Real-time Sound Source Localization
Based on Audiovisual Frequency Integration

Tokuo Tsuji, Kenkichi Yamamoto, and Idaku Ishii
Department of Artificial Complex Systems, Hiroshima University, 1-4-1 Kagamiyama,
Higashi-Hiroshima 739-8527, Japan
 ttsuji@hfl.hiroshima-u.ac.jp

Abstract

We propose a pixelwise sound source localization algorithm based on audiovisual frequency integration. The localization is realized by detecting the common vibration dynamics of sound sources in the audio and the brightness signal. In order to detect the common vibration dynamics, temporal correlation values between the two signals are calculated in the algorithm. Several experimental results are shown for vibrated objects, and the pixelwise sound source localization images are obtained.

1. Introduction

Sound waves are compressional waves generated by a periodic change in pressure in a medium such as air. The periodic change is directly related to the physical vibration dynamics of sound sources, such as the resonance phenomenon. Therefore, the temporal change in auditory information converted from sound waves is strongly correlated with the physical movements of the sound sources. The auditory sensation area in human beings is from 20 Hz to about 20 kHz in the temporal frequency domain. While most auditory sensors such as microphones cover this domain, most of the vision sensors cannot do so because of the limitation of the video signal. Most of the conventional vision sensors are designed to display images for human eyes, and their frame rates are constrained by the video signal (NTSC 30 Hz/PAL 25 Hz).

If vision sensors are speeded up, the physical vibrations of the sound sources can be observed by them. The brightness values at the pixels around the sound sources will then change periodically. The periodic changes in the brightness values correspond to the dynamically moving positions of the sound sources. Therefore, both audio and brightness signals contain common vibration dynamics when both vision and auditory sensors operate at sampling rates of the order of kHz or higher. The common dynamics of the two signals is caused by the physical vibrations of the sound sources.

In this paper, we propose a correlation-based algorithm of pixelwise source localization. The algorithm is implemented on a real-time high-speed vision system. Several experimental results of real-time source localization are shown, and the obtained pixelwise localized images for the sound sources are evaluated.

Our algorithm is useful for many industrial applications. Vibrations invisible to the human eye such as the abnormal sound source of machines in factories and low frequency noise in the outdoors that cause pollution. Auditory sensors can detect the presence of such abnormal sounds or noises, while they cannot localize the invisible vibrations accurately because of their low directivity.

2. Pixelwise source localization

2.1 Correlation-based Pixelwise Source Localization

We present pixelwise source localization using correlation values between vision and auditory sensor information. Source localization by using microphones has already been reported in many works [1, 2]. In general, such source lo-
Correlation-based Source Localization Image

Here, we describe a source localization image as a pixelwise source localization map, which is obtained by calculating the temporal correlation values at each pixel. Let the audio signal at time \( t \) be \( g(t) \), and the brightness signal of a vision sensor at pixel \((x, y)\) and time \( t \) be \( \hat{I}(x, y, t) \). We then define \( F(x, y; t_1, t_2) \) as a source localization image. The image is calculated using the correlation values between the audio and brightness signals from time \( t_1 \) to \( t_2 \) at all the pixels.

\[
F(x, y; t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \hat{I}(x, y, t) \hat{g}(t) \, dt \quad (1)
\]

Here, \( \hat{I}(x, y, t) \) and \( \hat{g}(t) \) are obtained by subtracting temporal averages as shown below. The temporal averages are dc components that do not depend on the vibration dynamics of the sound sources.

\[
\hat{I}(x, y, t) = I(x, y, t) - \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} I(x, y, t) \, dt \quad (2)
\]
\[
\hat{g}(t) = g(t) - \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} g(t) \, dt \quad (3)
\]

2.2. Frequency Properties of Temporal Correlation

For a source localization image, the audio signal \( g(t) \) and the brightness signal \( I(x, y, t) \) can be described in the frequency domain as follows.

\[
\hat{g}(t) = \sum_{i=1}^{\infty} \alpha_i \sin(2\pi f_i t + \phi_i) \quad (4)
\]
\[
\hat{I}(x, y, t) = \sum_{i=1}^{\infty} \beta_i(x, y) \sin(2\pi f_i t + \psi_i(x, y)) \quad (5)
\]

Here, \( f_i = i f_s \) is described discretely as the \( i \)-th frequency \((i: \text{non-zero integer}, f_s: \text{constant})\). \( \alpha_i \) and \( \phi_i \) are the amplitude and phase of the audio signal at frequency \( f_i \), respectively. \( \beta_i(x, y) \) and \( \psi_i(x, y) \) are the amplitude and phase of the brightness signal at pixel \((x, y)\) and frequency \( f_i \), respectively.

Next, we describe the relation between the frequency components and temporal correlation values. By substituting Exp. (4), (5) into Exp. (1) and using a trigonometric formula, the following relation is obtained for a source localization image of \( F(x, y) \). Here, \( t_1 = -T, t_2 = T \), and we suppose that \( T \to \infty \).

\[
F(x, y) = \frac{1}{2T} \int_{-T}^{T} \hat{I}(x, y, t) \hat{g}(t) \, dt = \left| \sum_{i=1}^{\infty} \alpha_i \beta_i(x, y) \cos(\phi_i - \psi_i(x, y)) \right| \quad (6)
\]

The correlation value is positive as long as the audio and brightness signals contain a common non-zero frequency component at a certain frequency. This holds true even if the waveforms of the signals do not coincided completely.

3. Real-time Experiments for Pixelwise Source Localization

3.1. Pixelwise Source Localization System

Here, we introduce a real-time source localization system that uses high-speed vision. The source localization system developed by us consists of a high-speed vision system, a microphone, and a PC(CPU: Intel Celeron 2.4 GHz, Memory: 512 MB, and OS: Microsoft Windows 2000). Fig. 2 shows an overview of the system.

The MmVision (mega-pixel and milli-second vision) prototype[3] is a real-time high-speed vision system in which the required image areas can be intelligently selected and processed in the order of kHz. The audio signal is obtained by using a dynamic microphone, and it is transferred to a personal computer via a sound card, Sound Blaster Live 5.1. This computer calculates temporal correlation values.
between the brightness and audio signals. The source localization images can be monitored in real-time on an LCD display connected to the computer.

3.2. Source localization

Several source localization experiments are performed under various conditions. In the experiments, 8-bit gray-level images are captured at a frame rate of 1000 Hz. Audio signals are recorded as 16-bit monaural sound, and they are sampled at the rate of 1000 Hz.

The 6th String Vibration For the string vibration of a guitar, image size is set to $256 \times 16$ pixels. First, we show the source localization result for the vibration of the 6th string. The fundamental frequency of the 6th string is about 82 Hz. Three strings – 4th, 5th and 6th string – are observed in the captured images. $t = 0.484$ s is the time at which the 6th string starts to vibrate. Fig. 3 is an image sequence for $t = 1.500 \sim 1.510$ s. The interval is 0.002 s.

A coordinate value $(x, y)$ represents a pixel position. Three pixels – $P_1(25,0), P_2(53,0)$, and $P_3(100,0)$ – are selected as shown in Fig. 3. Fig. 4 shows the brightness values at the three pixels and the audio signal for $t = 1.500 \sim 1.600$ s.

The brightness values change in a period of approximately 0.012 s at $P_1$ and $P_2$. The periodic change in brightness depends on the string vibration because $P_1$ and $P_2$ are located around the 6th string. A similar periodic change is observed in the audio signal. On the other hand, $P_3$ is located in an area that is unaffected by the 6th string vibration; therefore, we observe no vibration at $P_3$.

Fig. 5(a) shows a source localization image obtained by calculating temporal correlation values between the brightness and audio signals. The correlation values are calculated for $t = 1.500 \sim 2.000$ s. Here, black pixels are assigned when the correlation values are high; and white ones when they are low. Fig. 5(b) shows the intersection in the horizontal direction $(y = 0)$.

With a Dynamically Moving Object We show the source localization result when there is a periodically moving object near the guitar strings. A fan rotating at about 50 rpm is placed near the vibrating 6th string, as shown in Fig. 6. A rectangular area corresponding to the fan and the 6th string is observed in the captured images. Fig. 7 is an image sequence for $t = 2.500 \sim 2.505$ s. The interval is 0.001 s. At $t = 1.313$ s, the 6th string starts to vibrate.

Fig. 8(a) shows a source localization image. The correlation values are calculated for $t = 2.500 \sim 3.000$ s. Fig. 8(b) shows the intersection in the horizontal direction $(y = 0)$. The correlation values are high at several pixels around the 6th string, and low at the pixels around the rotating fan. Thus, pixelwise source localization is correctly realized even if a dynamically moving object exists near the vibrating string.

Vibrating Motor Source localization experiments are carried out for a vibrating motor. In these experiments, images of $96 \times 96$ pixels are captured. We show the source localization result when a motor vibrates steadily at about
70 Hz. The size of the motor is $21 \times 6$ mm, and the motor is fixed on the table using couple-face tapes. The rectangular area shown Fig. 9 is captured using high-speed vision. At $t = 0.710$ s, the motor starts to vibrate.

Fig. 10 shows the brightness values at three pixels and the audio signal for $t = 1.000 \sim 1.100$ s. The selected pixels are $P_1(36,22)$, $P_2(45,40)$, and $P_3(10,70)$ as shown in Fig. 9. The brightness values change in a period of about 0.014 s at $P_1$ and $P_2$. The periodic change in brightness depends on the vibration of the motor because $P_1$ and $P_2$ are located near the motor. A similar periodic change is observed in the audio signal. On the other hand, $P_3$ is located in the unaffected by the vibration of the motor.

Fig. 11(a) shows a source localization image. The correlation values are calculated for $t = 1.000 \sim 1.500$ s. Fig. 11(b) shows a three-dimensional plot of the source localization image. The correlation values are high at several pixels around the motor, and the values are low at the pixels in the area unaffected by the motor. Thus, pixelwise source localization is correctly realized for a vibrating motor.

4. Conclusion

In this paper, we have introduced a correlation-based algorithm for pixelwise source localization. In this algorithm, sound source areas are localized by detecting the common vibration dynamics between the brightness and audio signals. The source localization algorithm has been implemented on a real-time vision system with frame rate of 1000 Hz. We have also presented basic experimental results for several vibrating objects. We are currently planning to apply source localization algorithms to several practical cases such as the abnormal sound localization of machines in factories, low-frequency noise localization in the outdoors.

References