TR - H - 139

Learning to Localize Sounds Using Vision

Ed Gamble
David Rainton

1994. 4. 3

ATR人間情報通信研究所
〒619-02　京都府相楽郡精華町光台2-2　☎ 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

Learning to localize sounds using vision

E. B. Gamble, D. Rainton
ATR Human Information Processing Laboratories
2-2 Hikaridai
Seika-cho Soraku-gun
Kyoto 619-02
Japan
E-mail rainton@hip.atr.co.jp

March 31, 1995

Abstract

Auditory visual spatial registration is a prerequisite for any subsequent data fusion. This paper describes an algorithm for autonomous learning of a common auditory visual perceptual space.

Contents

1 Introduction

With the recent emergence of multi-media information systems comes the potential for the synergistic assimilation of data from multiple sensory channels. Sight and sound are arguably those modalities of greatest importance for human/machine communication. Developing algorithms for exploiting auditory visual multisensor integration and fusion is one of our main goals here at ATR.

The importance of auditory visual interaction has long been recognized in the experimental psychology community. However it is only relatively recently that attempts have been made to exploit such interactions for improving the man/machine interface [19] [13]. This being mainly due to the prohibitively expensive nature of both the necessary input devices and computational hardware. However, recent advances in both technologies now permit a new level of sophistication for practical multisensor data fusion [18] [17].

Interest in auditory visual data fusion arises primarily from the observation that auditory and visual information sources are complementary and any noise sources largely orthogonal. Humans in particular are highly adept at exploiting both the complementary and redundant information provided by their eyes and ears, relying heavily on visual cues in acoustically noisy, ambiguous or reverberant environments [15] [2]. Constructing multisensory systems with both acoustic and visual inputs will facilitate the perception of features which are difficult or impossible to obtain independently from either modality in isolation.

However in order to exploit the synergistic combination of both acoustic and visual information it is important that object perception occur within a common perceptual space. For example, the perceived visual location of a speaker's mouth should coincide with the perceived acoustic locus of any uttered sounds. This perceptual spatial registration is a fundamental prerequisite for any subsequent data fusion.

To date the question as to how to achieve such spatial registration has been largely ignored. Typically it is assumed that the object of interest is trivially centered in both the acoustic and visual fields. Learning the spatial correspondence between randomly located objects is still an open research issue.

Given the seemingly effortless way in which humans and animals both learn and exploit auditory visual spatial registration we begin with a brief

overview of the relevant psychophysiological literature. The reader should be aware however that this summary is far from comprehensive and reflects mainly the interests and familiarity of the authors.

2 Visual factors in human auditory spatial perception

The acoustic sensory space does not project directly onto the sensory surface of the ear, as it does for example in the case of the eye. As a consequence the auditory system is forced to derive spatial information indirectly from such cues as inter-aural time delays and inter-aural intensity differences. Given however that these spatial cues depend on both the size and shape of the head and ears, how is the cue/position mapping learnt and maintained? The evidence summarized in this section suggests that perceived acoustic spatial location is directly influenced by perceived visual location.

Experimental psychologists have long been studying the influence of visual factors in human auditory spatial perception. Some of the earliest documented experiments were those in which subjects visual worlds were "reversed" by means of 180 degree rotating prismatic glasses [3] [16]. After a few days of continuously wearing such glasses subjects found that sounds were localized as coming from where the source was seen, as opposed to its real physical location.

A variation of the above theme employed a "pseudophone" to reverse the subjects auditory field by 180 degrees [21]. In the absence of vision, it was found that sound localization was similarly reversed. Yet with normal sight sounds were heard as originating from their real locations. More recently in a less extreme version of the above experiments it has been shown that a subject's auditory mid-line setting for dichotically-presented clicks is also directly influenced by visual displacements produced by prismatic glasses [11].

In another experiment subjects were sat stationary in the center of a rotating circular screen [20]. After a while, due to the motion in their visual field, the subjects perceived themselves as rotating in the opposite direction to the screen. After this state of self-induced ego-movement had been attained, a sound was played at some distance beyond the screen from directly in front. However, subjects perceived the sound source as originating

vertically above, this being a location in which the relative position of the sound source would remain fixed if the subjects really were in motion, as they believed themselves to be.

Other experiments have shown that the accuracy of auditory localization is increased if subjects are allowed to move their eyes in the direction of the target [6]. Hence our ability to point to the source of a sound more accurately in the light than in the dark.

The physiological maps of auditory and visual space are mutually aligned, with the visual map dominant, in the sense that visual spatial distortions tend to introduce acoustic spatial distortions, but not visa-versa. The next question is how is this alignment learnt?

3 Visual supervision of auditory localization learning

Although it is not well understood how humans acquire their auditory visual spatial alignment, experiments have shown that new born infants as young as 2 days old are capable of directing their visual attention to off center sounds [12]. Also it has been shown that infants become visibly distressed upon observing their mothers speak to them while the mother's voice is displaced in space [1]. Their ability to perceive this discrepancy indicates the existence of some initial, albeit crude, perceptual alignment.

The situation is better understood in the case of barn owls where Knudsen and his colleagues have done a great deal of work in elucidating how the auditory and visual senses combine [8] [7] [9]. Their work suggests that although owls are capable of crude acoustic localization soon after birth, improving this initial crude localization capability is a supervised learning process, with the visual system acting as the supervisor. Studies of barn owls have shown that animals raised with one ear plugged make systematic errors in auditory localization when the earplug is removed [10]. These errors are soon corrected for in the case of those birds with normal vision. However, birds deprived of vision by blindfolding never learn to correct their constant auditory errors. Even more interestingly, birds fitted with prismatic lenses immediately after removal of the stopper adjust their auditory localization

to match the errors induced by the prisms[1]. Here then is direct evidence of the role vision plays in acoustic localization learning, at least in the case of owls. It is probably not unreasonable to assume a similar learning process in humans.

4 A learning paradox

An interesting paradox arises if we accept the basic premise that accurate acoustic localization is learnt by seeing the location of an acoustic visual object and then recording the mapping between this perceived visual location and the corresponding acoustic sensory inputs. The paradox is this; the auditory visual spatial mapping can be learnt using only those acoustic visual inputs which arise from spatially coincident sources. However to decide if this is the case accurate acoustic localization must have already been learnt!

One possible solution to this paradox requires the assumption that temporally correlated acoustic and visual signals are spatially coincident, irrespective of the reality. An existence proof for such an assumption is the "ventriloquism" effect. That is, when viewing a dummy which has its mouth movement synchronized with a ventriloquist's voice, it is common to perceive the voice as coming from the direction of the dummy's mouth, not the ventriloquists [14]. A similar effect occurs when watching television, i.e. speech sounds are usually perceived as coming from the actor's visual images on the screen, not the speakers at the side of the television. The most important variable for achieving such an effect is synchronized movement between the mouth and sounds. A delay of even 0.2 seconds between mouth movements and speech sounds leads to a large decrease in the ventriloquism effect [5]. That the effect breaks down at separation angles greater than about 30 degrees may well be due to the resolving power of any initial crude localization capability. It seems reasonable to speculate that the ventriloquist effect is simply a side effect of the temporal/spatial correlation assumption required for learning a common auditory visual spatial mapping.

[1] As long as the visual and auditory errors were within the same quadrant of directions.

5 A learning scenario for the head/eye/ear system

Here at ATR we are in the process of constructing a head/eye/ear system capable of autonomously learning its own common auditory visual spatial mapping. Very briefly, the system comprises of two color cameras (eye retina) mounted upon 2 degree of freedom servo motors (eye muscles). Servo motors being chosen in favor of stepper motors to allow rapid human like eye motion.

Acoustic input is provided by two miniature omni-direction microphones mounted inside Bruel and Kjaer ear pinna simulators (see figure 1). Just as in a biological system the spacing between the ears and the relative delays introduced by the acoustic system are unknown. Acoustic visual stimuli are provided by a computer controlled speaker/light array.

Based on our interpretation of the psychophysiological evidence and practical engineering constraints the system employs the following simple learning algorithm. First visually reactive saccades are learnt. For those unfamiliar with the terminology, a visually reactive saccade refers to the ability to center a visual object on the retina based solely on the visual error signal [4]. In the simplest example, a light hits the retina where its position is encoded in retinal (pixel) coordinates. The visual error signal is simply the offset from the center of the retina. The eye motors move in response to this visual stimulus and a new error signal recorded. Gradually the system learns to move its eyes to center the light in the retina. This learning is completely self contained within the visual motor system. With this ability the visual motor system can thus provide the eye motor coordinates of any given visual object by performing a visually reactive saccade to the object and reading off the resulting motor coordinates.

Once the ability to perform visually reactive saccades has been acquired the system is presented with a series of temporally and spatially coincident acoustic visual stimuli. Currently such stimuli are produced by playing a sound from a speaker upon which a light is mounted. The eyes saccade to center the light, and the eye motor coordinates are recorded. Simultaneously the auditory system records the corresponding acoustic stimuli and a mapping is built up between the acoustic stimuli and eye motor coordinates. In the current system the acoustic stimuli comprise of the left and right ear power spectra and the mapping is represented by a Gaussian probability

distribution of acoustic stimuli for each motor position.

Importantly the learning process is entirely autonomous; the system simply responds to the visual, or acoustic visual stimuli detected by its sensors. Obviously it is necessary for the majority of the acoustic visual stimuli to be spatially coincident in order to learn the correct spatial mapping. We postulate however that this is simply mirroring the real world. If this were not the case then humans would not have evolved to make the temporal spatial correlation assumption.

Once the above auditory visual spatial mapping has been learnt the system is capable of moving its eyes to visually center a sound source, even in the absence of any associated visual stimulus. In addition sounds can be used to attract the attention of the eyes to acoustic visual objects outside their visual field.

6 Future

Having learnt a common auditory visual perceptual space, the system is potentially able to exploit the subsequent integration of acoustic and visual information arriving at its sensors [18]. We are currently working on developing such algorithms.

References

[1] E. Aronson and S. Rosenbloom. Space perception in early infancy: perception within a common auditory-visual space. *Science*, 172:1161–1163, 1971.

[2] N. Erber. Auditory-visual perception of speech. *Journal of speech and hearing disorders*, XL:481–492, 1975.

[3] P. Ewart. The effect of inverted retinal stimulation on spatially coordinated behaviour. *Genet. Psychol. Monog.*, 1(3):177–193, 1930.

[4] S. Grossberg and M. Kuperstein. *Neural dynamics of adaptive sensory-motor control.* Pergamon Press, 1989.

[5] C. E. Jack and W. R. Thurlow. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills*, 37(967-979), 1973.

[6] B. Jones and B. Kabanoff. Eye movements in auditory space perception. *Perception and Psychophysics*, 17(3):241–245, 1975.

[7] E. Knudsen. Early auditory experience aligns the auditory map of space in the optic tectum of the barn owl. *Science*, 222:939–942, November 1983.

[8] E. Knudsen. The role of auditory experience in the development and maintenance of sound localization. *TINS*, pages 326–330, September 1984.

[9] E. Knudsen. Experience alters the spatial tuning of auditory units in the optic tectum during a sensitive period in the barn owl. *The Journal of Neuroscience*, 5(11):3094–3109, November 1985.

[10] E. Knudsen and P. Knudsen. Vision guides the adjustment of auditory localization in young barn owls. *Science*, 230:545–548, November 1985.

[11] J. Lackner. Visual rearrangement affects auditory localization. *Neuropsychologia*, 11:29–32, 1973.

[12] D. Muir and J. Field. Newborn infants orient to sounds. *Child development*, 50:431–436, 1979.

[13] E. Petajan. Experiments in automatic visual speech recognition. In *Proceedings of the 7th Symposium of hte Federation of Acoustical Societies of Europe(FASE)*. Institue of Acoustics, Edinburgh, 1988.

[14] M. Radeau and P. Bertelson. The after-effects of ventriloquism. *Quarterly Journal of Experimental Psychology*, 26:63–71, 1974.

[15] K. Sekiyama and Y. Tohkura. McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am.*, 90(4):1797–1805, 1991.

[16] G. Stratton. Vision without inversion of the retinal image. *Psychol. Rev.*, 14:341–389 and 463–481, 1897.

[17] A. Moiseff T. Takahashi and M. Konishi. Time and intensity cues are processed independently in the auditory system of the owl. *The Journal of Neuroscience*, 4(7):1781–1786, July 1984.

[18] K. Takahashi and H. Yamasaki. Real-time sensor fusion system for multiple microphones and video camera. In *Proceedings of the Second International Symposium on Measurement and Control in Robotics*, pages 249–256, 1992.

[19] J. Travis. Building a baby brain in a robot. *Science*, 264:1080–1082, 1994.

[20] H. Wallach. The role of head movements and vestibular cues in sound localization. *Experimental Psychology*, 27(4):339–368, October 1940.

[21] P. Young. Auditory localization with acoustical transposition of the ears. *J. Exp. Psychol.*, 11:399–429, 1928.

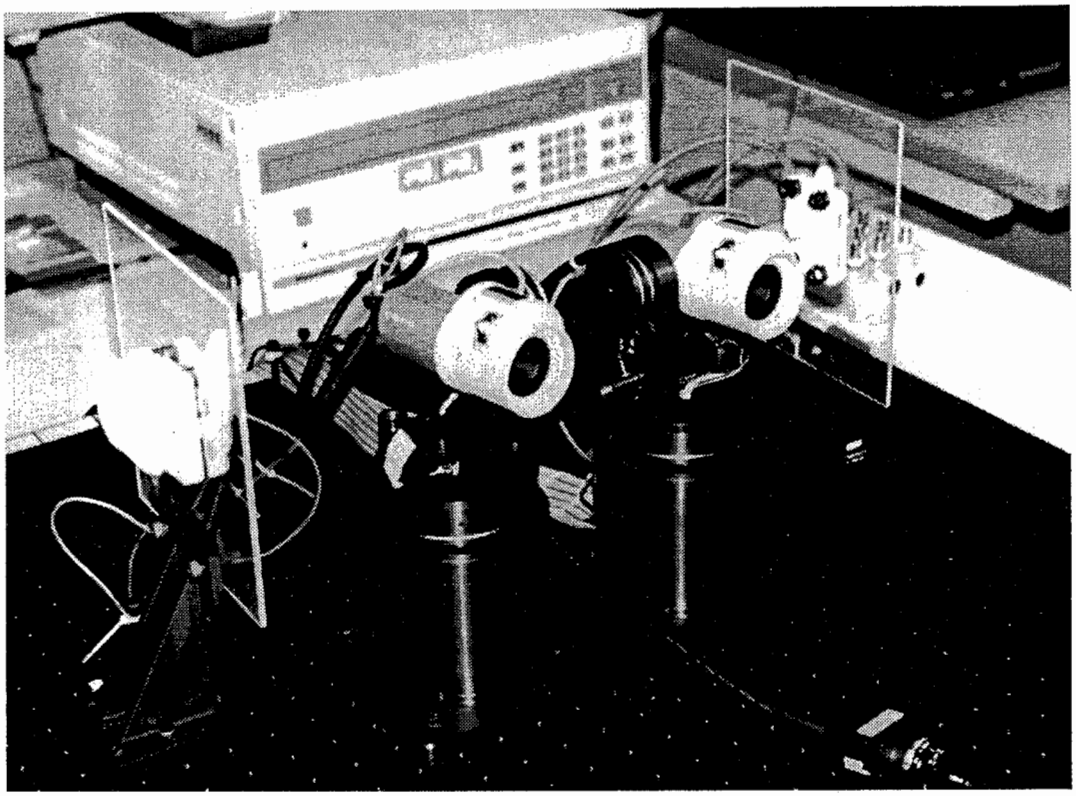

Figure 1: The ATR head/eye/ear system