Audio-Vision: Using Audio-Visual Synchrony to Locate Sounds

John Hershey *
jhershey@cogsci.ucsd.edu
Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92093-0515

Javier Movellan
movellan@cogsci.ucsd.edu
Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92093-0515

Abstract

Psychophysical and physiological evidence shows that sound localization of acoustic signals is strongly influenced by their synchrony with visual signals. This effect, known as ventriloquism, is at work when sound coming from the side of a TV set feels as if it were coming from the mouth of the actors. The ventriloquism effect suggests that there is important information about sound location encoded in the synchrony between the audio and video signals. In spite of this evidence, audiovisual synchrony is rarely used as a source of information in computer vision tasks. In this paper we explore the use of audio visual synchrony to locate sound sources. We developed a system that searches for regions of the visual landscape that correlate highly with the acoustic signals and tags them as likely to contain an acoustic source. We discuss our experience implementing the system, present results on a speaker localization task and discuss potential applications of the approach.

Introduction

We present a method for locating sound sources by sampling regions of an image that correlate in time with the auditory signal. Our approach is inspired by psychophysical and physiological evidence suggesting that audio-visual contingencies play an important role in the localization of sound sources: sounds seem to emanate from visual stimuli that are synchronized with the sound. This effect becomes particularly noticeable when the perceived source of the sound is known to be false, as in the case of a ventriloquist's dummy, or a television screen. This phenomenon is known in the psychophysical community as the *ventriloquism effect*, defined as a mislocation of sounds toward their apparent visual source. The effect is robust in a wide variety of conditions, and has been found to be strongly dependent on the degree of "synchrony" between the auditory and visual signals (Driver, 1996; Bertelson, Vroomen, Wiegeraad & de Gelder, 1994).

*To whom correspondence should be addressed.