Scene Analysis by Clues from the Acoustic Signals

Hiroaki Kudo†, Jinji Chen††, and Noboru Ohnishi†

† Graduate School of Information Science, Nagoya University
†† Graduate School of Engineering, Nagoya University

kudo@is.nagoya-u.ac.jp

A several-month-old infant can relate speech sound (audio information: temporal power spectrum structure) to mouth movement (visual information: temporal brightness change in a scene).

Thus, it is favorable for an artificial intelligent system to correspond and integrate multi-modal sensor information in order to realize the function of sensor fusion and the automatic acquisition of knowledge without being taught (including entire object movement and sound location change), like humans.

The purpose of this presentation is to relate multiple audio-visual events observed by a camera and a microphone according to general laws without object-specific knowledge. Namely, we show how to obtain the knowledge of audio-visual information of movement automatically, and to understand the environment through an observation. As corresponding cues, we use Gestalt's grouping laws: simultaneity of sound onsets and direction changes in movement, and similarity of repetition between sound and movement.

We have designed a system that consists of three parts: 1) audio information processing, 2) visual information processing, and 3) correspondence-determination processing.

In the audio information processing, we obtain time series of sound onsets for each separated sound source. In the visual information processing, we obtain the STI (Space-Time Invariant) time series of STI sequence for each object movement. Based on the correlation coefficient between auditory and visual time series, the component of frequency at a sound's onset is related to the STI sequence of movement.

We experimented in the real environment and obtained satisfactory results that proved the effectiveness of the proposed method. We show the results in Figure 1. The upper-left figure shows a frame of a real scene, and the lower-left figure shows the mixed sound signal caused by the two sound sources (a metronome and a left leg). The result is that one sound source with the power spectrum shown in the upper-center figure is related to the STI sequence, which is shown in the lower-center figure. Another sound power spectrum (the upper-right figure) is related to the STI sequence shown in the lower-right figure.

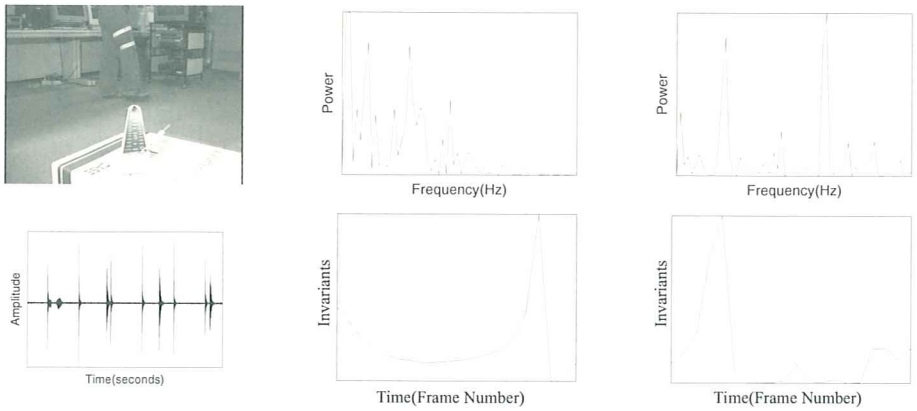


Figure 1: The correspondence of two movements and two sounds (a metronome and a left leg)

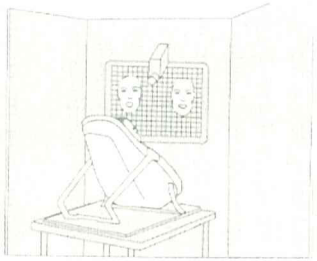
Scene Analysis by Clues from the Acoustic Signals

Hiroaki Kudo †　　Jinji Chen † †
Noboru Ohnishi †

† Graduate School of Information Science, Nagoya University
† † Graduate School of Engineering, Nagoya University

Background

P.K.Kuhl & A.N.Meltzoff:
Science,Vol.218,pp.1138-1140 (1982.12)
The Bimodal Perception of Speech in Infancy



Infant(18~20 week-old)
- Audio information (spectrum of speech sound)

⇕ relating

Visual information (mouth movement)
- Imitating speech

Realizing the function of sensor fusion and acquisition of knowledge

Related research

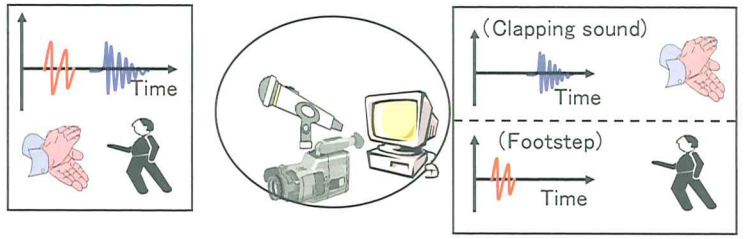
- Takahasi et al., Real-Time Sensor Fusion System for multiple Microphones and Video Camera(1992)
  One movement and one sound
- Mukai et al., Grouping Corresponding parts in Vision and Audition Using Perceptual Grouping among Different Sensations(1996)
  One movement and one sound
- Hayakawa et al., Finding correspondence between vision and audition based on physical law(1999)

Multiple movements and one sound

- Chen et al.,Finding the Correspondence of Audio-Visual Events Caused by Multiple Movements(2001)
  Multiple movements and multiple sound

Objective

Relating the audio-visual events caused by more than one movement.



The correspondence between more than one movement and more than one sound.

Clue of audio-visual correspondence

	Visual information	Auditory information	Gestalt's grouping factor
1	Category, Material	Category, Sound tone	None
2	Size, Distance	Loudness	Common fate
3	Location	Direction	Similarity
4	Change in movement	Sound onset	Common
5	Repetition of movement	Repetition of sound	Similarity

Correspondence principle

(one camera and one microphone)

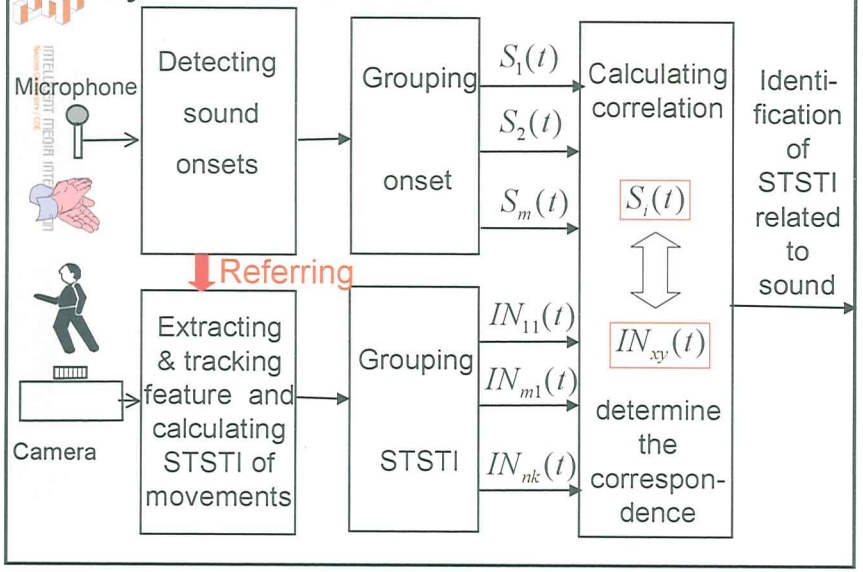
- Simultaneous occurrence of the sound and the change in movements for the same event
- Similarity of repetition between sound and movement

The occurrence time of the sound.

Correlations

The occurrence time of corresponding movement

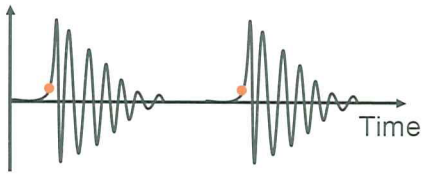
System overview



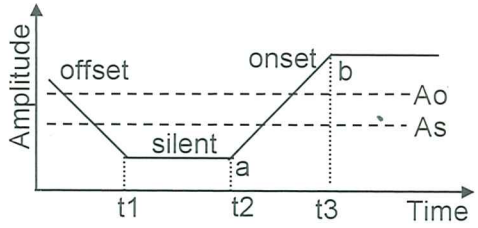
Sound information processing

Input signal

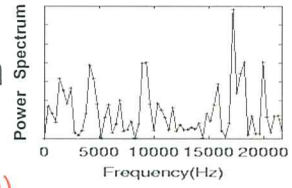
(Monaural sound signal)



Detecting onsets

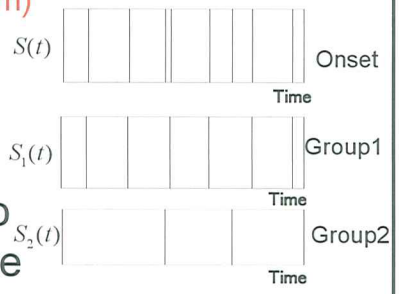


Calculating spectrum of the onset



(Using correlation of spectrum)

Separating mixed sound signal



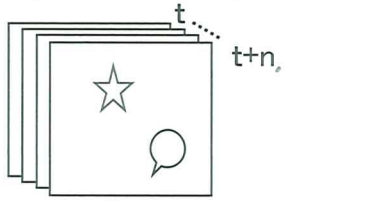
Fitting sampling rate to the image sampling rate

Getting sound onset time series $S_i(t)$

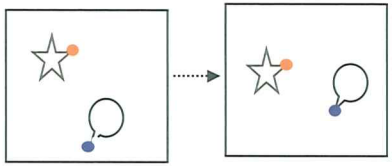
$$S_i(t)=\begin{cases}1 & \text{if } t \text{ is onset}\\ 0 & \text{otherwise}\end{cases}$$

Visual information processing

Inputting image series
(Around sound onset)



Extracting and tracking of feature points

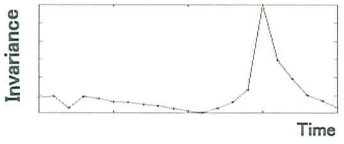


Calculating STI
(space-time invariant)

$$I(m_i,m_l)=\frac{Volume(m_i,m_j,m_k,m_n)}{Volume(m_i,m_j,m_k,m_l)}$$

(m_t, feature point)

Making STSTI
(short-term space-time invariant)



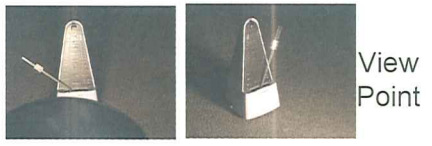
Separating STSTI
(similarity)

Getting the series $IN_k(t)$ of STSTI
(k=1, …,n)

$$IN_i(t)=\begin{cases}1 & \text{if } t \text{ has a STSTI}\\ 0 & \text{otherwise}\end{cases}$$

Space-time invariant

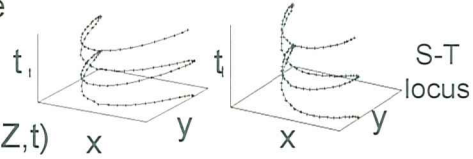
•Point M(X,Y,Z) in space



View Point

Affine camera

Point m(x,y) on image

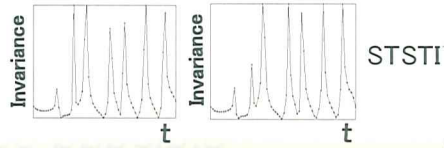


S-T locus

• Space-time locus (X,Y,Z,t) of point M

Movement on plane

Affine projection



STSTI

Space-time locus (x,y,t) of point m

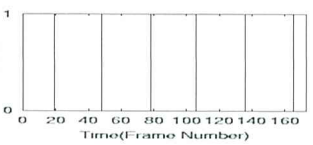
Correspondence determination

$S_i(t)$:Time series of sound onset of the i-th sound ($i=1,2,\ldots,m$)

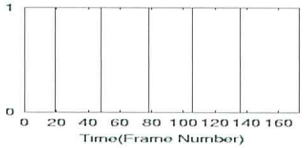
$INxy(t)$:Time series of STSTI ($x=1,2,\ldots,m;\ y=1,2,\ldots,n$)

$$\gamma_{ik} = \frac{Cov(Si, INxy)}{\sqrt{Var(Si)\cdot Var(INxy)}} > TH$$

γ_{ik}: Correlation coefficient $Var(\bullet)$:Variance $Cov(\bullet)$:Covariance



Time series of sound

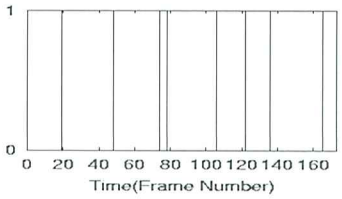


Time series of STSTI

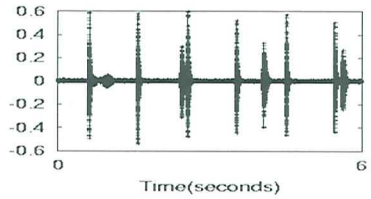
Experiment

- One camera and one microphone
- Sampling frequency

Sound 44.1kHz
Image 30 frames/sec.

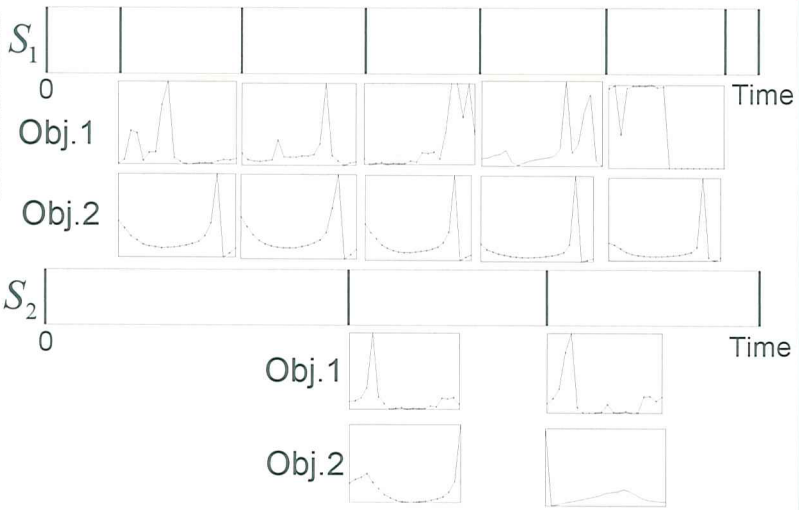


Extracted onsets

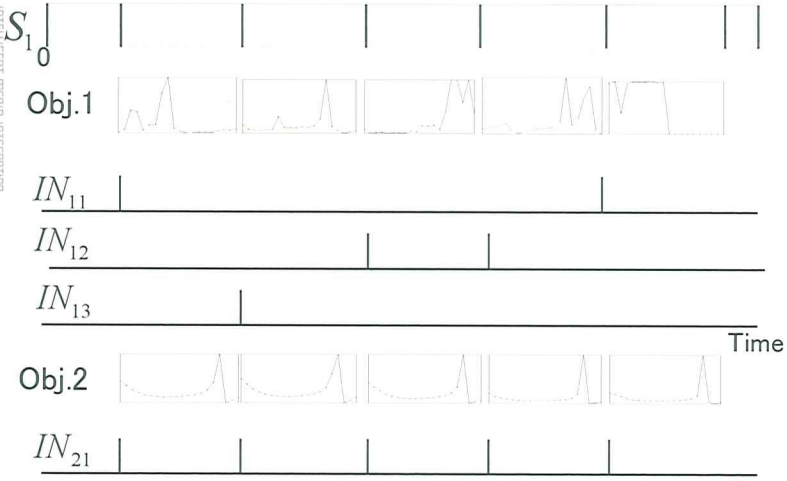


Mixed signal

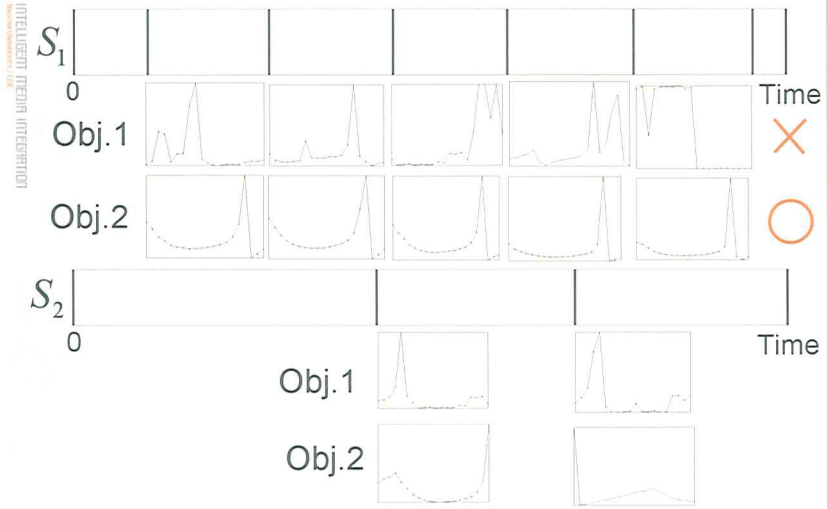
Result(time series of onset and STSTI)



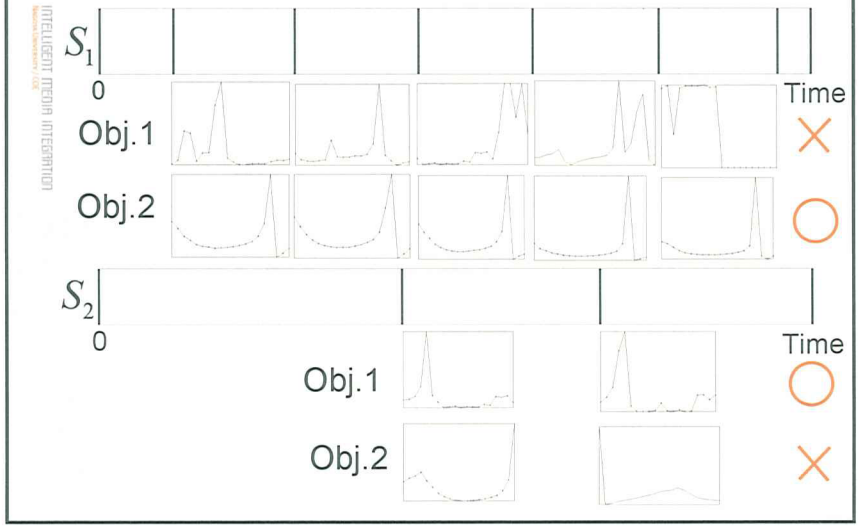
Result(S and IN)



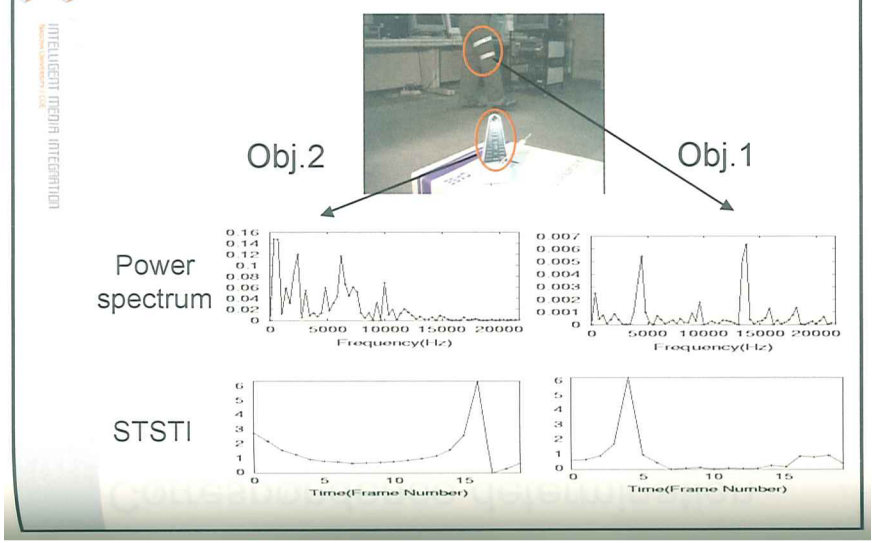
Result (time series of onset and STSTI)



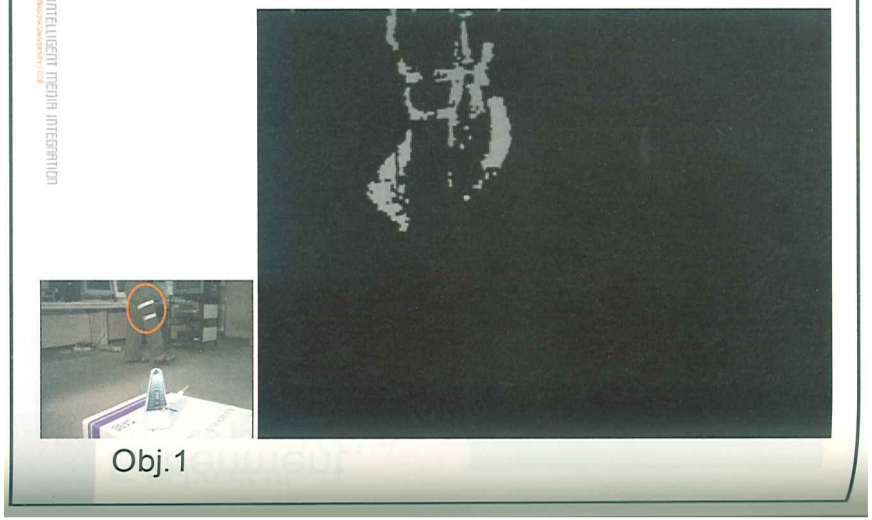
Result (time series of onset and STSTI)



Results (power spectrum and STSTI)

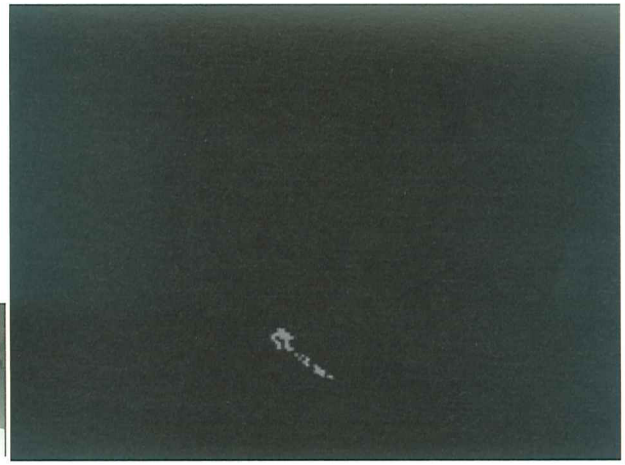


Results (Movement 1)



Results（Movement 2）

Obj.2



Conclusion

- We proposed the method for correspondence between more than one movement and more than one sound. Even the sound location changes and entire object moves.
- We showed the system can make audio-visual correspondence in the real environment.

Further Works

- Applying our method in real time using a robot