people detection in the camera view.

Long-range audio signal acquisition and event detection using an LDV offers benefits for many applications, such as surveillance, search and rescue, and military applications. Audio background modeling based on the properties of surface vibration improves the signal quality for audio event detection using an LDV. When this is combined with video information, the final decision can be further validated in both modalities. In future work, we will develop more advanced audio-visual integration methods considering both the signal characteristics and the LDV sensor properties.

Author Information

Tao Wang, Zhigang Zhu, and Yufu Qu
Department of Computer Science
The City College and Graduate Center
The City University of New York
New York, United States

Tao Wang is a PhD student. Since 2006, he has been a research assistant in the City College Visual Computing Research Laboratory (CCVCL), working on multimodal sensor design and integration and audio and video surveillance.

Zhigang Zhu is a full professor whose research interests include 3D computer vision, multimodal sensing, and video representation. He has published over 100 technical papers in related fields.

Yufu Qu received his PhD in instrument science and technology from Harbin Institute of Technology, China, in 2004. Since 2008, he has been visiting the CCVCL. His research interests include optical sensing and multimodal sensor design and integration.

Ajay Divakaran
Vision and Multi-Sensor Systems
Sarnoff Corporation
Princeton, United States

Ajay Divakaran is a technical manager. Before joining Sarnoff Corporation, he worked at Mitsubishi Electric Research Laboratories from 1998–2008. His research interests include multimedia content analysis, audio analysis, and computer vision. He has two books and over 100 journal and conference publications to his credit.

Continued on next page
References


