博　士　論　文

# Synchrony-based Audiovisual Analysis
# 同期性に基づく音と映像の統合解析

東京大学大学院
情報理工学系研究科
電子情報学専攻

48-67413　　劉　玉宇

指導教員　　佐藤　洋一　准教授

平成 21 年 9 月

# Synchrony-based Audiovisual Analysis

by

Yuyu Liu

Bachelor of Science
*Beijing University of Posts & Telecommunications*
2000

Master of Science
*Tsinghua University*
2003

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Information and Communications Engineering

in the

GRADUATE DIVISION

of

THE UNIVERSITY OF TOKYO

September 2009

Synchrony-based Audiovisual Analysis



by

Yuyu Liu

## Abstract

This thesis presents a computational framework to jointly analyze auditory and visual information. The integration of audiovisual information is realized based on synchrony evaluation, which is motivated by the neuroscience discovery, that synchrony is a key for human beings to perceive across the senses of different modalities. The works in this thesis focus on answering two questions: how to perform and where to apply this audiovisual analysis with synchrony evaluation.

To answer the first question, we develop novel effective methods to analyze the audiovisual correlation, and perform a classification and an experimental comparison of the existing techniques, including the ones we developed. Since this is the first work that classifies and experimentally compares the methods of this field, it supplies a basis for designing algorithms to computationally analyze the audiovisual correlation.

To answer the second question, we apply audiovisual correlation analysis to solve three different problems.

The first problem is the detection of a speaker's face region in a video, whose previous solutions either require special devices like microphone array or supply only highly fragmental results. Assuming that speaker is stationary within an analysis time window, we introduce a novel method to analyze the audiovisual correlation for speaker using newly introduced audiovisual differential feature and quadratic mutual information, and integrate the result of this correlation analysis into graph cut-based image segmentation to compute the speaker face region. This method not only achieves the smoothness of the detected face region, but also is robust against the change of background, view, and scale.

The second problem is the localization of sound source. General sound sources are diverse in types and usually non-stationary while emitting sounds. To solve this problem, we develop an audiovisual correlation maximization framework to trace the sound source movement, and introduce audiovisual inconsistency feature to extract audiovisual events

for all kinds of sound sources. We also propose an incremental computation of mutual information to significantly speed up the computation. This method can successfully localize different moving sound sources in the experiments.

The third problem is the recovery of drifted audio-to-video synchronization, which used to require both special device and dedicated human effort. Considering that the correlation reaches the maximum only when audio is synchronized with video, we develop an automatic recovery method by analyzing the audiovisual correlation for a given speaker in the video clip. The recovery demonstrates high accuracy for both simulation and real data.

While the theoretical justification and experimental justification are performed independently, this thesis taken as a whole lays a necessary groundwork for jointly analyzing audiovisual information based on synchrony evaluation.

To Danhua and Xiangbao

# Acknowledgements

I would like to extend my deepest thanks to my advisor, Yoichi Sato, for his discussions, guidance and encouragement over the past three years. He directed my research, proofread all my submissions, and gave me, a foreigner in Japan, generous help in many aspects. His comments and suggestions were invaluable to me.

I am indebted to many of my lab members who were often a source of inspiration to my research and life during my studies. Special thanks to Imari Sato and Takahiro Okabe, for their advices to my research. I thank Kris Kitani for giving me advices on campus life and correcting my English errors. I thank Daisuke Sugimura for proofreading my paper and helping improve my Japanese. I would like to thank Yusuke Sugano for our free discussions on various topics and getting me to think differently. I thank Michihiro Kobayashi for answering my questions about Matlab and discussing optical flow with me. I also thank Shiro Kumano, Fei Du, Gabriel Pablo Nava, Lulu Chen, and Chung-Lin Wen, for spending the time to help me work through different ideas.

I would like to express my thanks to my ex-colleagues in Sony research for their kind help even after I have quitted the company. I thank Takayuki Yoshigahara and Weiguo Wu for their advices and help to my research. I would also like to thank Keisuke Yamaoka, Yoshiaki Iwai, and Akira Nakamura for their enthusiastic encouragement.

I thank my parents for their encouragement. Special thanks to my wife Danhua. She gives me quiet and consistent support, and hides her needs to let me focus on my work. I have not even noticed many of these needs until a great difficulty fell to her. I should never wait for a convenient time to take care of the people who love me.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

We human beings can naturally fuse the auditory and visual senses. What we hear is unconsciously associated to what we see, and vice versa. With this astonishing ability, we can efficiently localize an approaching predator, can feel the beauty of a rhythmic dance with music, and can communicate with each other with both speech and nonverbal signals.

However, such a natural ability is still difficult for a machine to realize, regardless of the progresses independently made in audio processing and computer vision. Recently, motivated by the mechanism of how human beings jointly perceive audiovisual information, researchers are getting believed that synchrony is the key to understand this audiovisual integration. The remained problems following this idea are how to computationally analyze this synchrony and where to apply the audiovisual analysis by using a machine. This thesis focuses on finding the answers for these two questions.

## 1.1 Motivation

Techniques have been developed to utilize both auditory and visual information to enhance the processing that used to utilize only one of them. An example of this idea is to improve speech recognition with a visual observation of the lips (Petajan, 1984).

Other examples include audiovisual person authentication (Poh and Korczak, 2001), audiovisual human tracking (Li et al., 2004), audiovisual face detection (O'Donovan et al., 2007), and so on.

These approaches took advantage of audiovisual information by directly using the classical processing framework of either audio or visual information and incorporating the information of the other modality as an additional feature. An advantage of these approaches is that they can adopt the framework that has been researched thoroughly in audio or visual processing. On the other hand, a critical defect lies in that it is usually difficult to extract this feature from the signals of another modality.

The difficulty to extract this feature is resulted from the substantial difference between audio and visual signals. Audio signal captured by a microphone varies in temporal domain only and is omni-directional, which means that microphone summates audio signals from all directions. While visual signal captured by a camera varies in both spatial and temporal domains. Such difference is demonstrated in Figure 1.1. Consequently, if the feature is to be extracted from visual signal, the spatial location of the sound source is required as a prerequisite. This is why that audiovisual speech recognition demands the camera to focus on the mouth of a speaker (Petajan, 1984; Rivet et al., 2006). Conversely, if the feature is to be extracted from audio signal, an array of microphones is usually required because only one microphone cannot supply spatial information (Li et al., 2004; O'Donovan et al., 2007). The use of such special devices not only severely limits the applicability of the techniques based on this approach, but also demands a complicated process to calibrate the world coordinate of the microphone array to that of the camera.

This problem remains unresolved until Hershey and Movellan introduced the discoveries of neuroscience into this field (Hershey and Movellan, 1999). In the second half of the twentieth century, a series of neuroscience discoveries have revealed a fact that human beings sense the auditory and visual information in a fairly interactive way. For instances, we used to think that speech is perceived by hearing only, although vision supplies auxiliary information. This opinion was proven to be wrong by the McGurk effect discovered in 1976 (McGurk and MacDonald, 1976), which suggests that speech

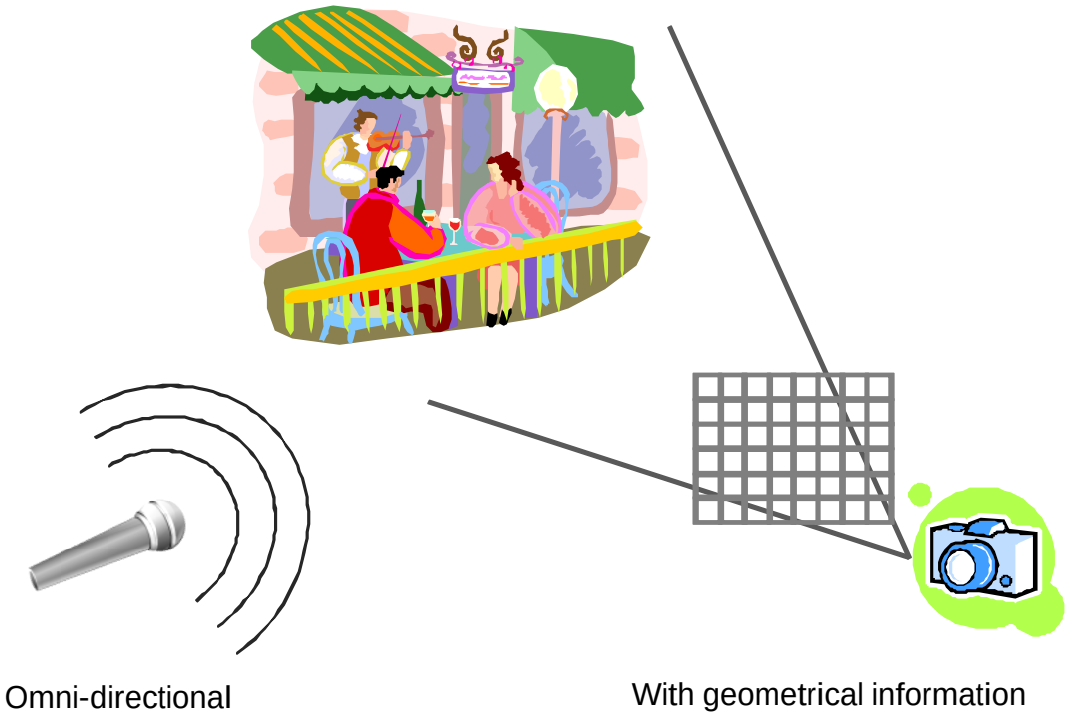



Figure 1.1. Camera and microphone.

perception involves at least both hearing and vision. Similarly, the discovery of illusory mis-location (Bertelson et al., 1994; Driver, 1996) revealed that the position of a sound source can be mis-localized to its visible place, although the sound in fact comes out from a different position (e.g., consider a show of ventriloquism). These phenomena are robust in a wide variety of conditions, and have been found to be strongly dependent on the degree of "synchrony" between the audio and visual signals (Bertelson et al., 1994; Driver, 1996; McGurk and MacDonald, 1976). Inspired from this fact, Hershey and Movellan developed a method to localize a sound source by evaluating the degree of synchrony (Hershey and Movellan, 1999).

The contribution of the work in (Hershey and Movellan, 1999) lies in not only a new method to localize a sound source, but also a novel framework to process the audiovisual information, where problems are solved by optimizing an audiovisual objective function rather than the one in only one modality. This objective function is related to the eval-

uation of synchrony of the events between the audio and visual signals, which has its root in the neuroscience discoveries (Bertelson et al., 1994; Driver, 1996; McGurk and MacDonald, 1976).

Since this new framework optimizes an objective function over both the audio and visual signals, it can probably remove the above special requirements, such as the necessity of a microphone array. That is to say, with the novel framework, we can process the audiovisual data that are captured by the set of one camera and one microphone. This relaxed condition can be satisfied by the data taken by any off-the-shelf video camera, which implies a wide application foreground of this new framework.

Our research interests focus on this new framework by trying to answer two questions: how to perform and where to apply this audiovisual analysis with synchrony evaluation. It is my hope that this work will be one of endeavors to gain interests on the synchrony-based audiovisual analysis.

## 1.2 Overview

We begin this thesis by reviewing related works in this field in Chapter 2. Following this review, we also classify and make an experimental comparison on the methods to analyze the audiovisual correlation. The results of this comparison answer the question on how to analyze the audiovisual correlation by supplying objective evidence. Based on these results, we design our methods to analyze the audiovisual correlation.

Chapter 3, Chapter 4, and Chapter 5 answer the question on where to apply this audiovisual correlation analysis. Three different problems are solved in the three chapters by analyzing the audiovisual correlation. All of them used to require special devices or dedicated human effort. The effective solution to these problems demonstrates the usability of this audiovisual correlation analysis by synchrony evaluation.

Face region segmentation of a stationary speaker in a video is introduced in Chapter 3. For us human beings, speaker is a most important class of sound sources. The necessity to localize a speaker happens frequently in our daily communications. However,

previous solutions either require special devices like microphone array or supply only highly fragmental results. Assuming that speaker is stationary within an analysis time window, we introduce a novel method to find this region robustly against the changes of view, scale, and background. The main thrust of our technique is to integrate audiovisual correlation analysis into a video segmentation framework. We analyze the audiovisual correlation locally by computing quadratic mutual information between our audiovisual features. The computation of quadratic mutual information is based on the probability density functions estimated by kernel density estimation with adaptive kernel bandwidth. The results of this audiovisual correlation analysis are incorporated into graph cut-based video segmentation to resolve a globally optimum extraction of the speaker's face region. The setting of any heuristic threshold in this segmentation is avoided by learning the correlation distributions of speaker and background by expectation maximization. Experimental results demonstrate that our method can detect the speaker's face region accurately and robustly for different views, scales, and backgrounds.

The localization of a non-stationary sound source is introduced in Chapter 4. Sound source here indicates visual objects that emit or induce sound, such as a walker, a hand playing a piano, and certainly a speaker. The motion pattern of sound source may be more complex than that of a speaker. Additionally, a sound source is usually not stationary, i.e., it moves its position while emitting sound (for instance, a hand that plays a piano). Consequently the localization of a non-stationary sound source is much more difficult than that of a stationary speaker. We develop a method to achieve this goal. This method can also be adopted to localized a stationary speaker, except that, being general, this method requires more complex computation than the one introduced in Chapter 3. The localization problem is formulated as independently finding the optimal visual trajectories that best represent the movement of the sound source over the pixels in a spatio-temporal volume. Using a beam search, we search these optimal visual trajectories by maximizing the correlation between the newly introduced audiovisual features of inconsistency. An incremental correlation evaluation with mutual information is developed here, which significantly reduces the computational cost. We would like to mention that this incrementally computation is in fact general and can be applied to other applications to speed

up the computation of mutual information at stages. The correlations computed along the optimal trajectories are finally incorporated into a segmentation technique to localize a sound source region in the beginning visual frame. The experimental results demonstrated that our method could localize different types of non-stationary sound sources.

A new way to recover audio-to-video synchronization by using audiovisual correlation analysis is proposed Chapter 5. This recovery used to require a human to elaborately adjust audiovisual data by using a special device. Based on audiovisual correlation analysis, we develop a method of recovering drifted AV-sync in a video clip with only minor human interactions. Users just need to specify the time window for a stationary speaker. We search the optimum drift within this time window that maximizes the average audiovisual correlation inside the speaker region by shifting audio and computing the correlation for different drift hypotheses, and then recover the state of synchronization based on the refined optimum drift. The audiovisual correlation is analyzed by quadratic mutual information with kernel density estimation, which is not only robust against audiovisual changes in scale, but also independent of the language. The experimental results demonstrate that our method could effectively recover audio-to-video synchronization.

Finally, the conclusions and discussions of this thesis are presented in Chapter 6.

# Chapter 2

# Preliminaries

## 2.1 Related works

Motivated by the neuroscience discovery that synchrony is a key for human beings to integrate audiovisual information (Bertelson et al., 1994; Driver, 1996; McGurk and MacDonald, 1976), Hershey and Movellan first developed a method to computationally analyze the audiovisual correlation by evaluating the degree of synchrony (Hershey and Movellan, 1999). This evaluation was performed by computing mutual information (Shannon, 1951) between the temporal samples of pixel intensity and audio energy. Generally, the computation of mutual information requires the estimation of probability density function. To avoid this estimation, they assumed that pixel intensity and audio energy obey a joint normal distribution. With this, mutual information can be computed directly from the Pearson correlation coefficient as

$$MI(a;v) = -\frac{1}{2}\log(1-\rho^2(a,v)), \tag{2.1}$$

where $a$ and $v$ represent the audio energy and pixel intensity, respectively. $\rho(a,v)$ is the Pearson correlation coefficient. Given the temporal samples of $a$ and $v$ as $a_i$ and $v_i$, $i = 1,\cdots,N$, Pearson correlation coefficient $\rho(a,v)$ is given by (Rodgers and Nicewander, 1988)

$$\rho(a,v) = \frac{\sum_i (a_i-\bar{a})(v_i-\bar{v})}{\sqrt{\sum_i (a_i-\bar{a})^2 \sum_i (v_i-\bar{v})^2}}, \tag{2.2}$$

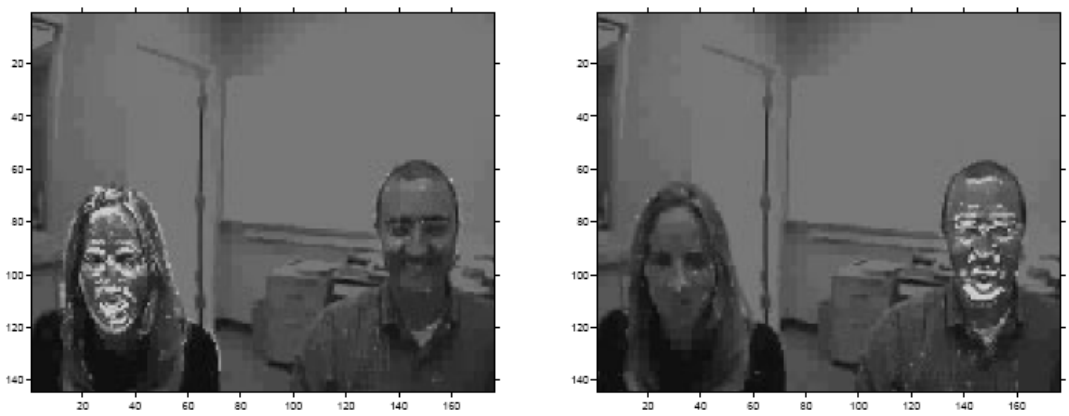


Figure 2.1. The audiovisual correlation analyzed by (Hershey and Movellan, 1999).

where $\bar{a} = \frac{1}{N}\sum_i a_i$, and $\bar{v} = \frac{1}{N}\sum_i v_i$. They computed this audiovisual correlation by Equation (2.1) for all the pixels. The result was regarded as the degree how much that pixel belonged to sound source. The analyzed audiovisual correlation of their experimental results is demonstrated in Figure 2.1.

Following this seminal work, many methods were developed to analyze the audiovisual correlation more effectively and apply this analysis to solve problems. We briefly introduce several representative ones of them below.

In 2003, Smaragdis and Casey (Smaragdis and Casey, 2003) avoided the explicit analysis of the audiovisual correlation by using principal component analysis and independent component analysis to localize the sound source. Pixel intensities of a visual frame were aligned into a vector. The spectrum of the audio signal that corresponds to this visual frame was also aligned into another vector. They then combined these two vectors into a large one and applied principal component analysis to the temporal samples of this large vector such that they could find a linear transform to convert these vectors into low dimensional ones. Finally, they applied independent component analysis to the converted low dimensional vectors. The detected independent components were regarded as different sound sources. By projecting them back to the large vector, they could find the positions of these sound sources in the visual image and their spectral elements. Their detected sound sources are shown in Figure 2.2 (a).

Instead of analyzing the value of audiovisual correlation, Fisher and Darrell (Fisher and Darrell, 2002, 2004) localized speaker by taking the audiovisual correlation as an objective function to maximize. Like (Smaragdis and Casey, 2003), they also aligned pixel intensities and audio spectrum into two vectors. They showed that inequality

$$MI(h_a^T X_a; h_v^T X_v) \leq MI(X_a; X_v) \tag{2.3}$$

holds true for any projection vectors $h_a$ and $h_v$. $X_a$ and $X_v$ are the audio and visual vectors, respectively. Since $h_a^T X_a$ and $h_v^T X_v$ result in two low dimensional vectors, the computation of $MI(h_a^T X_a; h_v^T X_v)$ is much easier than $MI(X_a; X_v)$. The value of mutual information was computed by using the temporal samples of $X_a$ and $X_v$. Note that $h_a$ and $h_v$ were assumed to be same for the temporal samples of $X_a$ and $X_v$. They maximized this $MI(h_a^T X_a; h_v^T X_v)$ with respect to $h_a$ and $h_v$. This maximization of $MI(h_a^T X_a; h_v^T X_v)$ was regarded as the maximization of the lower bound of $MI(X_a; X_v)$. Since $h_v$ was a weighting vector whose dimension was the same as that of $X_v$, the found optimum $h_v$ were regarded as a spatial distribution that indicated the likelihoods how much a pixel belonged to speaker. Their computed likelihoods are shown in Figure 2.2 (b).

Similarly, Kidron and Schechner (Kidron et al., 2005, 2007) also localized sound source by searching a projection vector that can maximize their objective function. At the first step, visual image was converted into wavelet coefficients, and audio energy at the time of each visual frame was computed. Using the temporal samples of the audio and visual features, they initially planned to analyze the audiovisual correlation by canonical correlation analysis. However, they found that, for a short time window, say, several seconds, the correlation analyzed by canonical correlation analysis always reached its maximum. Hence they instead took it as a constraint that the correlation is maximized, and optimized the norm of the projection vector. By comparing $L0$-, $L1$-, and $L2$-norms, they claimed that $L1$-norm is the best since the found optimum projection vector can be spatially sparse. That is to say, the number of the detected positions belonging to sound source can be as small as possible. Based on these considerations, their objective

function was designed as

$$\begin{cases} \arg\min_{\mathbf{w}} ||\mathbf{w}||_1 \\ \mathbf{V}\mathbf{w} = \mathbf{A} \end{cases}, \tag{2.4}$$

where $\mathbf{V} = (\mathbf{V}_1, \cdots, \mathbf{V}_N)^T$ was the packed visual features of $N$ frames. $\mathbf{A} = (a_1, \cdots, a_N)^T$ was the vector of audio energies of $N$ frames. $\mathbf{w}$ was the projection vector. The computed optimum $\mathbf{w}$ was projected back to the image by the wavelet transformation to give out the positions of sound source. Their localization results are shown in Figure 2.2 (c).

All the approaches introduced above based the visual feature on the change of the appearance of photographed visual objects. However, in (Monaci et al., 2005; Monaci and Vandergheynst, 2006), Monaci *et al.* claimed that the movement of photographed sound source conveys better correlation with auditory information than the change of their visual appearance. Consequently, they first detected local visual objects in the first video frame and then tracked the movements of all these visual objects, including scale change, translation, and rotation. Both the detection of visual objects and their tracking were computed by using matching pursuit algorithm, which decomposes an image into two-dimensional anisotropic atoms (Vandergheynst and Frossard, 2001). The audiovisual correlation was analyzed independently for all these visual objects by computing Pearson correlation coefficient (Rodgers and Nicewander, 1988) between the change of the parameters of visual movements and the change of audio energy. The visual objects whose audiovisual correlations were higher than a pre-defined threshold were regarded as sound source, which are shown in Figure 2.2 (d).

Based on the localization result of sound source, Casanovas furthermore tried to separate the sound from different sources (Casanovas, 2006). He first used the same matching pursuit algorithm to decompose audio signal into Gabor atoms. He also binarized the movements of each visual object into whether or not there was a visual event, which was based on the magnitude of the displacement of the visual movement. If an audio atom happened simultaneously with a visual event, they were linked. The audiovisual correlation of this visual object was also incremented. After this process was finished, visual objects whose audiovisual correlations were beyond a pre-defined threshold were

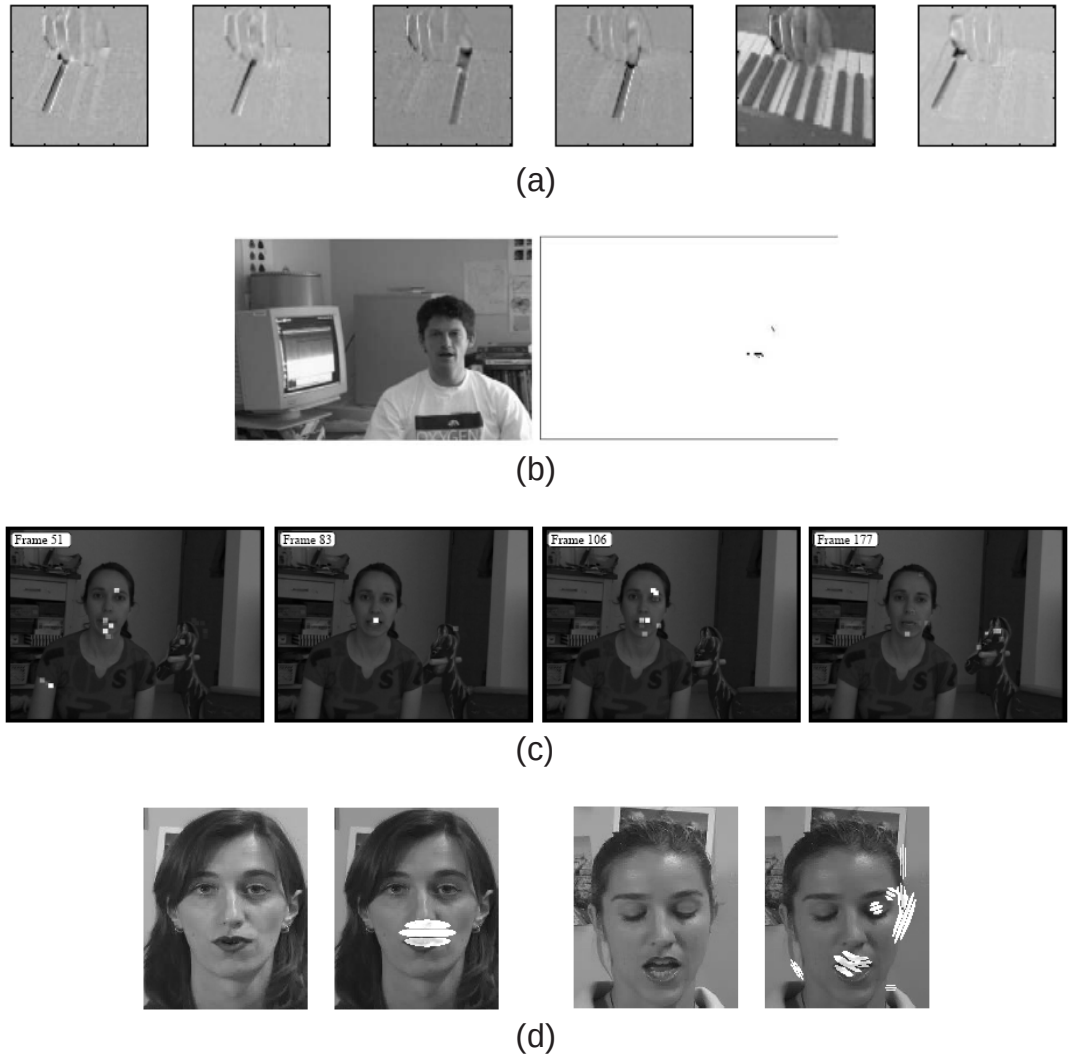



Figure 2.2. Experimental results of localized sound sources. Figures (a), (b), (c), and (d) show the localization results of a sound source of (Smaragdis and Casey, 2003), (Fisher and Darrell, 2004), (Kidron et al., 2007), and (Monaci et al., 2005), respectively.

regarded as sound source and clustered according to their spatial positions. The number of the clusters was taken as the number of sound sources. In the next step, all the audio atoms that were linked to the visual objects that belonged to the same sound source were considered to be the ones resulted by the sound of that sound source. Using inverse matching pursuit algorithm, the sound from that source was reconstructed with these audio atoms. The reconstructed sounds of a man and a woman who uttered simultaneously are shown in Figure 2.3 (a). However, this method required an important assumption that each audio atom corresponded to only one sound source, which was difficult to satisfy in usual cases. That is why he only tried to separate the mixed sounds from a man and a woman. Generally speaking, woman's voice covers high frequencies, while man's voice covers low frequencies. Their mixed sound has small overlaps in the frequency domain. Hence each audio atom can be regarded as belonging to only one sound source.

Barzelay and Schechner (Barzelay and Schechner, 2007) also developed a method to separate sounds from different sources. Compared to the extracted visual objects used in (Casanovas, 2006), they tracked the movements of several featured points, which was extracted by the method in (Shi and Tomasi, 1994). The accelerations of the movements of these featured points were computed and binarized by a pre-defined threshold. The binarized values at different frames formed a vector $\mathbf{v}$. On the other hand, audio signal was transformed into frequency domain. For each frequency, they verified whether or not there was an audio onset at the time of each visual frame. The binary flag of this existence formed a vector $\mathbf{a}$ also. The audiovisual correlation between $\mathbf{v}$ and $\mathbf{a}$ was analyzed by

$$C = 2\mathbf{a}^T\mathbf{v} - \mathbf{1}^T\mathbf{v}, \tag{2.5}$$

where $\mathbf{1}$ was a vector with all its elements to be 1. They then developed a framework to optimize this audiovisual correlation with respect to the number of sound sources and the separation of sounds in an iterative procedure. Note that, differing from (Casanovas, 2006), they did not clustering the featured points, but only selected a representative one for a sound source to perform sound separation. Unfortunately, their method still required the similar assumption as (Casanovas, 2006) that the sounds from different sources cover different frequencies without any overlap. The separation result is shown in Figure 2.3 (b).

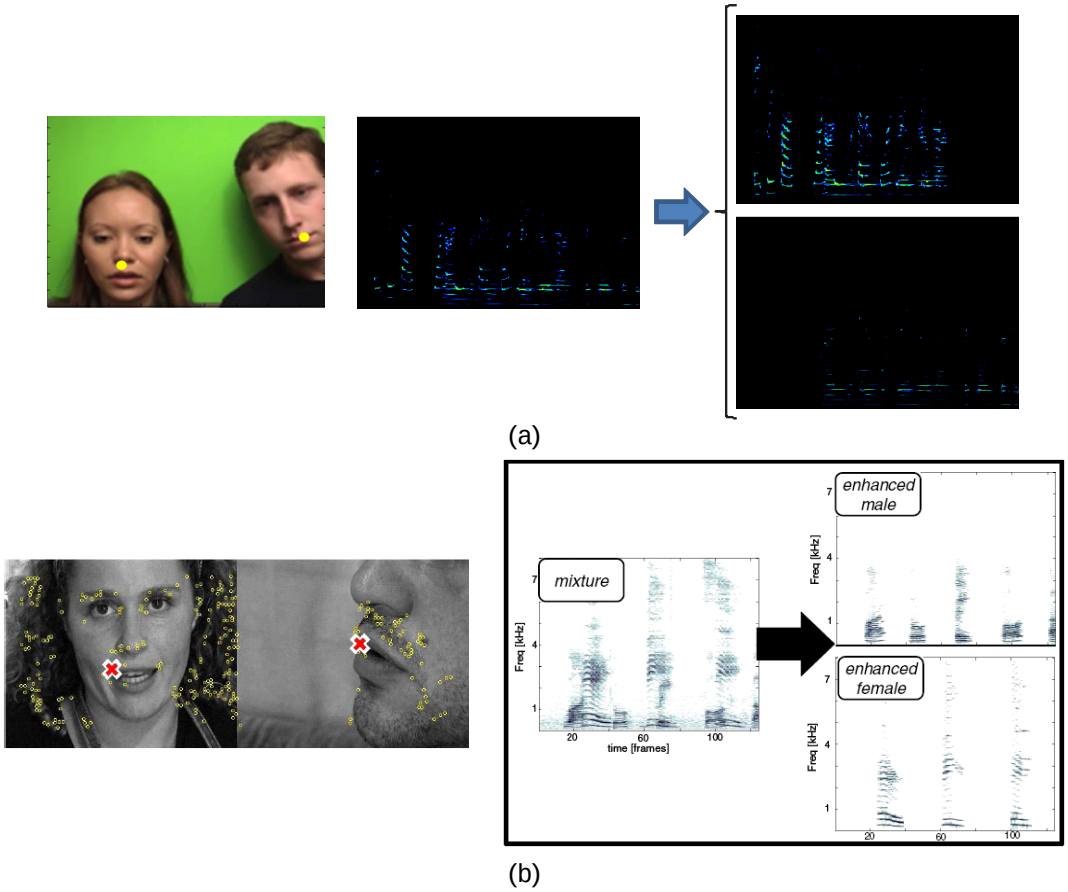



Figure 2.3. Experimental results of sound separation. Figures (a) and (b) show the experimental results of sound separation in (Casanovas, 2006) and (Barzelay and Schechner, 2007), respectively. In both figures, separation results are shown in spectrograms, whose horizontal and vertical axes represent time and frequency, respectively.

## 2.2 Classification of audiovisual correlation analysis

Based on the previous review of the existing techniques, we give a classification on the methods that analyze the audiovisual correlation. To make this classification complete, we include the techniques developed by us, which will be introduced in details in the following chapters. Here we include them by citing our published papers, which is the same as the references to the works of other persons.

This classification is divided into the classification of "feature" and the one of "measure". Generally speaking, feature and measure drive audiovisual correlation analysis. As mentioned before, audiovisual correlation analysis is based on the neuroscience discoveries that synchrony is the key for human beings to jointly perceive audiovisual information. However, it remains unknown between which and how we should evaluate this synchrony. Hence, we have to answer these two questions first in order to be able to computationally analyze the audiovisual correlation. The development of feature gives the author's answer to the first question, and the development of measure gives the answer to the second question.

### 2.2.1 Different features

We classify visual and audio features on two axes. One axis is the physical meaning which a feature represents. Another axis is the order of temporal differential which a feature belongs to. At the axis of physical meaning, visual and audio features are independently classified because they usually have different forms even when they describe a same physical meaning. The reason of this lies in the substantial difference between visual and audio signals. On the contrary, the classification at the axis of differential order is same for both visual and audio features. Although audio signal has a much higher sampling frequency than visual signal, audio samples are usually divided into frames for the sake of feature extraction. The frame duration is set to be the same as the duration of a visual frame. Audio feature is extracted from the samples inside each frame. Therefore

audio feature is aligned with visual feature in temporal domain and can have the same classification as the visual feature at the axis of differential order.

In the sense of physical meaning, visual features can be classified into appearance-related and motion-related. Appearance-related visual features describe the change of visual appearance observed in the camera, such as pixel intensity (Hershey and Movellan, 1999; Smaragdis and Casey, 2003; Fisher and Darrell, 2004) and wavelets (Kidron et al., 2005). While motion-related visual features describe the motion of taken visual objects, which cannot be got directly from visual signal but has to be analyzed by other algorithms, such as optical flow (Liu and Sato, 2008), tracking (Barzelay and Schechner, 2007), and matching pursuit (Monaci et al., 2005). Motion-related visual features include visual translation (Monaci et al., 2005), optical flow (Liu and Sato, 2008), acceleration (Barzelay and Schechner, 2007), and visual inconsistency (Liu and Sato, 2008).

Compared to visual features, all the audio features have the same physical meaning. They all try to describe the temporal change of the magnitude of audio signal. However, according to whether or not this magnitude is computed individually for different frequencies, audio features can be classified into spectrum-related and energy-related. Spectrum-related audio features describe the change of the magnitude with respect to not only time but also frequency. They are usually multi-dimensional vectors for each frame, with each dimension represents the magnitude of energy at different frequencies. Spectrum-related audio features include spectrum itself (Fisher and Darrell, 2004; Smaragdis and Casey, 2003), pursuits (Casanovas, 2006), and audio onset (Barzelay and Schechner, 2007). Energy-related audio features describe the magnitude of audio energy inside a frame. They usually have only one dimension and are computed by summing the energy of audio samples in a frame, although methods have difference in the division of frames and weights for different samples. Energy-related audio features include energy (Hershey and Movellan, 1999; Kidron et al., 2005), differential energy (Liu and Sato, 2008) and audio inconsistency (Liu and Sato, 2008).

In the sense of the differential order, both audio and visual features can be classified into zero-, first-, and second-order differential ones. Zero-order differential features represent their states at current time, which can be extracted with only one-frame audio-

Table 2.1. Feature classification.

| | Visual feature | | Audio feature | |
|---|---|---|---|---|
| | Value | Motion | Spectrum | Energy |
| Zero-order | Pixel intensity, Wavelets | Translation | Spectrum, Pursuit | Energy |
| First-order | - | Optical flow | - | Differential energy |
| Second-order | - | Acceleration, Inconsistency | Onset | Inconsistency |

visual data. Zero-order differential features include visual features like pixel intensity (Hershey and Movellan, 1999) and audio features like spectrum (Smaragdis and Casey, 2003). First-order differential features represent the velocity of the change of their states at current time, whose extraction requires at least two-frame audiovisual data. First-order differential features include visual features like optical flow (Liu and Sato, 2008) and audio features like differential energy (Liu and Sato, 2008). Second-order differential features represent the acceleration of the change of their states at current time, whose extraction requires at least three-frame audiovisual data. Second-order differential features include visual features like visual acceleration (Barzelay and Schechner, 2007) and audio features like onset (Barzelay and Schechner, 2007).

Taking physical meaning and the order of temporal differential as two axes, we summarize a table of this feature classification, which is listed in Table 2.1. The place where no feature has been developed is denoted as "-".

### 2.2.2 Different measures

Compared to features, the classification of measures is simple, which has only two classes: linear and nonlinear measures. Linear measures include Pearson correlation coefficient (Casanovas, 2006; Monaci et al., 2005), canonical correlation analysis (Kidron et al., 2005), linear approximation of mutual information (Fisher and Darrell, 2004),

and inner product (Barzelay and Schechner, 2007). Nonlinear measures include mutual information (Hershey and Movellan, 1999; Liu and Sato, 2008) and quadratic mutual information (Liu and Sato, 2008). Since nonlinear measures can evaluate higher order correlation between two random variables, they should have a better performance than linear ones. However, nonlinear measures require an estimation of probability density function. The accuracy and stability of this estimation significantly affects the correlation analysis by the nonlinear measures.

## 2.3 Evaluation of features and measures

We first compared the performance of different measures because this comparison can be done by using generated random numbers. In contrast to measures, the performance of features cannot be evaluated without a measure. We selected the measure that has the best performance in the comparison to evaluate the performance of features.

### 2.3.1 Evaluation of measures

Among all the linear correlation measures, we selected Pearson correlation coefficient as a representative one to evaluate the performance because Pearson correlation coefficient is the best measure if the two random variables are linearly correlated under a same coordinate system.

We evaluated the performance of all the two non-linear correlation measures — mutual information and quadratic mutual information. Both the former and the latter one can be adopted to indicate the correlation between either discrete or continuous random variables. However, when evaluating the correlation for continuous random variables, the former one requires integration over the probability density function, which usually cannot be computed analytically. This in fact restricts the usage of the former one in continuous situations. In most cases, random variables need to be first quantized to compute their mutual information in a discrete form. On the contrary, the latter one can be computed analytically for continuous random variables also. Therefore we evaluate them

both. The details on quadratic mutual information and its computation can be referred to Chapter 3 of this thesis.

The comparison experiments were performed by using random numbers that are uniform distributed between 0 and 1. The random numbers are generated by Mersenne twister algorithm (Matsumoto and Nishimura, 1998). Mersenne twister provides for fast generation of very high-quality pseudorandom numbers, having been designed specifically to rectify many of the flaws found in older algorithms.

In the first experiment, we compared different measures by evaluating the correlation between two random variables $x$ and $y$ which have deterministic relationship. We generated 100 uniformed distributed numbers for random variable $x$, and computed the samples of random variable $y$ with a deterministic mapping function. The mapping functions include linear and nonlinear maps. The generated 100 samples of $x$ and $y$ are illustrated in Figure 2.4.

The computation of Pearson correlation coefficient from the samples of $x$ and $y$ is straightforward following the definition (Rodgers and Nicewander, 1988). While the computation of mutual information and quadratic mutual information is not direct. Using the method introduced in Chapter 3, we first estimate an appropriate bandwidth to analyze the probability density function of the random variables $x$ and $y$. The bandwidths are independently estimated for $x$ and $y$. This bandwidth is then used as the bin size to quantize the samples of $x$ and $y$ and compute their mutual information in a discrete form, or used as the kernel bandwidth to compute quadratic mutual information directly from the samples of $x$ and $y$.

The evaluation results of the three measures are listed in Table 2.2. For linear maps like $y = x$ and $y = 2x + 3$, all measures demonstrated an invariance against the change of map by giving the same evaluated correlations. For nonlinear maps, this invariance is broken for all the measures. The measure whose evaluated correlation changed most is Pearson correlation coefficient. This is reasonable because its invariance is only guaranteed for linear maps. The second one is mutual information. Although mutual information is theoretically invariant to both linear and non-linear deterministic one-to-one

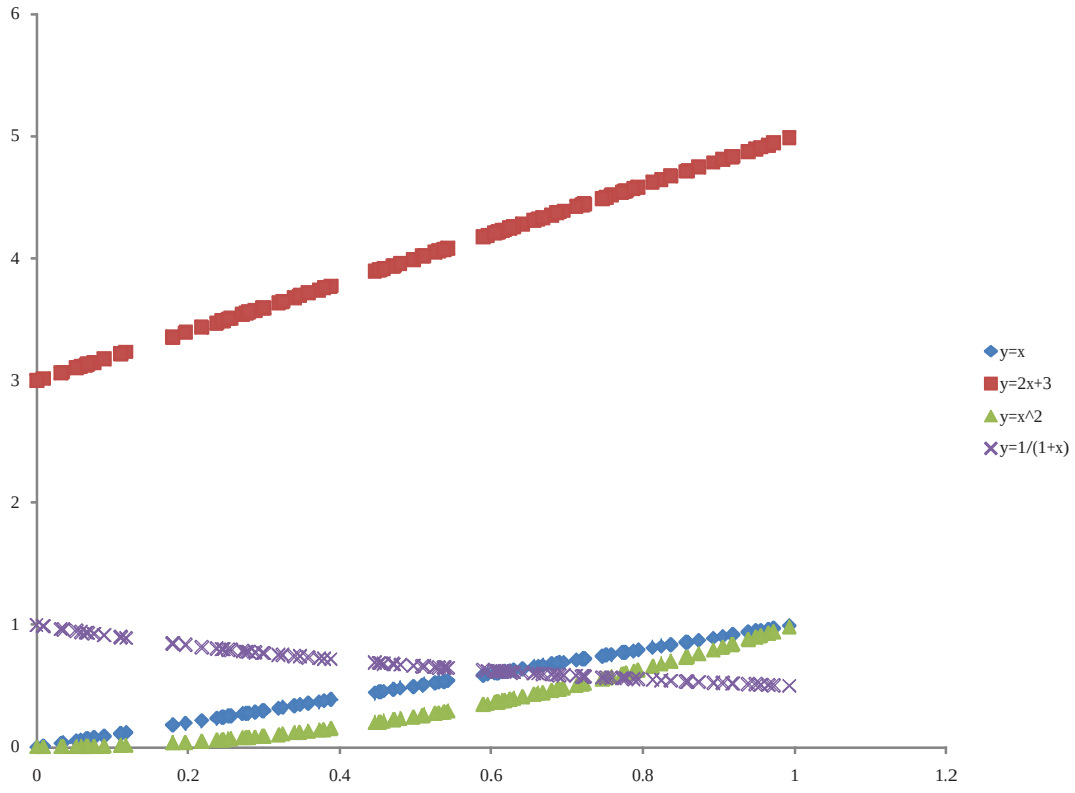



Figure 2.4. An illustration of the generated samples with different mapping functions.

Table 2.2. A comparison on measures with deterministic maps.

| | Pearson coefficient | Mutual information | Quadratic mutual information |
|---|---|---|---|
| $y = x$ | 1.0 | 0.90 | 0.0044 |
| $y = 2x + 3$ | 1.0 | 0.90 | 0.0044 |
| $y = x^2$ | 0.96 | 0.41 | 0.0041 |
| $y = 1/(1 + x)$ | -0.98 | 0.52 | 0.0042 |

maps, the requirement of quantization degrades this invariance significantly in our experiments. Among the three measures, quadratic mutual information demonstrated best invariance to both linear and nonlinear maps.

In the second experiment, we tested the performance of measures by evaluating the correlation between two random variables $x$ and $y$ whose relationship were partly deterministic. "Partly" here means that $y$ is related to not only $x$ but also an additive or multiplicative random noise. The noise $n$ was assumed to be Gaussian with $\mu = 0$ and $\sigma = 1$. The generated samples of $x$ and $y$ are illustrated in Figure 2.5.

The evaluation results of three measures are listed in Table 2.3. Again, quadratic mutual information demonstrated best invariance against noise. Interestingly, we found that all measures were more robust against multiplicative noise than additive noise. This follows the intuitive observation on the data samples illustrated in Figure 2.5.

### 2.3.2 Evaluation of features

Strictly speaking, the performance of a feature is not only determined by the feature itself, but also by the measure used. That is to say, a feature may work better with one measure. However, the number of the combinations of features and measures are so large that it is difficult to make a thorough experimental comparison for all of them. Therefore, we use quadratic mutual information as the only measure to evaluate all the features since it has been shown to be most robust among the three representative measures.

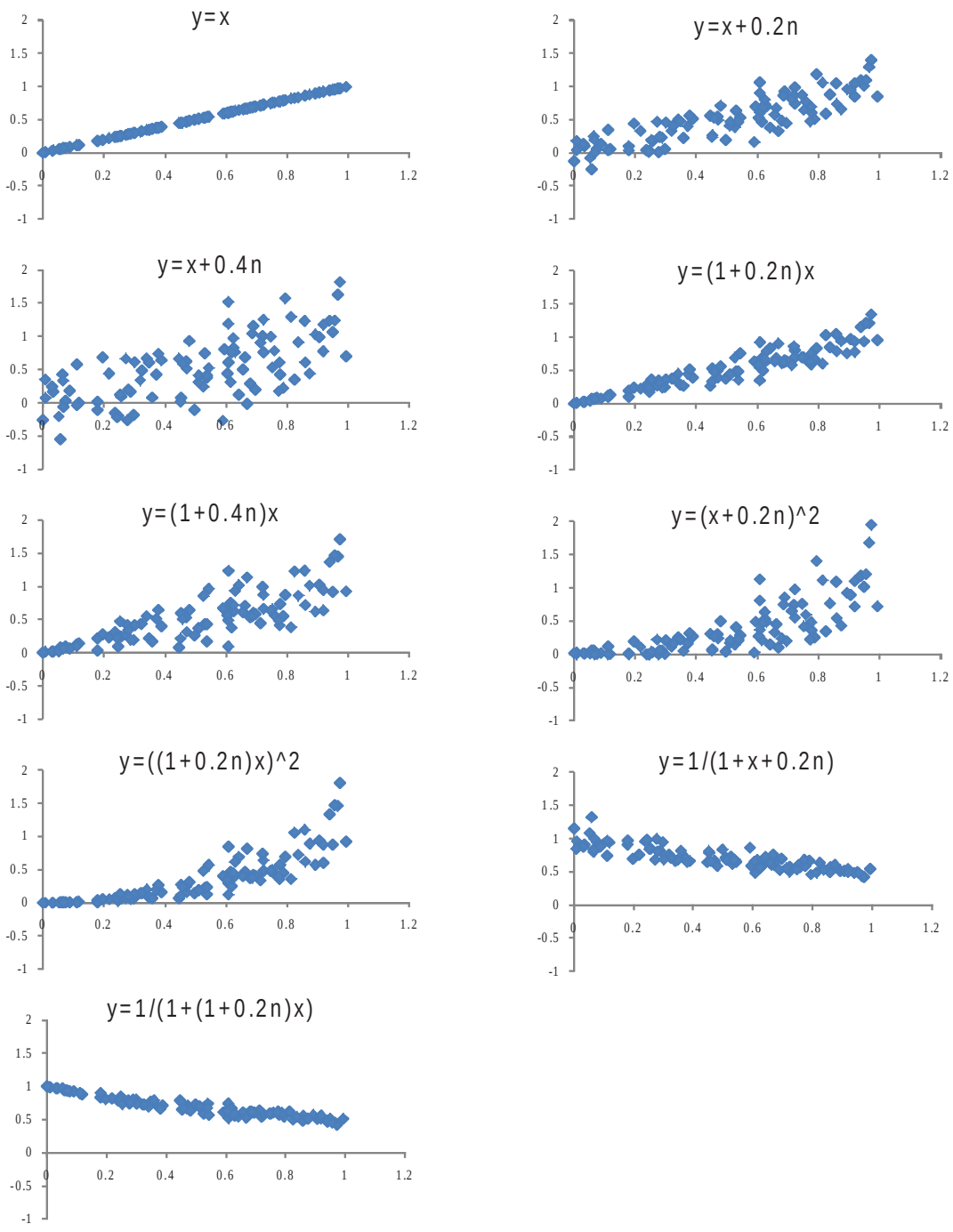



Figure 2.5. An illustration of the generated samples with different noise and mapping functions.

Table 2.3. A comparison on measures with noise.

| | Pearson coefficient | Mutual information | Quadratic mutual information |
|---|---|---|---|
| $y = x$ | 1 | 0.90 | 0.0044 |
| $y = x + 0.2n$ | 0.86 | 0.32 | 0.0032 |
| $y = x + 0.4n$ | 0.66 | 0.11 | 0.0019 |
| $y = (1 + 0.2n)x$ | 0.93 | 0.38 | 0.0037 |
| $y = (1 + 0.4n)x$ | 0.81 | 0.28 | 0.0027 |
| $y = (x + 0.2n)^2$ | 0.77 | 0.19 | 0.0024 |
| $y = ((1 + 0.2n)x)^2$ | 0.83 | 0.26 | 0.0028 |
| $y = 1/(1 + x + 0.2n)$ | -0.84 | 0.24 | 0.0030 |
| $y = 1/(1 + (1 + 0.2n)x)$ | -0.95 | 0.54 | 0.0039 |

The comparison experiments were performed with both simulation and real data. The simulated and real audiovisual data are demonstrated in Figure 2.6 and Figure 2.7, respectively.

The simulation data simulated an ideal situation where both audio and video were modulated by a same signal. The modulating function is a multiplication of two sine functions with a long and a short period to simulate a sound emission action. The coefficient of this modulation was given at each time $t$ by

$$c(t) = \max\{0.4 + 0.6 \sin(2\pi f_1 t) \sin(2\pi f_2 t), 0\}, \tag{2.6}$$

where $f_1 = 3$ s, $f_2 = 0.3$ s. Visual signal was synthesized by vertically shaking a random dot image (320×240) following the modulation coefficients computed by Equation (2.6). While audio was synthesized by sampling a modulated 2 KHz sine wave at 44.1 KHz. The modulation coefficients were the same as the visual ones.

The real data included two video clips. One photographed a stationary speaker who uttered English numbers from zero to ten in front of a green background. The other photographed a scene where a hand was playing a piano. Both visual data were monochromic

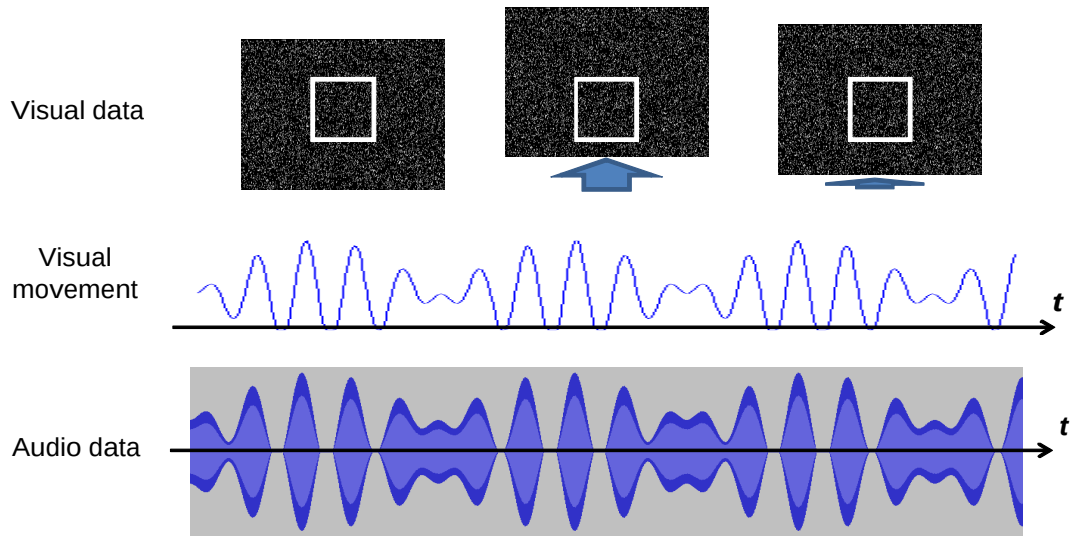



Figure 2.6. Audiovisual simulation data.

at a resolution of 240×160, and audio data were recorded at 44.1 KHz with stereo. We adopted its left channel only to compute the audio features.

Using these data, we performed experiments to compare the performance of features. Since there are too many combinations of audio and visual features, as listed in Table 2.1, we tested a subset of the combinations based on two rules. The first rule is that audio and visual feature should be at the same differential order, i.e., $n$-order audio feature is combined with $n$-order visual feature. This is reasonable since audiovisual feature should describe a same physical phenomenon such that the synchrony between them can be evaluated. The second rule is that we ignore the audio features that are based on the spectrum. Generally speaking, audio spectrum supplies a more intricate description on audio energy, i.e., it gives the magnitude of audio energy at each frequency. Thus methods with audio spectral feature often used this property to analyze the audiovisual correlation in a more subtle degree such that they can separate the sound individually for different frequencies. Yet, to analyze the audiovisual correlation subtly also degrades its robustness. If there is no noise in audio signal, we believe that using audio energy would be the best choice.

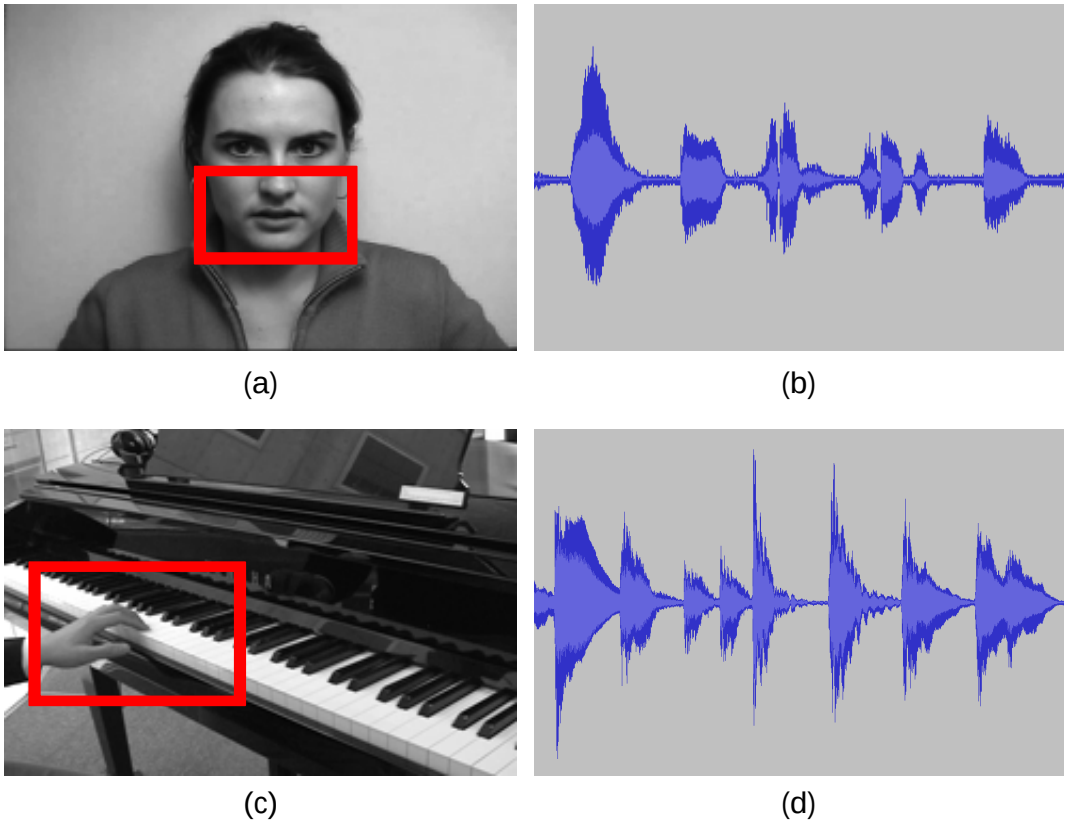



Figure 2.7. Audiovisual real data. Figures (a) and (b) respectively show the visual and audio data of the speaker video, and figures (c) and (d) show the data of the piano video. Red rectangles show the region where the audiovisual correlations are analyzed and averaged.

Table 2.4. A comparison on the magnitude of audiovisual features.

| | Intensity ($\times 10^{-6}$) | Translation ($\times 10^{-6}$) | Velocity ($\times 10^{-6}$) | Acceleration ($\times 10^{-6}$) | Inconsistency ($\times 10^{-6}$) |
|---|---|---|---|---|---|
| Simulation | 1200 | 290 | 150 | 67 | 2300 |
| Speaker | 630 | 300 | 150 | 39 | 170 |
| Piano | 230 | 190 | 88 | 29 | 100 |

Based on the above two rules, we got five combinations of audiovisual features. The experimental results of their performance comparison with simulation and real data are listed in Table 2.4 and Table 2.5.

In the lists, intensity represents value-related feature, and the others represent motion-related ones. Translation, velocity, acceleration, and inconsistency respectively represent the zero-, first-, second-, and transformed second-differential features. As a more detailed description, intensity, translation, velocity, acceleration, and inconsistency respectively correspond to the combination of pixel intensity and audio energy, the combination of visual translation and audio energy, the combination of optical flow and differential audio energy, the combination of visual acceleration and second-order differential of audio energy, and the combination of visual inconsistency and audio energy inconsistency.

A problem is that visual features like translation, velocity, and acceleration are two-dimensional vectors, and need to be converted into a scalar value to analyze the audiovisual correlation. Some works adopted the magnitude of this vector (for instance, (Barzelay and Schechner, 2007; Monaci et al., 2005)). While in (Liu and Sato, 2008), it was found that vertical element of this vector conveys higher correlation with audio than magnitude or horizontal element. Therefore, to make this comparison fair, we divide the comparison into the two lists. The Table 2.4 adopts the magnitude of all the visual features, and the Table 2.4 adopts the vertical element of all the visual features. Since visual intensity and inconsistency features have magnitude only, they were not included in the Table 2.4.

Table 2.5. A comparison on the vertical element of audiovisual features.

| | Translation ($\times 10^{-6}$) | Velocity ($\times 10^{-6}$) | Acceleration ($\times 10^{-6}$) |
|---|---|---|---|
| Simulation | 1300 | 840 | 53 |
| Speaker | 450 | 950 | 61 |
| Piano | 230 | 190 | 220 |

By comparing the experimental results of the two lists, it can be observed that the vertical element demonstrate a much higher correlation than that of the magnitude. This fact strongly supported the conclusions of (Liu and Sato, 2008) that the vertical element of the two-dimensional motion vector conveys higher correlation with audio than the others.

In the comparison, all features gave better performances on simulation data. Among them, inconsistency feature demonstrated an astonishing performance. Yet the performance degraded largely on real data. We believe that the reason of this lies in the visual motion discontinuity. Although visual inconsistency feature is designed to describe the visual acceleration of a visual motion, if it is extracted at the border of the areas with different visual motions, it demonstrates a high inconsistency. Visual simulation data do not have visual motion discontinuities. While such situations happen frequently in real data when we extract visual inconsistency features at a fixed position. That is why Liu and Sato adopted a visual path optimization to trace the changing positions of pixels by using inconsistency feature.

For speaker, taking the vertical element of optical flow gave the best performance. We believe that the reason of this lies in not only that speaking actions happen mainly vertically, but also the simplicity to compute the first-order differential feature. Compared to it, the computation of the zero-order differential feature requires a tracking of the positions of pixels for all the visual frames, and the computation of the second-order differential feature requires this tracking for at least three frames.

An interesting discovery in these two lists is about the intensity feature. After Monaci *et al.* (Monaci et al., 2005) claimed that the movement of photographed sound source conveys better correlation with auditory information than the change of their visual appearance, most works have changed to base their visual feature on the motion. However, our comparison data demonstrated that intensity feature can give comparative performance as well. Considering the easiness to extract an intensity feature, we believe that it is worthy of researching in the future. In face the same phenomenon has been addressed in facial expression recognition. Although most researches in this field try to analyze facial expression based on shape deformation, a comparative performance was demonstrated by using the change of intensity of facial pixels in (Kumano et al., 2007).

# Chapter 3

# Face region segmentation of a stationary speaker

## 3.1 Introduction

The ability to detect the position of a speaker is useful for various applications, such as video processing and content analysis. For example, a video-teleconferencing system may need to focus on a speaker, or a video analysis system may have to associate uttered words to a speaker. Being able to identify the speaker's face region is furthermore preferred because this makes various effects possible, such as to automatically emphasize a speaker by blurring all other persons and background, or, on the contrary, to impose mosaic over an interviewee to protect privacy. An example is shown in Figure 3.1 based on the results of our method.

However, regardless of the great progresses made in face and human detection (e.g., (Viola and Jones, 2004) and (Dala and Triggs, 2005) respectively) in recent years, speaker detection is still under development. As the purpose is to distinguish a person from others, either face detection or human detection fails in this area. One solution (Luettin et al., 1996) is to detect the face, locate the mouth, and check its movement. The weakness of this method is the requirement of a frontal view. View dependency is also a

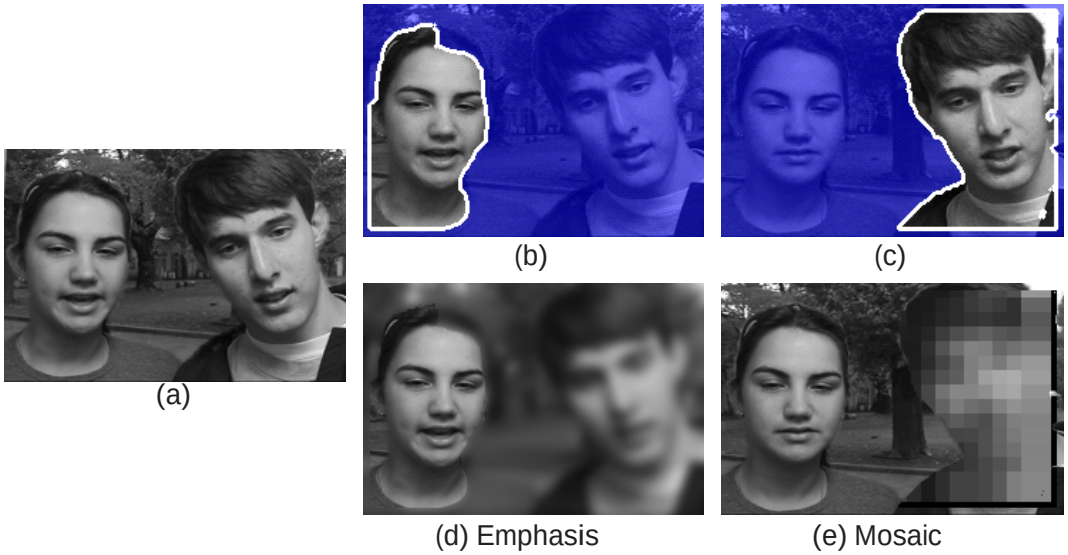



Figure 3.1. Special effects given the speaker's face region. Figure (a) shows the original image, figures (b) and (c) show the speaker face region localized by our method, and figures (d) and (e) show the special effects imposed based on our estimation.

challenging problem for face and human detections (Huang et al., 2005). Additionally, it is prone to be disturbed by unconscious movements from other persons.

In this work, we develop a novel technique to find the speaker's face region for different time windows, which is robust against the changes of view, scale, and background. This technique is based on the recent developments in sound source localization by audiovisual correlation analysis and an segmentation technique of multiple video frames.

To localize an audio source by audiovisual correlation analysis is a relatively new research topic and has drawn much attention in recent years. Based on neuroscience discoveries (Bertelson et al., 1994; Driver, 1996; McGurk and MacDonald, 1976), many approaches have been developed to analyze the audiovisual correlation (for instance, (Fisher and Darrell, 2004; Hershey and Movellan, 1999; Kidron et al., 2005; Monaci et al., 2005)), which have been reviewed in Chapter 2.

Unfortunately, all existing localization methods suffer a common problem: the estimated mask of sound source is highly fragmental. Therefore, they experience difficulties in designating a correct speaker position, much less identifying a reliable speaker region.

Most of them only detect pixels that are supposed to be the sound source, except that Casanovas (Casanovas, 2006) clustered the detected pixels and adopted the cluster center as the speaker position. Yet clustering may be vulnerable to the outliers that appear often (Fisher and Darrell, 2004; Kidron et al., 2005; Monaci et al., 2005).

To be able to detect a reliable speaker region, we consider a novel technique, whose key idea is to integrate audiovisual correlation analysis into a video segmentation framework. In contrast to sound source localization, image segmentation has been researched for decades (for instance, (Kass et al., 1988; Zhu and Yuille, 1996)). Recently, Boykov and Funka-Lea made an important progress step (Boykov and Funka-Lea, 2006), in which a globally optimum segmentation, which balances pixel likelihood and image region information, is found efficiently using graph cut. The method works for not only a single image, but also for multiple video frames with inter-frame continuity considered (Boykov and Funka-Lea, 2006). A weakness of graph cut is the requirement of a manual operation to designate seeds of foreground and background. Fortunately, our incorporation of audiovisual correlation analysis not only takes advantage of the effective optimization of graph cut, but also removes the necessity of this manual operation.

Other works have also been developed to incorporate information into graph cut-based segmentation to enhance the performance and remove the manual operation. Kolmogorov *et al.* (Kolmogorov et al., 2005) adopted stereo depth information to segment foreground. Yu *et al.* (Yu et al., 2007) based their method on face detection to segment people. Schoenemann and Cremers (Schoenemann and Cremers, 2008) took advantage of motion information to divide motion layers. However, their incorporated information is still based on visual signal and cannot supply the cues beyond the visual signal. For example, if both people move their mouths, without audio, one can hardly tell who the real speaker is. Fusing audio not only resolves this ambiguity, but also improves the robustness compared to the usage of visual signal only. To the best knowledge of the authors, this is the first trial to fuse other modality information into the Graph Cut-based segmentation.

We analyze the audiovisual correlation by computing quadratic mutual information between our audio and visual features. We extract visual features locally, whose locality

helps our method to be robust to the change of view, and compute quadratic mutual information to analyze the audiovisual correlation. We also estimate the kernel bandwidth from data when computing quadratic mutual information, which makes our method adaptive to the changes of visual scale and audio gain. The audiovisual correlation analyzed locally is incorporated into a global optimization framework to extract the speaker's face region, which is based on video segmentation by graph cut. To avoid a heuristic decision of a segmentation threshold, we learn the distributions of the audiovisual correlation of speaker and background by using expectation maximization. The likelihoods of each pixel to these two distributions are combined with image smoothness constraints to form the energy function in the graph cut segmentation.

Our system requires that the speaker must stay nearly at the same position in the estimation time window for the sake of audiovisual correlation analysis, as was assumed in previous methods (Fisher and Darrell, 2004; Hershey and Movellan, 1999; Kidron et al., 2005). The time window is generally within 2–4 seconds.

We detect the speaker's face region but not the mouth region because, when speaking, many unconscious movements happen also on face parts, which are as highly correlated with the audio as mouth movements. Consequently, in most cases our method can detect the whole face region. However, as discussed in Section 3.4, if the speaker intentionally restrains these unconscious movements, only the mouth region of this speaker can be detected. If the speaker position only is needed, this will not be a problem as the mouth region center can be adopted. If the whole face region is needed, a manual extension of the segmentation mask is necessary in such cases.

The main contribution of this work can be summarized as follows: 1) to find the speaker's face region by incorporating audiovisual correlation analysis into video segmentation, including the method to locally analyze the audiovisual correlation and the learning of correlation distributions, and 2) to adopt audio information to eliminate the manual operations in Graph Cut-based segmentation and improve its robustness against complex backgrounds.

The rest of this work is organized as follows. In Section 3.2, we introduce the au-

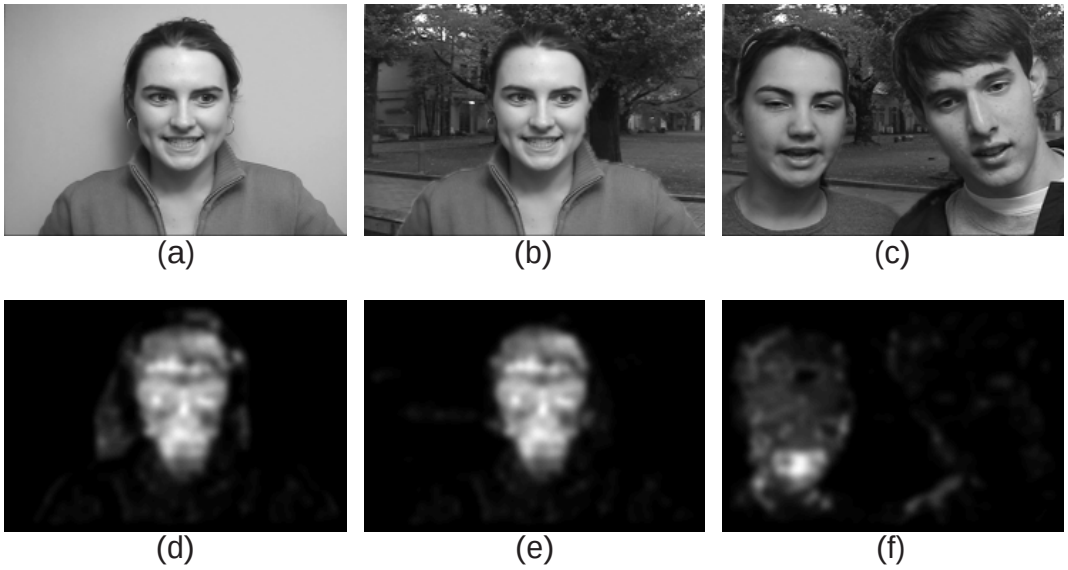



Figure 3.2. Analyzed audiovisual correlation for different video sequence. Correlation in (d–f) are normalized independently. The whiter a pixel, the higher its correlation.

diovisual feature and the correlation computation using quadratic mutual information. In Section 3.3, we explain how we find the speaker's face region by performing video segmentation based on the audiovisual correlation. In Section 3.4, we present and discuss our experimental results. In Section 3.5 we present our conclusions.

## 3.2 Audiovisual correlation analysis

Within a time window, we extract the visual feature at each local position $(x, y)$ and each time $t$, and the audio feature at each time $t$. The correlation between the temporal changes of the visual feature at $(x, y)$ and the audio feature is analyzed by using quadratic mutual information. After the analysis, we can get a table $C(x, y)$ which shows the audiovisual correlation of each image position $(x, y)$ in the current time window. An example of the audiovisual correlation table is shown in Figure 3.2.

Below we first introduce our audiovisual feature and then explain the correlation analysis by using quadratic mutual information. Note that our audiovisual correlation

analysis is based on the prerequisite that the audio and visual signals are recorded synchronously. Asynchronous data may degrade the accuracy of this analysis.

### 3.2.1 Our audiovisual feature

Both our audio and visual features describe the differential between two continuous frames. However, because of the substantial difference between the audio and visual signals, their extraction methods differ significantly.

#### Audio feature

Since audio is usually sampled at a much higher frequency than video, we first divide audio samples into frames to compute the audio feature. The frame duration, $T_a$, is set to be the same as the visual frame duration, $T_v$. In order to keep a temporal continuity, it is set such that each pair of two successive frames have an overlap of the duration of $T_a/2$. Additionally, to reduce the boundary effect, a Hamming window is multiplied (Rabiner and Juang, 1993), whose coefficients are computed by

$$w(i) = 0.54 - 0.46\cos(\frac{\pi i}{M}), \quad i = 1, \cdots, M \tag{3.1}$$

where $M$ is the number of the audio samples in a $2T_a$ duration. The audio energy $e(t)$ of frame $t$ is computed by

$$e(t) = \log\left(\frac{1}{M}\sum_{i=1}^{M}(w(i)s(t,i))^2\right), \tag{3.2}$$

where $s(t,i)$ refers to the processed audio sample $i$ in frame $t$ and the two sourrounding overlaps of frame $t$. This process is demonstrated in Figure 3.3.

The audio feature is defined as the differential energy between the current and next frames, which is given by

$$a_t = e(t+1) - e(t). \tag{3.3}$$

Since in silence durations the absence of audio information makes it impossible to analyze the audiovisual correlation, we ensure there is speech in all frames. This is done by

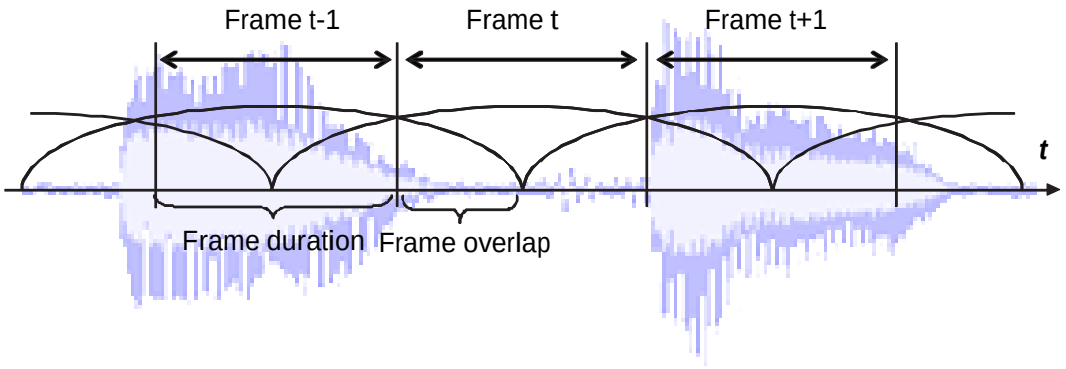



Figure 3.3. Division of audio frames.

checking whether or not audio energy $e(t)$ is larger than a pre-defined threshold. Frames failing this test are regarded as silent and dropped, together with their corresponding visual frames. Only frames passing this test are buffered till the frame number reached a pre-defined value. If we discuss $N$ audiovisual frames in this chapter, it refers to frames that are buffered.

### Visual feature

The same as (Monaci et al., 2005), we believe that there is an audiovisual correlation mainly between the movements of visual objects and the change of audio. Therefore, optical flow is used as the visual feature in our method. In particular, we only take the vertical element of optical flow considering that most speaking actions move vertically. We have compared three methods to get a scalar visual feature from the 2D optical flow vector: horizontal element, vertical element, and amplitude. The results of analyzed audiovisual correlation are shown in Figure 3.4. The vertical element obviously has much higher correlation with the audio feature.

Visual feature $v_t(x, y)$ is defined as the vertical optical flow extracted at $(x, y)$ between frames $t$ and $t + 1$, which is computed by the Lucas-Kanade method (Lucas and Kanade, 1981). Since optical flow cannot be estimated stably in areas with less texture,

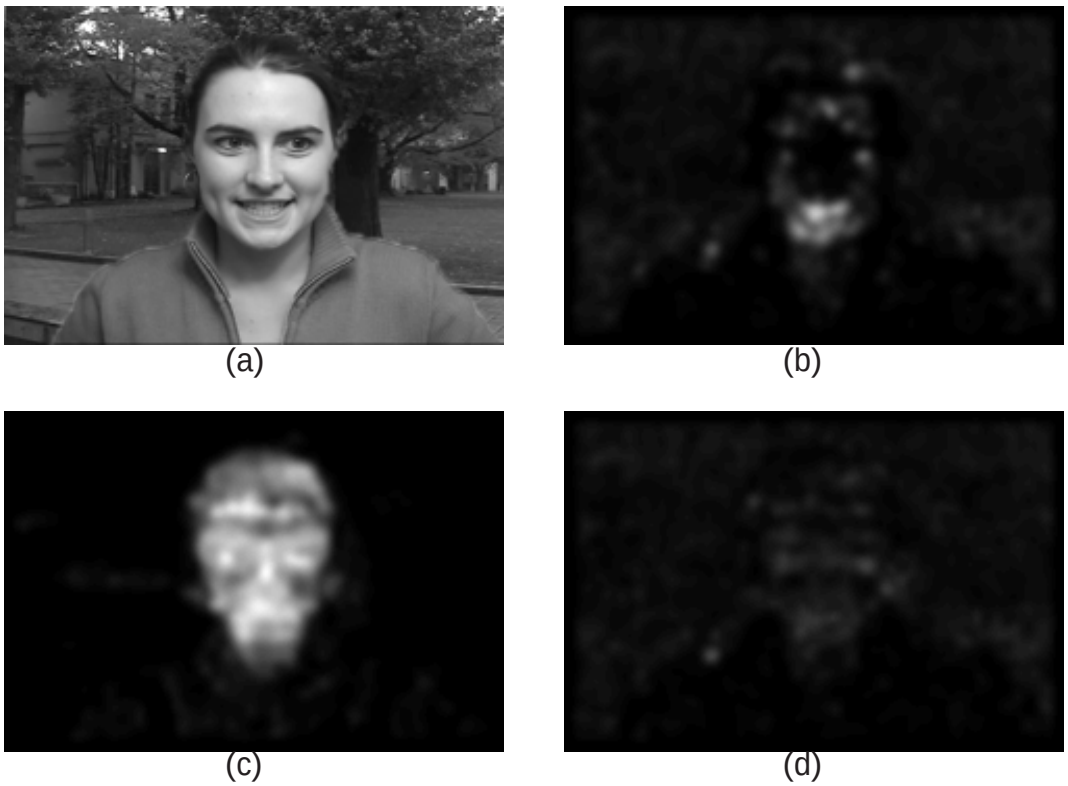



Figure 3.4. Audiovisual correlation with different optical flow elements. Figure (a) shows an original video frame, and figures (b), (c) and (d) show the analyzed audiovisual correlation by using horizontal element, vertical element, and amplitude, respectively.

we verify the variation of pixel intensities inside each window where we compute optical flow. If these are below a threshold, we set the flow value to be zero.

Adopting optical flow as the visual feature has three advantages for our system. First, it helps our system to be background robust. For a static background, movement is independent to its complexity. For a moving background, as its movements usually correlate marginally with audio, the influence can be suppressed in the subsequent correlation analysis also. Second, it helps our system to be view robust. No matter how different a face looks in different views, the local movements resulted from speaking action are similar. Optical flow captures this local movement and is thus view robust. Third, the locality of the optical flow also makes it possible for our method to achieve good segmentation boundary. Since optical flow describes the movement of each pixel, our segmentation can achieve an accuracy of every pixel.

### 3.2.2 Audiovisual correlation by quadratic mutual information

Many works have used mutual information to measure the audiovisual correlation (Fisher and Darrell, 2004; Hershey and Movellan, 1999). However, to analytically compute mutual information for continuous random variables, they either computed the second-order Taylor extension of mutual information (Fisher and Darrell, 2004), or assumed that audio and visual features obey normal distribution (Hershey and Movellan, 1999). The former one can approach only an approximation of mutual information and requires an iterative computation process. The latter one is arguable since obeying normal distribution is a strong assumption. We have applied a normality test (Doornik and Hansen, 1994) to our audio and visual features extracted inside a speaker mouth region. The results showed that $77.8\%$ of the tests fall into the refusal area $[9.49, +\infty)$ with $\rho$-value=0.05. That is to say, this assumption should be wrong in a confidence of 95%.

We use quadratic mutual information as a measure to analyze the audiovisual correlation, which can be computed analytically directly from the data without the necessity of any approximation and assumption. The computation of quadratic mutual information is based on the probability density functions (pdf) estimated by kernel density estimation

(Parzen, 1962). The bandwidth of the kernel density estimation is estimated from the variance of the data, which makes our method robust to the changes of visual scale and audio gain.

Quadratic mutual information is computed based on the temporal samples of the audio and visual features. This analysis is independently performed for different image positions by using the visual feature extracted at each image position $(x, y)$. After quadratic mutual information at all positions $(x, y)$ are computed, we can get the correlation table $C(x, y)$, which is used to segment out the speaker's face region in the next stage.

### Pdf estimation by kernel density estimation

Kernel density estimation (Parzen, 1962) (also known as Parzen window estimation) is a method of estimating the arbitrary pdf of a random variable (Parzen, 1962). Given $N$ data points $\{\mathbf{z}_i,\ i = 1, \cdots, N\}$, in $n$-dimensional space $R^n$, the multivariate kernel density estimation with kernel $K_{\mathbf{H}}(\mathbf{z})$ and a symmetric positive definite $n \times n$ bandwidth matrix $\mathbf{H}$, computed in point $\mathbf{z}$ is given by

$$p(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{z} - \mathbf{z}_i), \tag{3.4}$$

where $K_{\mathbf{H}}(\cdot)$ is the specified kernel function.

We adopt a Gaussian kernel with a diagonal bandwidth matrix, $\mathbf{H} = diag(\sigma_1^2, \cdots, \sigma_d^2)$. Compared to the other kernels, such as a triangle kernel, a Gaussian kernel results in an efficient computation of quadratic mutual information.

The selection of an appropriate bandwidth is important for kernel density estimation (Turlach, 1993). Small bandwidth values make the estimate look "wiggly" and show spurious features, whereas too big values will lead to an estimate which is too smooth in the sense that it is too biased and may not reveal structural features. Therefore, comparing an empirical decision of this kernel bandwidth, we estimate the bandwidth from data. For Gaussian kernel, a *rule of thumb* was proposed to estimate a proper bandwidth from the data (Turlach, 1993). We adopt this *rule of thumb* to compute the bandwidth as

$$\sigma = 1.06 \hat{\sigma} n^{-\frac{1}{5}}, \tag{3.5}$$

where $\hat{\sigma}^2$ is the sample variance.

### Correlation analysis by quadratic mutual information

Quadratic mutual information is proposed by Xu *et al.* (Xu et al., 1998) in 1998, which has its root in the quadratic form of Renyi entropy.

In 1961, Renyi (Renyi, 1961) showed that entropy of a random variable can be evaluated by a group of functions defined as

$$H_\alpha(x) = \frac{1}{1-\alpha} \log \left( \int p^\alpha(x) dx \right), \tag{3.6}$$

where $\alpha > 0, \alpha \neq 1$. As $\alpha \to 1$, $H_\alpha(x)$ approaches the Shannon entropy $H(x)$. In practice, the one most often used is its quadratic form, i.e., $\alpha = 2$, which can be efficiently computed based on the pdf estimated by kernel density estimation with Gaussian kernels. Supposing Gaussian kernel is represented as

$$K_\Sigma(\mathbf{x}) = G(\mathbf{x}, \Sigma), \tag{3.7}$$

it is easy to show that

$$\int G(\mathbf{x} - \mathbf{x}_i, \Sigma) G(\mathbf{x} - \mathbf{x}_j, \Sigma) d\mathbf{x} = G(\mathbf{x}_i - \mathbf{x}_j, 2\Sigma). \tag{3.8}$$

Substituting Equation (3.7) and Equation (3.8) into Equation (3.6) with $\alpha = 2$, quadratic Renyi entropy can be analytically computed as

$$H_2(x) = -\log \sum_{i=1}^{N} \sum_{j=1}^{N} G(x_i - x_j, 2\Sigma). \tag{3.9}$$

Although Renyi entropy can be analytically computed, it cannot analyze the correlation between two random variables. We need a measure like mutual information but in Renyi form to analyze this correlation. Such measure left undefined until 1998 when Xu *et al.* (Xu et al., 1998) proposed quadratic mutual information. Qudratic mutual information is extended from the quadratic form of Renyi entropy and thus has the same character of being able to be computed analytically based on the pdf estimated by kernel density estimation.

We use this quadratic mutual information to compute the audiovisual correlation between the audio and visual features. The same as mutual information, quadratic mutual information indicates the amount of information that one random variable conveys about another. At an image position $(x, y)$, quadratic mutual information is computed between the audio feature $a$ and visual feature $v$ by definition as

$$QMI(a; v) = \log \frac{\iint p^2(a, v)dadv \iint p^2(a)p^2(v)dadv}{(\iint p(a, v)p(a)p(v)dadv)^2}. \tag{3.10}$$

It can be shown that $QMI(a; v) \geq 0$ and the equality hold true if and only if $p(a) = p(v)$ using Cauchy-Schwartz inequality (Xu et al., 1998).

As mentioned before, quadratic mutual information can be computed analytically directly from the data. Given the temporal samples of the audio and visual features to be $\{(a_t, v_t), t = 1, \cdots, N\}$, it has been shown (Xu et al., 1998) that quadratic mutual information can be computed as

$$QMI(a; v|\{a_t, v_t\}) = \log \frac{V_c(\{a_t, v_t\})V_m(\{a_t\})V_m(\{v_t\})}{V_{nc}^2(\{a_t, v_t\})} \tag{3.11}$$

where $V_c(\{a_t, v_t\})$, $V_m(\{a_t\})$, $V_m(\{v_t\})$, and $V_{nc}(\{a_t, v_t\})$ are the terms computed from the data samples, which are given by

$$V_c(\{a_t, v_t\}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K_{2\sigma_a^2}(a_i - a_j)K_{2\sigma_v^2}(v_i - v_j), \tag{3.12}$$

$$V_m(\{a_i\}) = \frac{1}{N} \sum_{j=1}^{N} V_s(a_j, \{a_i\}), \tag{3.13}$$

$$V_m(\{v_i\}) = \frac{1}{N} \sum_{j=1}^{N} V_s(v_j, \{v_i\}), \tag{3.14}$$

$$V_{nc}(\{a_t, v_t\}) = \frac{1}{N} \sum_{j=1}^{N} V_s(a_j, \{a_i\}V_s(v_j, \{v_i\}). \tag{3.15}$$

Inside them, $V_s(a_j, \{a_i\})$ and $V_s(v_j, \{v_i\})$ are computed as

$$V_s(a_j, \{a_i\}) = \frac{1}{N} \sum_{i=1}^{N} K_{2\sigma_a^2}(a_j - a_i), \tag{3.16}$$

$$V_s(v_j, \{v_i\}) = \frac{1}{N} \sum_{i=1}^{N} K_{2\sigma_v^2}(v_j - v_i). \tag{3.17}$$

Here $K_\sigma(\cdot)$ is a one-dimensional Gaussian kernel. $\sigma_a$ and $\sigma_v$ are the estimated bandwidths of the audio and visual features obtained by using Equation (3.5).

Using Equation (3.11), we can compute quadratic mutual information directly from the samples without even the necessity to explicitly formulate pdf. Additionally, although the complexity of quadratic mutual information computation by using the definition in Equation (3.10) is O($N^4$), by using Equation (3.11) we can compute it at a complexity of O($N^2$) by removing duplicated computation.

As shown in Appendix A, with our bandwidth estimation, this correlation analysis is invariant to the scale changes of both audio and visual features. This invariance makes our method robust against the changes of both visual image scale and audio signal gain.

## 3.3 Segmentation of speaker's face region

We incorporate the analyzed audiovisual correlation into graph cut-based video segmentation to segment speaker's face region.

Again using the retrieved $N$ video frames, we build a N-D image as defined in (Boykov and Funka-Lea, 2006) and perform the video segmentation. To avoid a heuristic threshold for this segmentation, we learn the correlation distributions of speaker and background. The segmentation is performed based on the likelihood of each pixel to these two distributions. Note that, since there is only one scalar correlation value at each image position $(x, y)$, the computed distance is same for all the pixels at $(x, y)$, regardless of in which frame $t$ they are. On the other hand, as image information, like edge, pixel similarity and intra-frame continuity, is related to both $(x, y)$ and $t$, segmentation results can still be different in each frame and capture the face deformation when speaking.

### 3.3.1 Graph Cut-based segmentation

Segmentation of video frames by optimizing a global energy function was proposed in (Boykov and Funka-Lea, 2006). The global energy function is composed of two important terms: the sum of data costs of all the pixels, and the sum of the smoothness penalties between every two neighboring pixels in both temporal and spatial domains, whose definition is given by

$$E(l) = \sum_{p} D_p(l_p) + \lambda \cdot \sum_{\{p,q\} \in Ne} S_{pq}(l_p, l_q), \tag{3.18}$$

where $l$ represents the segmentation labels of all the pixels in the N-D image. $l_p = 1$ means pixel $p$ is labeled as speaker, while $l_p = 0$ means background. $Ne$ defines the neighborhood relationship between two pixels, which is discussed in detail in Section 3.3.3. $\lambda$ is a constant that adjusts the balance between the data costs and the smoothness penalties.

It has been shown that the energy function defined in Equation (3.18) can be efficiently optimized by calculating the minimum cut of a graph using a maximum flow algorithm (Boykov and Kolmogorov, 2004). Moreover, the optimization result is guaranteed to be the global minimum solution of the energy function (Boykov et al., 2001).

### 3.3.2 Data cost by audiovisual correlation

To compute the data costs, we first learn the correlation distributions of speaker and background by using expectation maximization algorithm. These two distributions are assumed to be one-dimensional Gaussian, whose parameters are learnt by the process below. First, the highest and lowest audiovisual correlation values are selected as two seeds. Then, by iteratively applying expectation maximization algorithm to all correlation values, we can compute an optimum estimation of the parameters of the two Gaussian distributions. The one trained from the seed of the lowest correlation is regarded as the distribution of background, denoted as $G(\mu_0, \sigma_0^2)$. The other is regarded as the distribution of speaker, denoted as $G(\mu_1, \sigma_1^2)$.

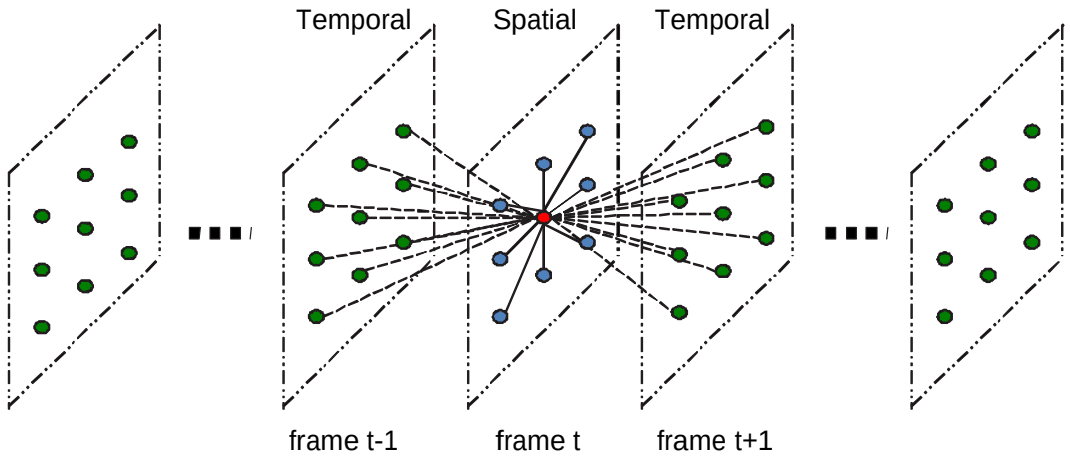



Figure 3.5. A demonstration of the N-D image and the neighborhood.

We find that the number of iterations performs the role of controlling the degree of how high an audiovisual correlation should have a higher likelihood to be speaker than the one to be background. If this number is too high, the background pixels that have relatively high audiovisual correlation may be wrongly segmented as speaker. We empirically find that three iterations are enough for this learning process.

The data cost of each pixel in Equation (3.18) is determined by the Mahalanobis distance to the correlation distributions of speaker and background, which is computed as

$$D_p(l_p) = \begin{cases} (C(x,y) - \mu_1)^2/\sigma_1^2 & l_p = 1 \\ (C(x,y) - \mu_0)^2/\sigma_0^2 & l_p = 0 \end{cases} . \tag{3.19}$$

### 3.3.3 Smoothness penalties by image information

Smoothness penalties are forced between every two neighboring pixels in both spatial and temporal domains. In the N-D image, the spatial and temporal neighborhoods are defined as shown in Figure 3.5. Each pixel can maximally have 26 neighbors.

The value of smoothness penalty is computed by

$$S_{pq}(l_p, l_q) = \exp\left(-\beta(I_p - I_q)^2\right) \cdot (d(p,q))^{-1} \cdot T[l_p \neq l_q], \tag{3.20}$$

where $p$ and $q$ are two neighboring pixels. $I_p$ and $I_q$ are their intensity values. The constant $\beta$ is chosen as in (Boykov and Funka-Lea, 2006) to be

$$\beta = \left(2\left\langle (I_p - I_q)^2 \right\rangle\right)^{-1}, \tag{3.21}$$

where $\langle \cdot \rangle$ denotes the expectation over the N-D image sample. This choice of $\beta$ ensures that the exponential term in Equation (3.20) switches appropriately between high and low contrast. $d(p, q)$ calculates Euclidean distance between $p$ and $q$ in the three-dimensional grid, which may be 1, $\sqrt{2}$ or $\sqrt{3}$ in our neighborhood model. $T[\cdot]$ is a boolean function returning 1 when the condition inside is true and 0 otherwise.

## 3.4 Experimental results

We adopted both simulation and real data to test the performance of our method. All videos of the data were or were supposed to be filmed at 30fps, while the audios were sampled at 44.1 kHz, since most recent off-the-shelf video cameras supply such audiovisual data. As for the algorithm parameters, in all our experiments, the balance constant in Graph Cut is set as $\lambda = 20$. The window for optical flow computation is of the size $9 \times 9$. The threshold for the texture verification in a window is set as 3. Except the experiments in Figure 3.7, we adopt 40 audiovisual frames to compute the results.

As for the computation time, it takes about 31 seconds to do segmentation for 40 frames at a resolution of $240 \times 160$ on our laptop, which is equipped with an Intel Core2 1.83 GHz CPU and a 1 GB RAM.

### 3.4.1 Simulation

To test the performance of our method when visual and audio signals change following an ideal pattern of a speaking action, we simulated a video clip and applied our method to it. Visual data photographed the movements of a random dot pattern, which was at a resolution of QVGA, $320 \times 240$. To include both mouth and background movements, we divided the dot pattern into two parts: a central rectangle face

Table 3.1. Segmentation performance on simulation data.

| | Correct rate (%) |
|---|---|
| Slow bg | 95.9 |
| Fast bg | 96.5 |

region and a background region. The central rectangle region was set to be slightly lighter than the background, as shown in Figure 3.6 (a). Both parts shook vertically. The central region moved synchronously with the audio change to simulate speaking movements. It was realized by computing the vertical shift at each time $t$ by the function $c(t) = \max\{\sin(2\pi f_f t), 0\}$, which was also adopted to modulate the magnitude of the audio. The audio, a 2 kHz modulated sine wave, was shown in Figure 3.6 (b). Alternatively, the vertical shift of the remained region was computed by another sine function $c(t) = \max\{\sin(2\pi f_b t), 0\}$.

As the central region simulates mouth movement, $f_f$ was set as $1/0.7$ Hz. For the remained region, we first chose a low frequency as $f_b = 1/2.3$ Hz to simulate a slow change background. The computed audiovisual correlation and the segmentation results are shown in Figure 3.6 (c) and (d), respectively. Furthermore, we chose a high frequency as $f_b = 1/0.4$ Hz to simulate a fast change background. The experimental results are shown in Figure 3.6 (e) and (f). In both cases, our method detected much higher audiovisual correlation in the central region and successfully segmented the region out.

The quantitative evaluation results were shown in Table 3.1. For both video clips, our method achieved high correct rate of the segmentation, which was the percentage of the number of correctly detected foreground and background pixels in the total number of the pixels.

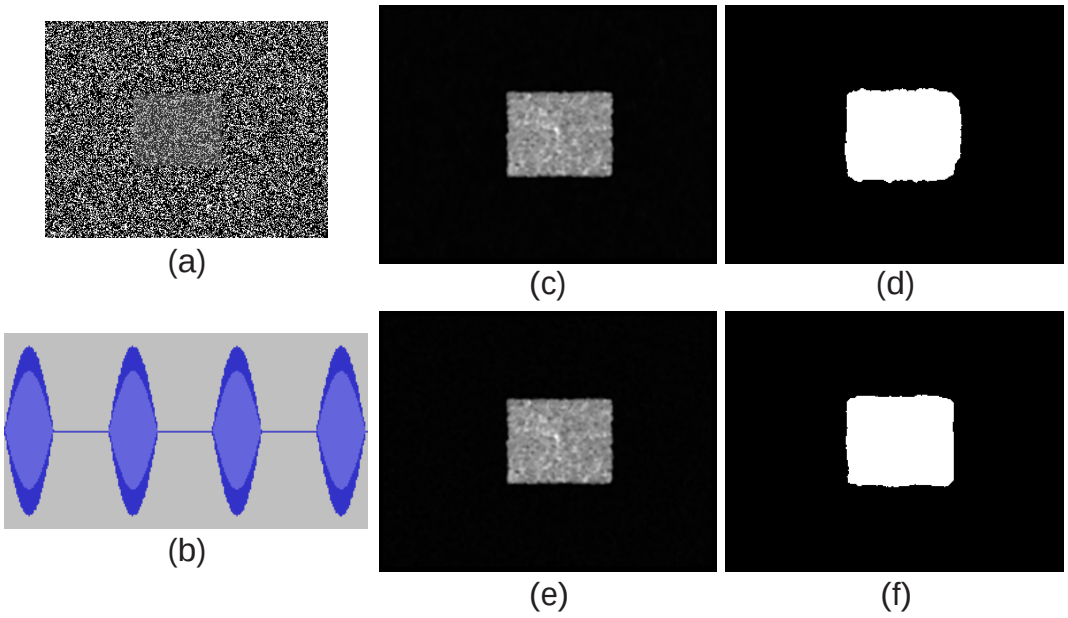



Figure 3.6. Segmentation results of simulation data. Figures (a) and (b) respectively show the visual random dot pattern and the audio, figures (c) and (e) show the analyzed audiovisual correlation, and figures (d) and (f) show the mask of the segmented face region.

### 3.4.2 Real data

For real data, we adopted CUAVE audiovisual database (Patterson et al., 2002), where 17 females and 19 males uttered English numbers in front of a green background with frontal, lateral, and moving views. An advantage of CUAVE database was that we could remove the green background by chroma-key and place other complex backgrounds to test the performance of our algorithm. Color images were then converted into gray images and down-sampled from $720 \times 480$ to $240 \times 160$ to make it possible to perform all the experiments in the 1 GB memory of our laptop.

We first investigated the relationship between the length of the time window and the analyzed audiovisual correlation. The experimental results are shown in Figure 3.7. It can be observed that a longer length of the time window helps to remove the ambiguity to determine the current speaker from Figure 3.7 (c) to Figure 3.7 (d). However, a longer time window causes our assumption more possible to be broken, as we assume that the speaker remains stationary in the processing time window. That is the reason we adopt 40 frames, which seems to be a good tradeoff.

We tested the segmentation performance with different backgrounds. The results are shown in Figure 3.8 (a–c), inside which only three of the 40 frames are shown. The segmented face region changed marginally under different backgrounds. To perform a comparison, we implemented the method in (Boykov and Funka-Lea, 2006). The manual seed we designated and the segmentation results over the same 40 frames are shown in Figure 3.8 (d–g). The segmented face region changed largely for different backgrounds. Note that this does not mean a comparison of performance, as the results by (Boykov and Funka-Lea, 2006) can be improved iteratively by adding seeds. However, the results here demonstrate that, through fusing audio information, our method can achieve better robustness to the change of background.

We tried to detect the speaker's face region with both frontal and lateral views. The results are shown in Figure 3.9. Since frontal and lateral views are two extreme cases of the view change, the success of our method to process them elegantly in the same framework demonstrated its robustness against different views.

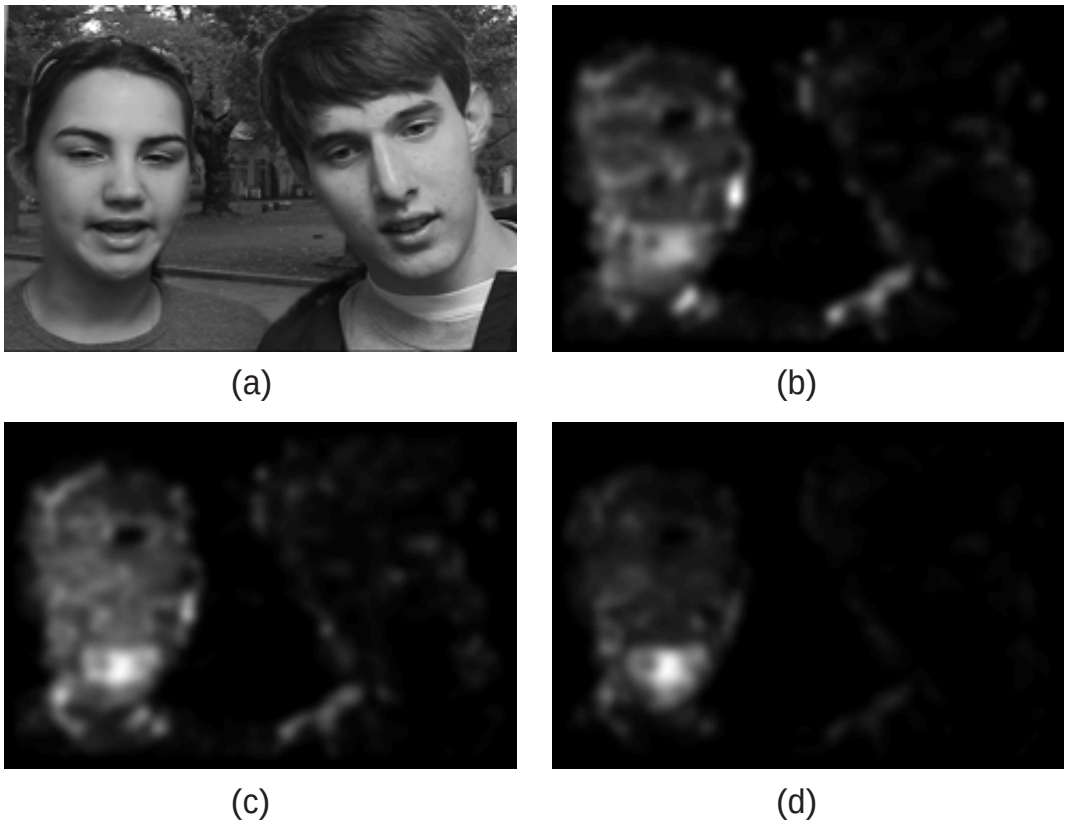



Figure 3.7. Statistical audiovisual correlation using different frame numbers. Figure (a) shows a video frame, and figures (b), (c), and (d) demonstrate the analyzed correlation using $20$, $40$ and $80$ frames. Correlation values are normalized independently for a better visualization.

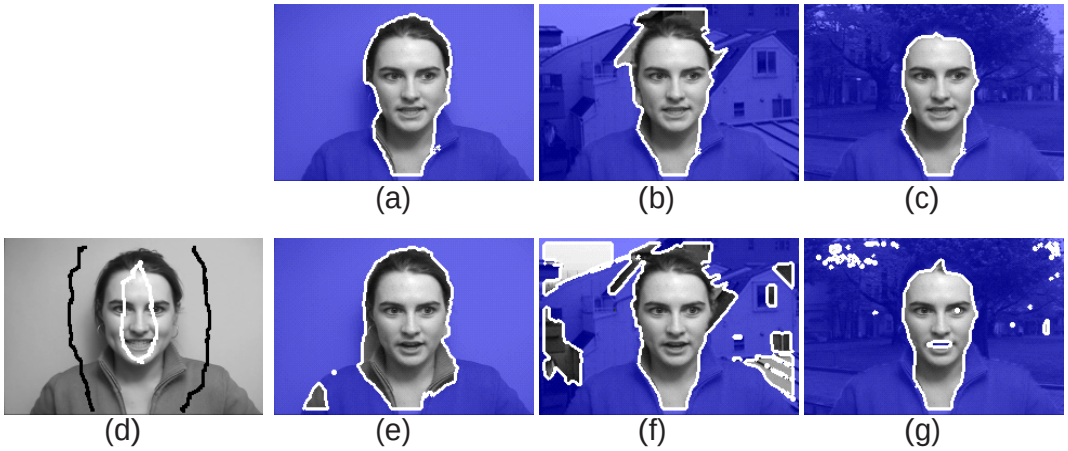



Figure 3.8. Estimated results by the method in (Boykov and Funka-Lea, 2006) and ours. The areas blended with blue represent the region of estimated background. The pixels located at the boundary between speaker and background are colored as white. Figures (a), (b) and (c) show the segmentation results of our method with different non-stationary backgrounds, figure (d) shows our designated segmentation seeds, and figures (e), (f) and (g) show the segmentation results by the method in (Boykov and Funka-Lea, 2006).


Figure 3.9. Segmentations for different views. Figures (a) and (c) show the segmented face region for a frontal view, and figures (b) and (d) show the results for a lateral view.

We tested our method when visual scale and audio gain were changed. The results are shown in Figure 3.10. As our method is adaptive to the scale or gain change, uniform segmentation results were achieved.

We also applied our method to other video clips in CUAVE database, where single or multiple persons were photographed. The backgrounds of some clips were intentionally replaced into complex ones to increase the difficulty of face region detection. Additionally, in the case of multiple persons, we applied to our method to different time windows within which different person was talking. The experimental results are shown in Figure 3.11. In most situations, our method successfully found out the speaker's face region within the time window when it was applied, except Figure 3.11 (c), where our method only segmented out the speaker's mouth region. The reason lies in that the man in (c) intentionally restrained the movements of all his face parts other than his mouth when he was speaking. His speaking manner thus looked a little unnatural, which may come from the tension of being before a camera. As discussed in Section 3.1, our method can localize the speaker's mouth region in such cases.

To give a quantitative evaluation of our detection result, we have manually labeled the face regions for the first frame of four video sequences. The ground truth and the detection rate of our method were shown in Figure 3.12. In most cases, our method can extract the face region with high correct rate.

## 3.5 Conclusions and future works

In this work, we have developed a method to find out the speaker's face region within time windows, which is robust against the changes of view, scale, and background. The main thrust of our idea was to integrate audiovisual correlation analysis into graph cut-based video segmentation. We have shown that our method is capable of finding less fragmented face regions than previous methods for both single and multiple persons under different conditions.

Our current evaluation of audiovisual correlation is sensitive to the noise. Visual

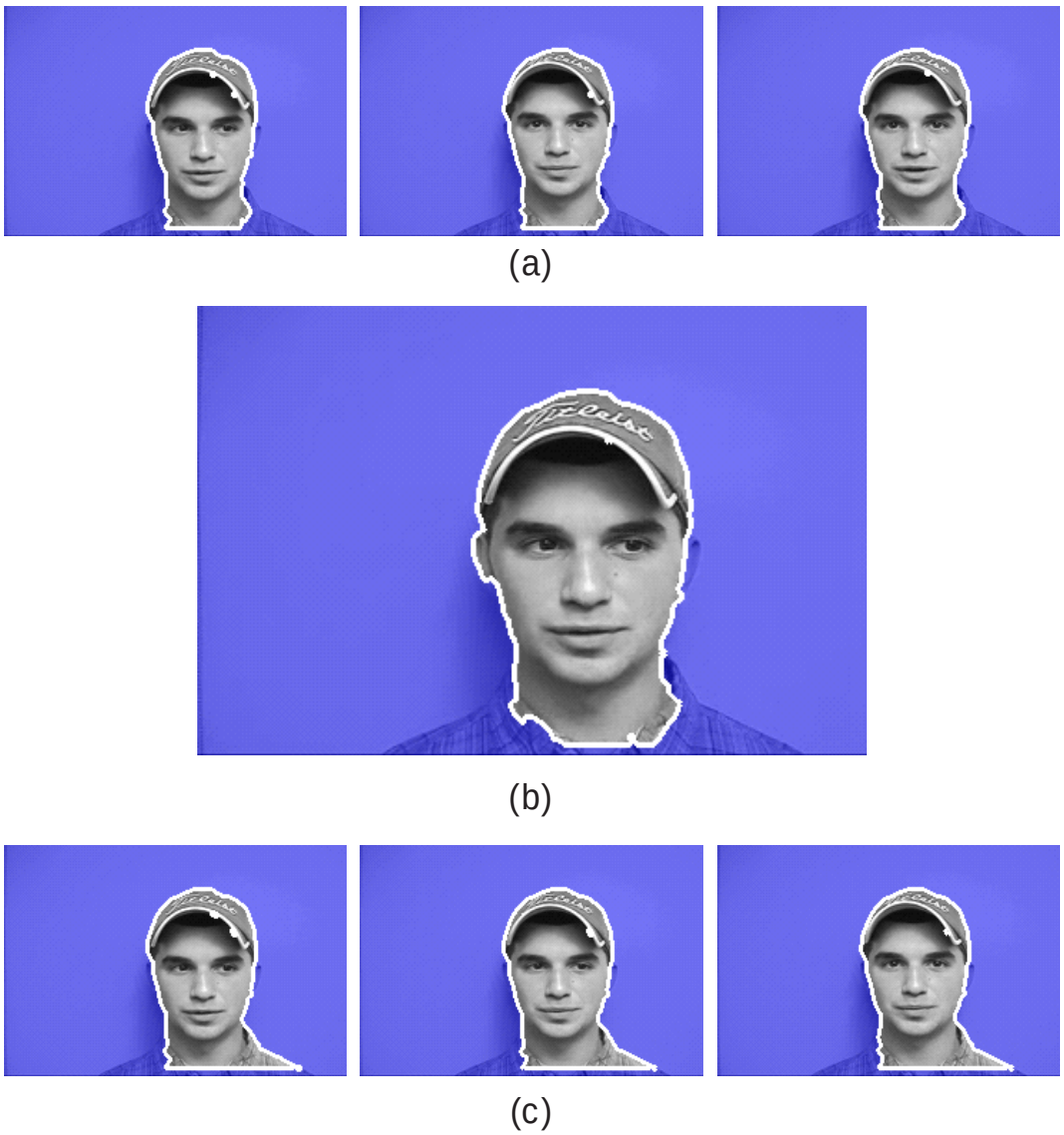



Figure 3.10. Segmentation for different visual scales and audio gains. Figure (a) shows the results with visual resolution 240x160 and original audio data, and figure (b) shows the results when the visual resolution was increased to 360x240, i.e., visual scale was changed to 1.5 times. Figure (c) shows the results when original audio was gained by 3.5dB, i.e., audio magnitude was increased by 1.5 times.

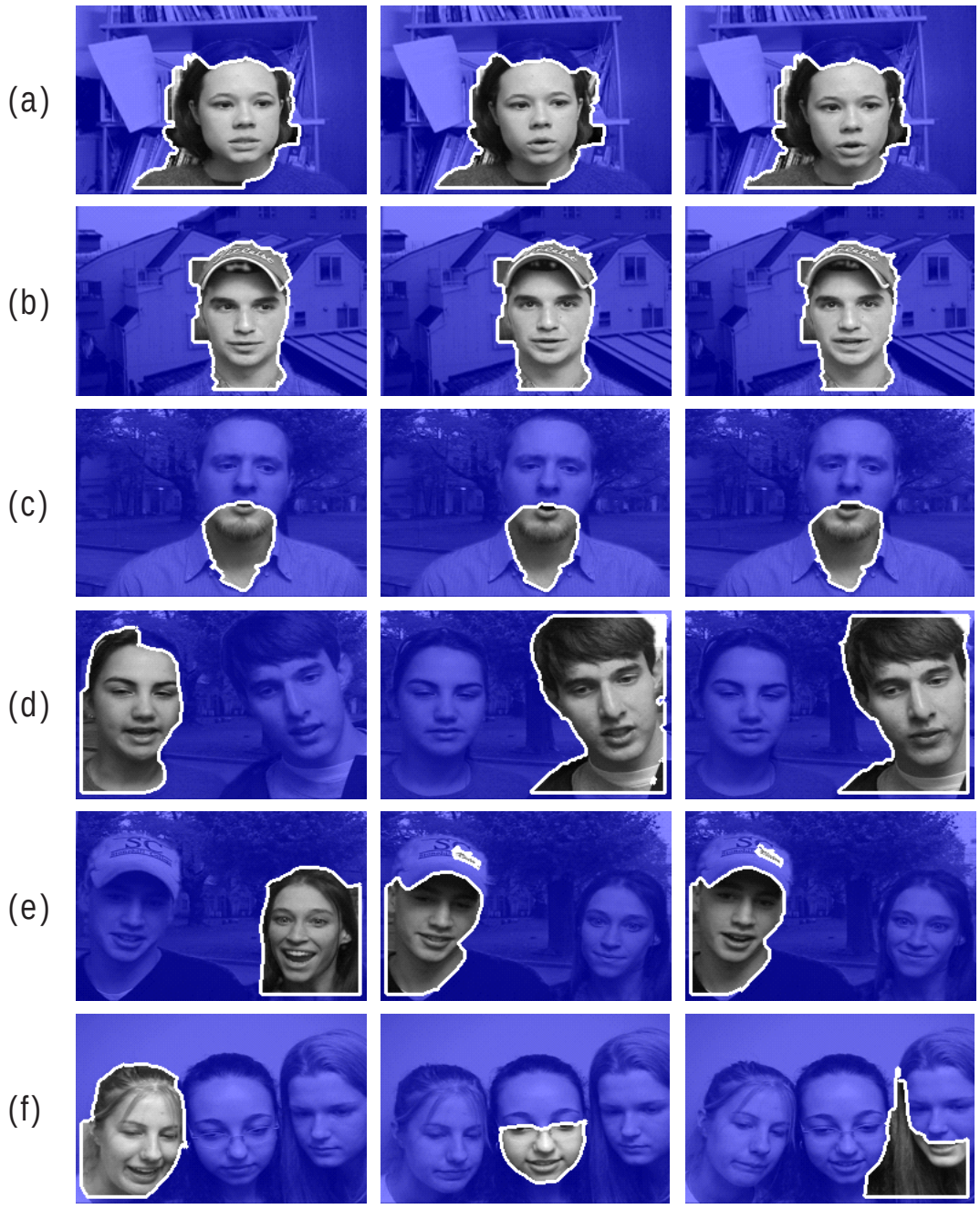



Figure 3.11. The experimental results for other persons. Figures (a), (b) and (c) show the results for a single person, and figures (d), (e) and (f) show the results for multiple persons within different time windows.

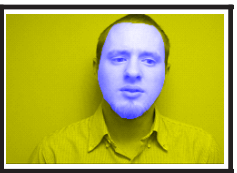
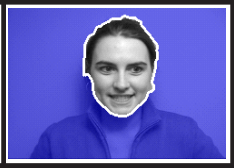
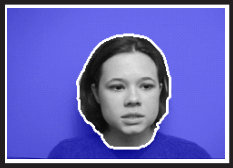
| | | | | |
|---|---|---|---|---|
| Ground truth |  |  |  |  |
| Detected result |  |  |  |  |
| Correct rate | 95.7% | 88.3% | 93.6% | 82.5% |

Figure 3.12. Ground truth and the detection rate of our method. The ground truth in the first row shows the manually labeled face region superimposed over the original image.

noise may yield incorrect optical flow in untextured regions. While Audio noise may disturb the frame energy estimation and make the audiovisual correlation inaccurate. We plan to try other reliable methods to compute optical flow and test other robust audio features, especially the audio features in frequency domain. Additionally, our method in this chapter requires the speaker to stay at relatively the same position in the statistical time span. The localization of a non-stationary speaker is discussed in Chapter 4.

# Chapter 4

# Visual localization of a non-stationary sound source

## 4.1 Introduction

The ability to visually localize sound sources captured by a camera is useful for various applications. For instance, the pan-tilt camera used with a video conferencing system can be controlled to follow a speaker. An interviewee could be automatically overlaid with mosaics to protect privacy. Other applications of sound source localization include surveillance, video analysis, and audio-to-video synchronization adjustment.

Microphone arrays are commonly used for sound source localization. However, the use of such special devices severely limits the applicability of techniques based on this approach. For this reason, much attention has been put on techniques that are based on the audiovisual correlation analysis (Barzelay and Schechner, 2007; Fisher and Darrell, 2004; Hershey and Movellan, 1999; Kidron et al., 2005; Liu and Sato, 2008; Monaci and Vandergheynst, 2006), because the localization can be achieved using only one microphone.

Originating from the discovery that audiovisual correlation lies in synchrony (Driver, 1996), previous works in this field have concentrated on methods for computation-

ally analyzing audiovisual correlation, where different audiovisual features (Barzelay and Schechner, 2007; Fisher and Darrell, 2004; Liu and Sato, 2008; Monaci and Vandergheynst, 2006) and correlation measures (Barzelay and Schechner, 2007; Fisher and Darrell, 2004; Hershey and Movellan, 1999; Kidron et al., 2005; Liu and Sato, 2008; Monaci and Vandergheynst, 2006) have been developed or introduced.

However, all of the existing techniques share a common limitation in that they cannot be used for non-stationary sound sources. When a sound source moves, visual features computed at a fixed position in different frames no longer correspond to the same sound source. The existing techniques choose to ignore this problem by assuming that a sound source is stationary within a given time window.

In this work, we develop a method to correctly analyze the audiovisual correlation for non-stationary sound sources. Within each time window, optimal visual trajectories starting from the pixels in the first frame are independently searched by maximizing the audiovisual correlation between the features extracted from local patches. The visual trajectory that is found in this search is regarded as the best possible motion of that pixel following the movement of the non-stationary sound source. The correlations of the pixels analyzed following their optimal visual trajectories are incorporated into a segmentation technique as (Liu and Sato, 2008) to localize the sound source region in the first visual frame. By shifting the time window, the sound source region in other frames can also be localized.

Two aspects drive our technique. First, we developed a method to efficiently search for optimal visual trajectory by using a beam search with an incremental analysis of the audiovisual correlation. Second, we introduce the inconsistency as an audiovisual feature to robustly analyze the magnitude of acceleration in a local patch.

The rest of this paper is organized as follows. In Section 4.2, we give an overview of our method. In Section 4.3, we introduce the audiovisual feature of the inconsistency. In Section 4.4, we explain the incremental analysis of correlation. In Section 4.5, we demonstrate and discuss the experimental results, and we present our conclusions in Section 4.6.

## 4.2 Outline of our method

First, we compute the visual feature, which evaluates the inconsistency of visual motion in a local Spatio-Temporal patch (ST-patch), and the audio feature, which evaluates the inconsistency of audio energy change in a local Temporal patch (T-patch). The visual and audio features are explained in detail in Section 4.3. Examples of an ST-patch and a T-patch are illustrated in Figure 4.1 (b).

Second, we search for the visual trajectory that maximizes the correlation between the visual and audio features by using a beam search. Given the number of beams $L$ and the search range of a pixel between two consecutive frames $d$, the beam search orders the correlation of the $L(2d+1)^2$ possible visual trajectories in every frame and retains only the $L$ best ones to begin from in the next frame as illustrated in Figure 4.1 (b). All the pixels in the first frame are regarded as starting points and are independently searched. The search continues till the last frame. An example of some search results is shown in Figure 4.1 (c).

Special care needs to be taken for silent frames since the absence of auditory information makes the audiovisual correlation analysis uninformative. Silent intervals can be detected using the method in (Liu and Sato, 2008). In this interval, we stop beam search and connect the path candidates from the last sound frame to the correspondent start pixels in the next sound frame based on a pixel correspondence map; as shown in Figure 4.1 (b). Note that the audiovisual feature is not extracted in the silent intervals. These frames are also not included in the audiovisual correlation analysis.

The correspondence map is computed by accumulating inter-frame optical flow computed by the method in (Lucas and Kanade, 1981). Since optical flow may have errors, the computed correspondence map has errors also. However, as we use this correspondence map to connect path candidates, our method is more robust to the errors in optical flow than the methods that directly use optical flow to track the movement of pixels.

One problem in the flow accumulation is the continuity problem as mentioned in (Chen and Tang, 2007), which addresses the confliction that optical flow computation

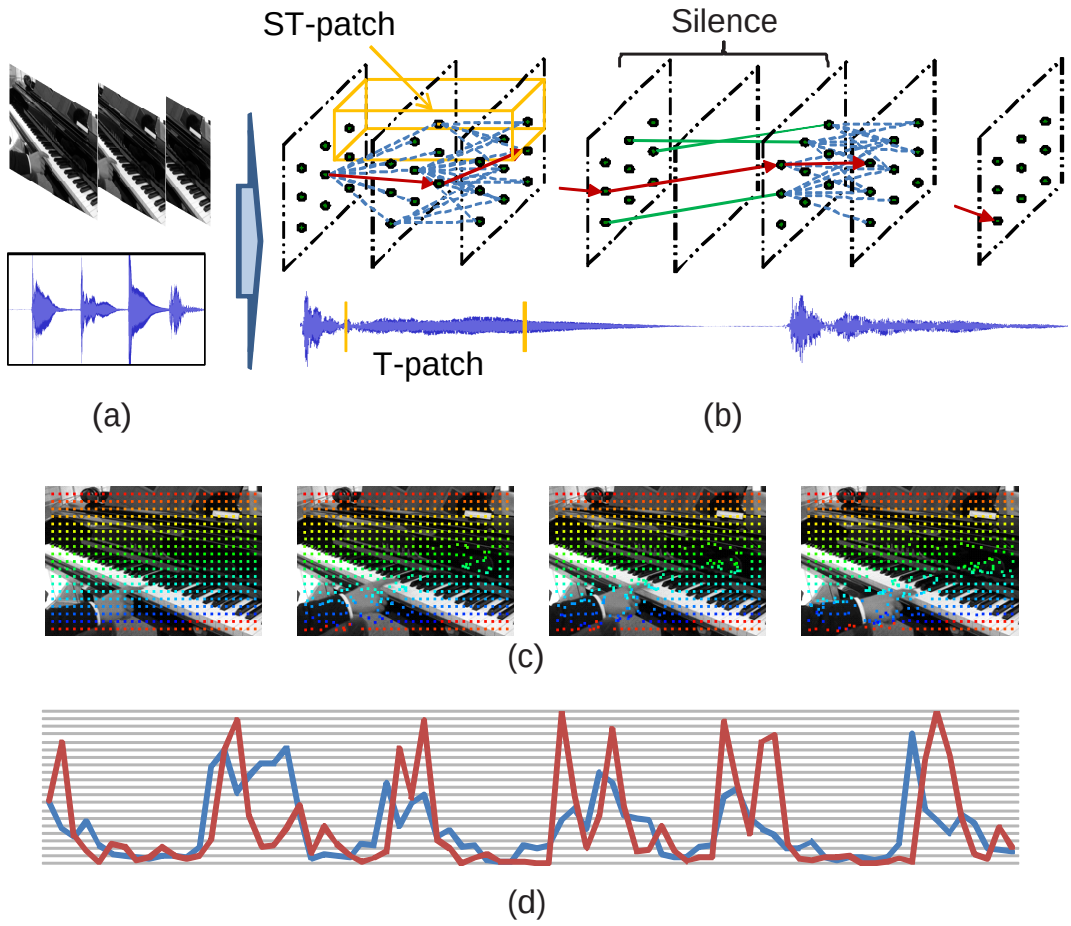



Figure 4.1. Audiovisual correlation maximization. (a) shows original audiovisual data, (b) demonstrates search of visual trajectory, (c) figures optimal visual trajectories starting from different pixels (differently colored), and (d) audio (red) and visual (blue) features following one of the optimal visual trajectories.


Figure 4.2. Accumulated optical flow. Figures (a) and (d) show the begin frames of two silent intervals, figures (b) and (e) show the end frames, and figures (c) and (f) show the accmulated optical flow of these two silent intervals, which become the correspondence maps.

must starts from pixel centers, regardless of the fact that previous pixels may move to sub-pixel positions. We solve this problem by accumulating optical flow in a recursive way. For each pixel $(x, y)$, the flow is accumulated as

$$\vec{O}^{(t+1)}(x,y) = \vec{O}^{(t)}(x,y) + \vec{o}^{(t)}(x + \vec{O}_x^{(t)}(x,y), y + \vec{O}_y^{(t)}(x,y)), \tag{4.1}$$

where $\vec{o}^{(t)}$ and $\vec{O}^{(t)}$ are the computed optical flow between frame $t$ and $t+1$ and the accumulated flow at frame $t$, respectively. The initial value of $\vec{O}^{(1)}$ is set to be zero. If $(x + \vec{O}_x^{(t)}(x,y), y + \vec{O}_y^{(t)}(x,y))$ results in a sub-pixel position, we interpolate $\vec{o}^{(t)}$ with bilinear interpolation.

Finally, we detect the sound source region in the first visual frame using the technique developed in (Liu and Sato, 2008), which evaluates the likelihood of each pixel as the sound source based on the analyzed audiovisual correlation and segments the sound source region out by graph cut.

## 4.3 Audiovisual feature

Differential features, like velocity and acceleration, have recently attracted a lot of interest (Barzelay and Schechner, 2007; Liu and Sato, 2008; Monaci and Vandergheynst, 2006). We believe that audiovisual correlation should be analyzed between the acceleration of the visual motion and that of the audio energy change. For instance, when a human beats a drum, sound is generated when a hand hits the drum. There are sudden changes in both the velocity of the hand and the energy of the sound, which implies the existence of acceleration. Playing pianos, walking, and so on also follow this pattern. Speaking is similar as well. Although the way that a human voice is generated is far more complex than beating a drum, it can still be regarded as a sound jointly modulated by the throat, tongue, teeth, and lips (Rabiner and Juang, 1993). When lip movements are accelerated, the energy of the modulated sound changes simultaneously.

Therefore, we base the audiovisual feature on the evaluation of the magnitude of acceleration, which is determined using the concept of motion inconsistency. As demonstrated in Figure 4.1 (d), high synchrony can be observed with our feature for a hand playing a piano.

### 4.3.1 Visual inconsistency

The usual way to compute the acceleration of visual motion (Barzelay and Schechner, 2007) is to use visual tracking (Lucas and Kanade, 1981) to estimate the visual translation first and then compute the acceleration. However, visual tracking is not stable in low textured areas and usually can be applied to only the featured points (Barzelay and Schechner, 2007).

For this reason, our method base the visual feature on the concept of motion inconsistency computed in a local ST-patch. Motion inconsistency was first introduced by Shechtman and Irani in (Shechtman and Irani, 2007), which measures the degree of the moving direction change in a local ST-patch. The consistent and inconsistent visual motions are demonstrated in Figure 4.3 (a) and (b), respectively. The size of a ST-patch

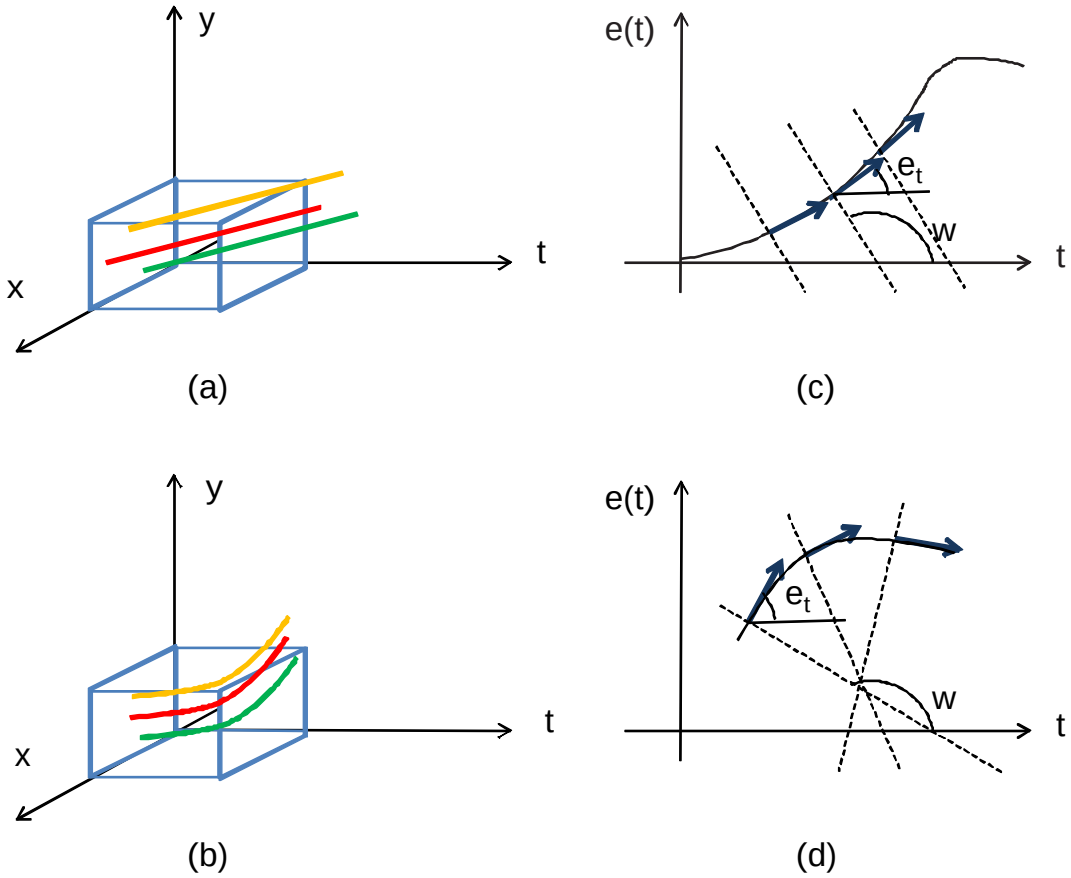



Figure 4.3. Audio and visual inconsistency. Figures (a) and (b) respectively show the consistent and inconsistent visual motions, and figures (c) and (d) respectively show the consistent and inconsistent changes of audio energy.

was recommended to be 7×7×3. The degree of motion inconsistency can be robustly evaluated by looking at the eigen-values of a gradient matrix M defined as

$$\mathbf{M} = \begin{pmatrix} \sum I_x^2 & \sum I_x I_y & \sum I_x I_t \\ \sum I_y I_x & \sum I_y^2 & \sum I_y I_t \\ \sum I_t I_x & \sum I_t I_y & \sum I_t^2 \end{pmatrix}, \tag{4.2}$$

where $I_x$, $I_y$, and $I_t$ respectively denote the partial derivative $\partial I/\partial x$, $\partial I/\partial y$, and $\partial I/\partial t$ of the intensity at each pixel. The motion inconsistency is then computed by

$$\Delta r = \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^{\diamondsuit} \cdot \lambda_2^{\diamondsuit}}, \tag{4.3}$$

where $\lambda_2$ and $\lambda_3$ are the second and the third eigenvalue of $\mathbf{M}$. $\lambda_1^{\diamondsuit}$ and $\lambda_2^{\diamondsuit}$ are the first and the second eigenvalue of the top left $2 \times 2$ submatrix $\mathbf{M}^{\diamondsuit}$ of $\mathbf{M}$. The computation can be speeded up as

$$\Delta \hat{r} = \frac{det(\mathbf{M})}{det(\mathbf{M}^{\diamondsuit}) \cdot ||\mathbf{M}||_F}, \tag{4.4}$$

where $||\mathbf{M}||_F = \sqrt{\sum M(i,j)^2}$ is the Frobenius norm of the matrix $\mathbf{M}$.

It is important to note that motion inconsistency is closely related to the magnitude of acceleration of an object. If the direction of movement is altered in a ST-patch, an inconsistent motion will be detected. The more the direction of movement is altered, the higher the degree of inconsistency. Accordingly, motion inconsistency can be used for measuring the magnitude of acceleration.

Therefore, we define the visual feature $v(x, y, t)$ as the degree of motion inconsistency computed in a ST-patch centered at $(x, y, t)$. The computation of motion inconsistency is the same as that of the motion consistency in (Shechtman and Irani, 2007).

Note that, as also mentioned in (Shechtman and Irani, 2007), there is another possibility that leads to inconsistent motion. When an ST-patch is located at the boundary of two different motion fields, it has inconsistent motion. Shechtman and Irani claimed that this is negligible considering the number of total pixels (Shechtman and Irani, 2007). Furthermore, it casts fewer effects on our system as the boundaries always demonstrate a high inconsistency and are less correlated with the audio. We also disregard this point just as (Shechtman and Irani, 2007).

In implementation, we first smooth visual images by a $5 \times 5$ Gaussian filter whose $\sigma$ is 0.8. Visual inconsistency is then analyzed with ST-patches whose size is $7 \times 7 \times 3$. Note that $\mathbf{M}$ is computed not as Equation (4.2), but by a weighted form, where weights are Gaussian with $\sigma_{space} = 1.5$ and $\sigma_{time} = 0.8$. Suppose that this Gaussian weight is $g_i$, which is determined by both spatial and temporal distance from the current pixel to the center of the current ST-patch. $\mathbf{M}$ will be computed as

$$\mathbf{M} = \begin{pmatrix} \sum_i g_i^2 I_{xi}^2 & \sum_i g_i^2 I_{xi} I_{yi} & \sum_i g_i^2 I_{xi} I_{ti} \\ \sum_i g_i^2 I_{yi} I_{xi} & \sum_i g_i^2 I_{yi}^2 & \sum_i g_i^2 I_{yi} I_{ti} \\ \sum_i g_i^2 I_{ti} I_{xi} & \sum_i g_i^2 I_{ti} I_{yi} & \sum_i g_i^2 I_{ti}^2 \end{pmatrix} \tag{4.5}$$

### 4.3.2 Audio inconsistency

Similarly to the visual feature, the audio feature is defined based on the inconsistency of audio signals in a local T-patch. The consistent and inconsistent changes of audio energy are demonstrated in Figure 4.3 (c) and (d), respectively.

First, we compute the audio energy $e(t)$ at each frame $t$. As audio usually has a much higher sampling frequency than video, we divide audio samples into frames. The frame duration $T_a$ is set to be the same as the $T_v$ of each visual frame. Audio energy $e(t)$ in each frame is computed using the method introduced in Chapter 3.

Second, we compute the inconsistency of the audio energy change in a local temporal patch. As $e(t)$ changes in temporal domain only, we can only compute the temporal derivative $e_t = de/dt$ that represents the slope of a tangent line of $e(t)$ at $t$. Suppose the slop of the line perpendicular to this tangent line to be $w$. Following Cartesian geometry we have

$$1 + we_t = 0. \tag{4.6}$$

If $e(t)$ changes in a same tendency in the local temporal patch, $w$ should be the same; as demonstrated in Figure 4.3 (c).

Packing Equation (4.6) for the derivatives in the temporal patch results in

$$\mathbf{P}\begin{pmatrix} 1 \\ w \end{pmatrix} = 0, \tag{4.7}$$

where $\mathbf{P}$ is defined as

$$\mathbf{P} = \begin{pmatrix} 1 & e_{t1} \\ 1 & e_{t2} \\ \vdots & \vdots \\ 1 & e_{tn} \end{pmatrix}. \tag{4.8}$$

Multiplying both sides of Equation (4.7) by $\mathbf{P}^T$ yields the condition that consistent change of audio energy must satisfy, which is

$$\mathbf{Q}\begin{pmatrix} 1 \\ w \end{pmatrix} = 0, \tag{4.9}$$

where $\mathbf{Q}$ satisfies that

$$\mathbf{Q} = \begin{pmatrix} \sum 1 & \sum e_t \\ \sum e_t & \sum e_t^2 \end{pmatrix}. \tag{4.10}$$

As analyzed in (Shechtman and Irani, 2007), the condition that $e(t)$ changes consistently is that $\mathbf{Q}$ is a rank deficient matrix. Consequently, the more $\mathbf{Q}$ becomes a rank full matrix, the higher the inconsistency of the change of audio energy in this T-patch. The degree of inconsistency in the local temporal patch can be evaluated based on the eigenvalues of $\mathbf{Q}$ (Shechtman and Irani, 2007).

The audio feature $a(t)$, which is defined as the degree of inconsistency, can thus be computed by

$$a(t) = \frac{\lambda_2}{\lambda_1^{\diamondsuit}} \tag{4.11}$$

where $\lambda_2$ is the second eigenvalue of the $\mathbf{M}$. $\lambda_1^{\diamondsuit}$ is the top left element of $\mathbf{Q}$.

Similarly as the visual feature, $\mathbf{Q}$ is computed not by Equation (4.10), but by a weighted form in the implementation. The weights are Gaussian with $\sigma = 0.8$ for the T-patch whose size is three. Suppose that this weight is $g_i$, which is determined by the

temporal distance from the current frame to the center of the current T-patch. $\mathbf{Q}$ will be computed as

$$\mathbf{Q} = \begin{pmatrix} \sum_i g_i^2 & \sum_i g_i^2 e_{ti} \\ \sum_i g_i^2 e_{ti} & \sum_i g_i^2 e_{ti}^2 \end{pmatrix} \tag{4.12}$$

### 4.3.3 Feature quantization

In order to analyze the audiovisual correlation, the audio and visual features introduced above are furthermore quantized. We explain this quantization process here.

We first show that the audio and visual features are normalized. As for the visual feature, inequalities $0 \leq \Delta r \leq 1$ and $0 \leq \Delta r \leq 1$ have been shown in (Shechtman and Irani, 2007). Based on them, we can get $0 \leq v(x, y, t) = \Delta\hat{r} \leq 1$, i.e., the visual feature is normalized. As for the audio feature, since $\mathbf{Q}$ is a symmetric definite matrix, from linear algebra we know that $0 \leq \lambda_2 \leq \lambda_1^{\diamondsuit} \leq \lambda_1$. Substituting Equation (4.11) into this inequality, we have $0 \leq a(t) \leq 1$, i.e., the audio feature is normalized.

For the normalized values of the audio and visual features, we quantize them uniformly in their range that is between zero and one. The number of quantization stages is denoted as $C$, which is fixed to be 20 in this work. We have tried other $C$ values but observed only minor changes when analyzing audiovisual correlation.

## 4.4 Incremental analysis of audiovisual correlation

It is possible to compute mutual information using its definition (Shannon, 1951) as

$$MI(a; v) = \frac{1}{N} \sum_i \sum_j h_{av}(i, j) \log \frac{N h_{av}(i, j)}{h_a(i) h_v(j)}, \tag{4.13}$$

where $h_{av}$ is the joint histogram of the quantized audio and visual feature. $h_a$ and $h_v$ are their marginal histograms, respectively. $N$ is the number of the samples. Although this computation just requires a sum over the histograms, it still calls for a lot of computation time in our case, since the correlation is evaluated millions of times in the maximization process.

In this work, we develop a method to compute mutual information incrementally in a beam search, which significantly speeds up the process. We would like to mention that this incremental computation is general and can be used in other situations where mutual information needs to be computed in multiple stages.

We explain the incremental computation of mutual information using entropy. Since mutual information can be divided into the sum of entropies following

$$MI(a;v) = H(a) + H(v) - H(av), \tag{4.14}$$

conclusions here can be easily applied to mutual information.

The entropy of a discrete random variable $z$ can be computed based on its histogram $h(z)$ by definition (Shannon, 1951) as

$$H(z) = -\frac{1}{N}\sum_i h_z(i)\log\frac{h_z(i)}{N}. \tag{4.15}$$

Suppose that we have the entropy computed in frame $k$ as $H^{(k)}(z)$. In frame $k+1$, a new sample is added to histogram $h_z^{(k)}$, which results in a new entropy $H^{(k+1)}(z)$. Following the equation deductions, we can represent $H^{(k+1)}(z)$ based on the known $H^{(k)}(z)$ as

$$H^{(k+1)}(z) = \frac{1}{k+1}\left(kH^{(k)}(z) - \log\frac{(1+n)^{1+n}}{n^n}\right) + C(k), \tag{4.16}$$

where $n = h_z^{(k)}(i^{(k+1)})$ is the number of the samples in the bin $i^{(k+1)}$ of $h_z^{(k)}$, $i^{(k+1)}$ is the index of the sample added in frame $k+1$, and $C(k) = \frac{k}{k+1}\log k - \log\frac{k}{k+1}$ is a coefficient decided by $k$. The correctness of Equation (4.16) is shonw in Appendix B.

We can thus adopt Equation (4.16) to incrementally compute entropy. The computation $H^{(k+1)}(z)$ by using Equation (4.16) is significantly faster than the computation by definition in Equation (4.15) because, based on the value of $H^{(k)}(z)$, the computation of incremental quantity $\log\frac{(1+n)^{1+n}}{n^n}$ requires the access of only one histogram bin. However, the computation of Equation (4.15) have to access all the histogram bins. As the histogram has 400 bins in our work, this speed-up is 400 times.

There is an undefined problem in the computation of Equation (4.16). When $n = 0$, the incremental quantity is undefined. However, if $n$ is a continuous variable, the limit

exists when $n$ approaches zero. Since we have

$$\lim_{n \to 0} \frac{(1+n)^{1+n}}{n^n} = 1, \tag{4.17}$$

whose correctness is shown in Appendix C, we can adopt this limit to define the value when $n = 0$ as $\log 1 = 0$.

We can substitue Equation (4.16) into Equation (4.14) to obtain the equation to incrementally compute mutual information. However, we have a simpler way to achieve this goal. Since $H(a)$ is invariant to the search of visual trajectories, minimizing

$$H(v|a) = H(av) - H(v) \tag{4.18}$$

is equal to maximizing $MI(a; v)$. Substituting Equation (4.16) into Equation (4.18), we can obtain the equation to incrementally compute $H(v|a)$ as

$$H^{(k+1)}(v|a) = \frac{1}{k+1} \left( kH^{(k)}(v|a) + \log \frac{\frac{(1+n)^{1+n}}{n^n}}{\frac{(1+m)^{1+m}}{m^m}} \right), \tag{4.19}$$

where $n = h_v^{(k)}(i^{(k+1)})$, and $m = h_{av}^{(k)}(i^{(k+1)}, j^{(k+1)})$. $j^{(k+1)}$ is the index of the quantized audio feature in stage $t+1$. The coefficient $C(k)$ in Equation (4.16) is cancelled in the subtraction. As the computation of $H(v|a)$ is simpler than the one of $MI(a; v)$, we in fact minimize $H(v|a)$ in the beam search instead of maximizing $MI(a; v)$.

The undefined problem exists also in Equation (4.19) when $n = 0$ and $m = 0$. We again use the limit in Equation (4.17) to define their values.

## 4.5 Experimental results

In this section we present the experimental results from using our method. We used the same parameters for all the experiments. The length of the time window was three seconds. The search range and candidate number in the beam search were $d = 1$ and $L = 10$, respectively. The video was sampled at 30 fps and converted to monochrome at a resolution of $240 \times 160$, while the audio was sampled at 44.1 KHz.

The localization results of non-stationary sound sources were demonstrated in Figure 4.4. The two clips in Figure 4.4 captured both a sound source (a hand or a walking man) and an ambiguous moving object (a rotating cover or a man riding a bicycle). Both sound sources were successfully localized. Interestingly, the positions that corresponded to the obscure reflections of the hand also demonstrated fragmentally high audiovisual correlation in Figure 4.4 (b). This is reasonable since both movements were synchronous with the audio change. Similar phenomenon was also discussed in (Kidron et al., 2005).

A speaker is an important class of sound sources. We used a CUAVE database (Patterson et al., 2002) to test the performance of our method for speaker localization.

The localization results of non-stationary speakers were demonstrated in Figure 4.5. Our method successfully found the facial region of the speaker. As a comparison, we compared this method with the one we developed in Chapter 3, which was designed to localize stationary speakers. The results are shown in Figure 4.5 (c) and (e), where we can observe how the analysis of audiovisual correlation failed for a non-stationary speaker and had the wrong localization.

We can detect the sound source at different times by applying our method to different time windows. The speaker localization results from multiple persons were demonstrated in Figure 4.6. Our method localized the current speaker.

Our method has to make a tradeoff between the tolerable moving speed determined by the search range and the computation time. To accommodate for fast movement, we need to set a large $d$ and $L$. When $d = 1$ and $L = 10$, our method failed to detect a fast moving sound source shown in Figure 4.7 (b). The detection was improved by increasing $d$ and $L$; as shown in Figure 4.7 (c). Yet, the computation time rises when $d$ and especially $L$ are increased. If we have many sound frames that need to be searched by the beam search, the rising speed is fast. For example, Figure 4.4 (a) has 69 sound frames in the 90-frame time window, which is the largest number in all the experimental data. The computation on our desktop, which has an Intel Core2 2.6 G CPU and a 3 G memory, took 87 seconds when $d = 1$ and $L = 10$, 195 seconds when $d = 2$ and $L = 10$, and 457 seconds when $d = 2$ and $L = 20$.

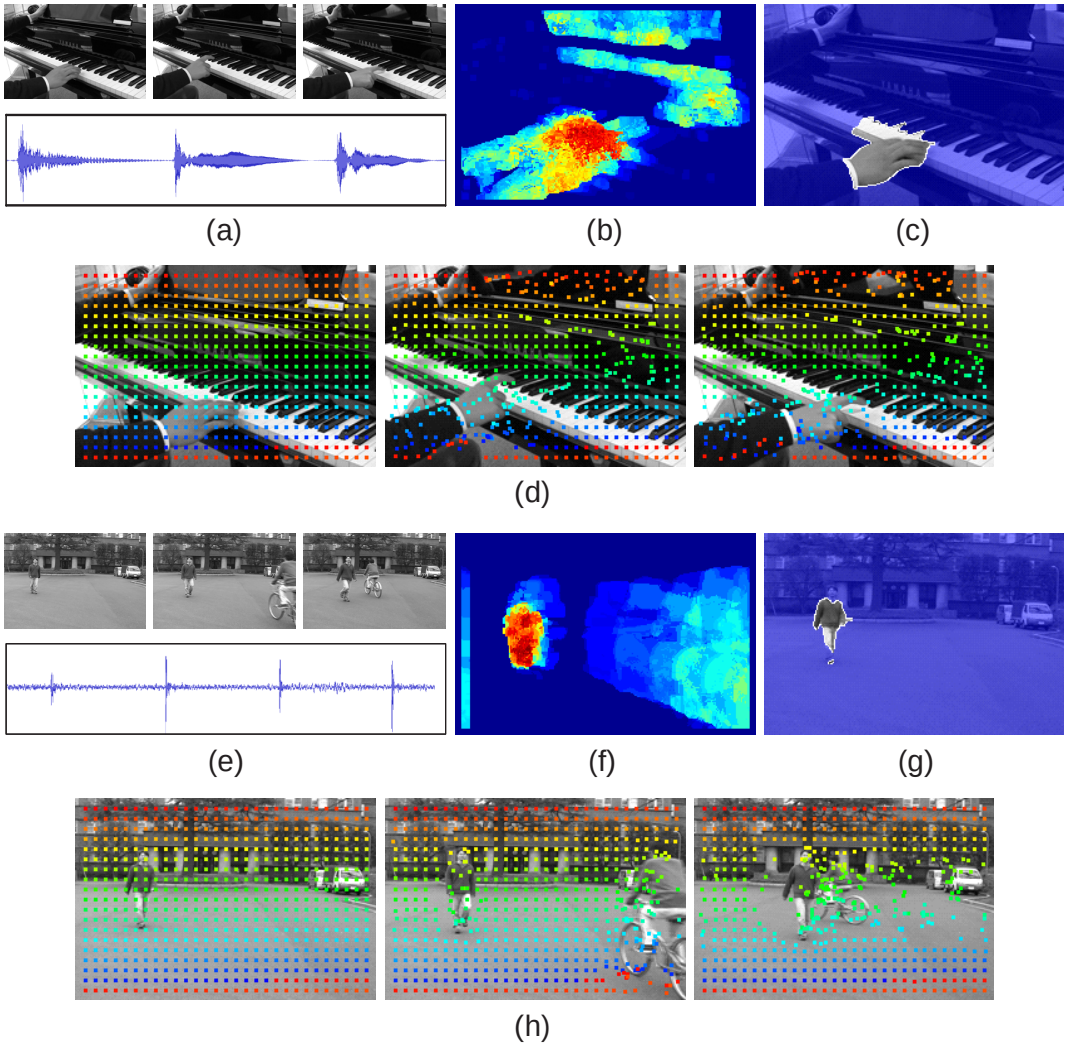



Figure 4.4. Localizations of non-stationary sound sources. Figures (a) and (d) show the original data. Figures (b) and (e) visualize the analyzed audiovisual correlation with jet color map. The redder a pixel, the higher its correlation. Figures (c) and (f) show the sound source region localized.

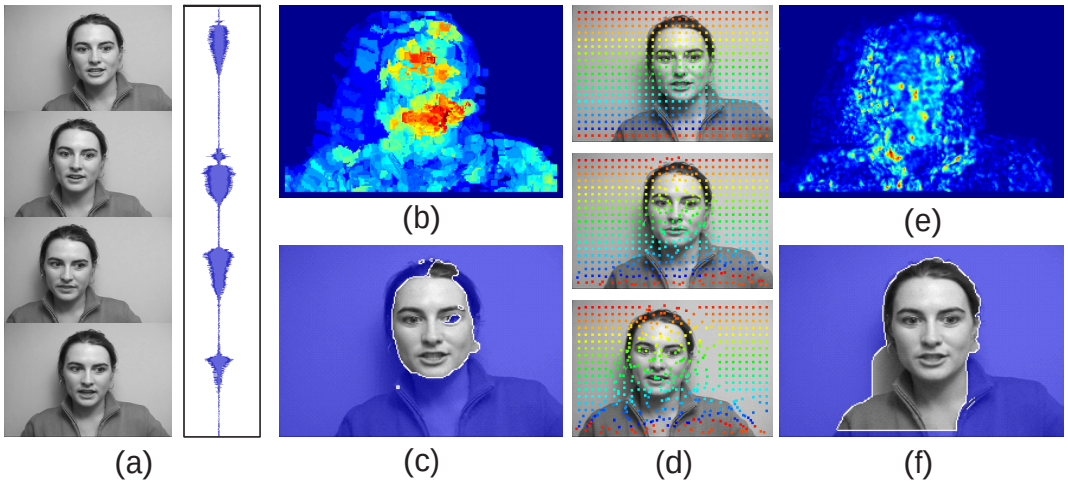



Figure 4.5. Localization of non-stationary speaker. Figure (a) shows the original audiovisual data. Figures (b) and (d) show the analyzed audiovisual correlation and localization results of our method, respectively. Figures (c) and (e) show the results when using the method in Chapter 3.

## 4.6 Conclusion and future work

We have developed a method to visually localize non-stationary sound sources by searching for the movements that maximize the audiovisual correlation. The search is efficiently conducted by using a beam search with the incremental analysis of the audiovisual correlation. We have also introduced inconsistency as an audiovisual feature. Our method is capable of localizing different kinds of sound sources for different time windows.

There is a tradeoff in our current method between the computation time and the tolerable motion speed, as discussed in Section 4.5. We are considering using an image pyramid and a coarse-to-fine policy to resolve this problem.

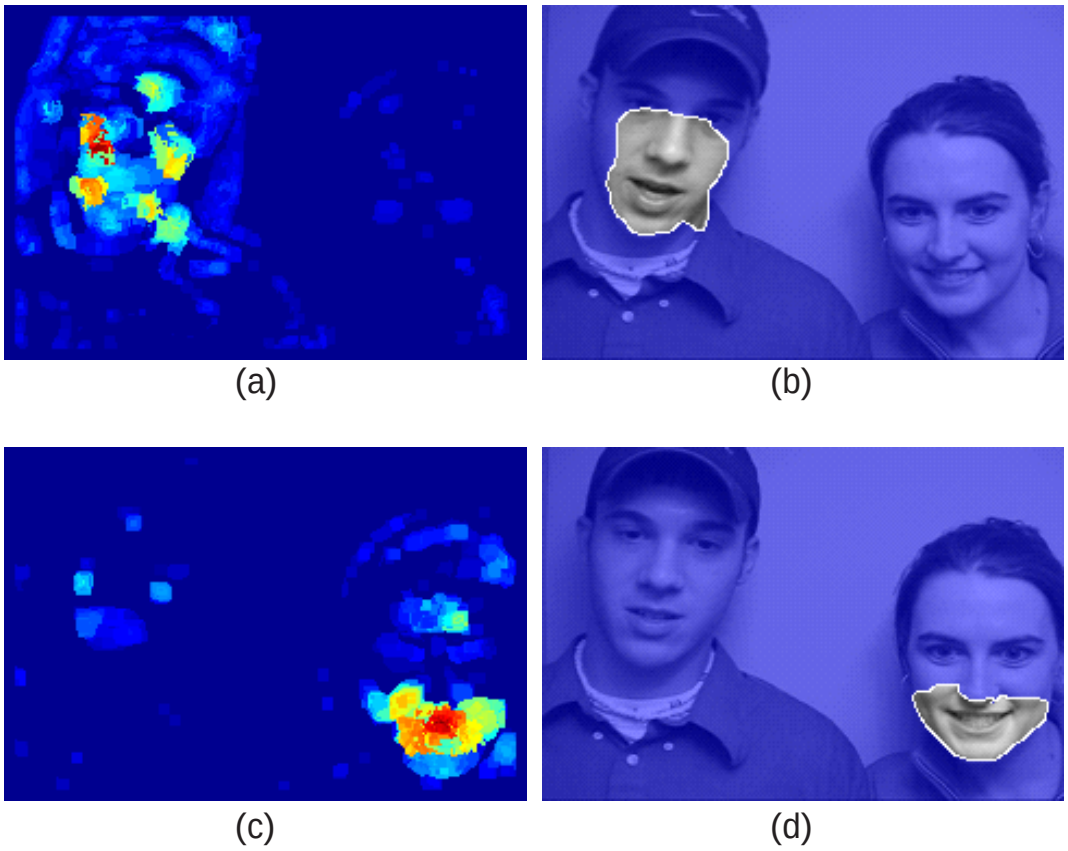
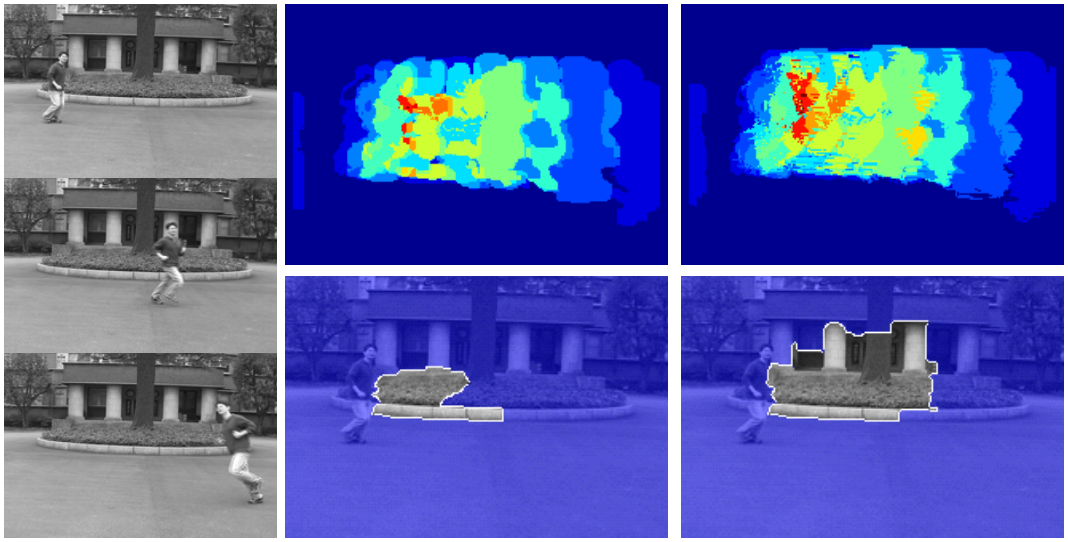



Figure 4.6. Speaker localization of different time windows. Figures (a) and (c) show the analyzed audiovisual correlation, and (b) and (d) show the localization results.



(a) (b) (c)

Figure 4.7. Localization of a fast moving sound source. Figure (a) shows the original data, figure (b) shows the result when $d = 1$ and $L = 10$, and figure (c) shows the result when $d = 3$ and $L = 20$.

# Chapter 5

# Recovery of audio-to-video synchronization

## 5.1 Introduction

Audio-to-video synchronization (AV-sync) is important for human sensing of AV perceptual cues. Reeves and Voelker discovered that, if AV-sync drifts, humans evaluate video content much more negatively (Reeves and Voelker, 1993). When audio preceded video by five video frames, satisfaction by viewers degraded about 84% (Reeves and Voelker, 1993). Television standards specify that audio should never be ahead of video by more than 15 ms, and should never lag behind video by more than 45 ms (ATSC group, 2003).

However, AV-sync may drift in real situations for many reasons, such as different processing times between video and audio, inconsistent network-transfer delays, and drift accumulating in concatenated processing stages. Previous efforts have mainly concentrated on avoiding drift by creating specifications for both hardware and software to maintain AV-sync (ATSC group, 2003). For example, video cameras record timecodes to maintain AV-sync for read out. MPEG-2 codecs print Presentation Time Stamps (PTS) into the data.

Few attempts have been made to recover AV-sync when drift has occurred. As video processing often includes several stages, a lack of capability for recovery means that all stages must be carefully designed to maintain AV-sync. Even though each stage only causes minor drift, drift can still accumulate into a form that is more obvious. Consequently, high-quality videos like commercial films always include a final stage for adjusting AV-sync, where producers employ a special device called an audio synchronizer to enable AV-sync to be recovered through dedicated efforts.

We developed a method of recovering AV-sync that had drifted in video clips that only required minor human interactions. Our method is not only robust against changes in the audiovisual scale, where changes in the audio scale mean that the audio signal is transformed by different gains, but it is also independent of language. The main thrust of our method was to detect the state of AV-sync by analyzing the cross-modality correlations between audio and video.

Our method utilizes audiovisual correlation analysis in a different way to detection of speakers. We assumed that the speaker was known. As synchrony is used to evaluate audiovisual correlation (Driver, 1996; Hershey and Movellan, 1999), our method finds the offset between audio and visual channels based on the assumption that the audiovisual correlation inside the speaker region reaches its maximum at the state of AV-sync. Our method works as follows. Given a video clip to recover AV-sync, we first ask a user to specify a rough time window during which a person is speaking. The speaker is supposed to be stationary in the time window, as has been required in all previous work (Fisher and Darrell, 2004; Hershey and Movellan, 1999; Kidron et al., 2005; Smaragdis and Casey, 2003). We then shift audio to make different drift hypotheses and analyze the average audiovisual correlation inside the speaker region identified by a face detection technique (Viola and Jones, 2004). Surrounding the optimum drift that maximizes the average correlation, we furthermore refine drift based on the correlation value to subframe accuracy and adopt this value to recover AV-sync.

The ability to analyze audiovisual correlation accurately is of great importance to precisely recover AV-sync. We developed a novel method of analytically computing this correlation making no assumptions on the distribution, which is robust against changes

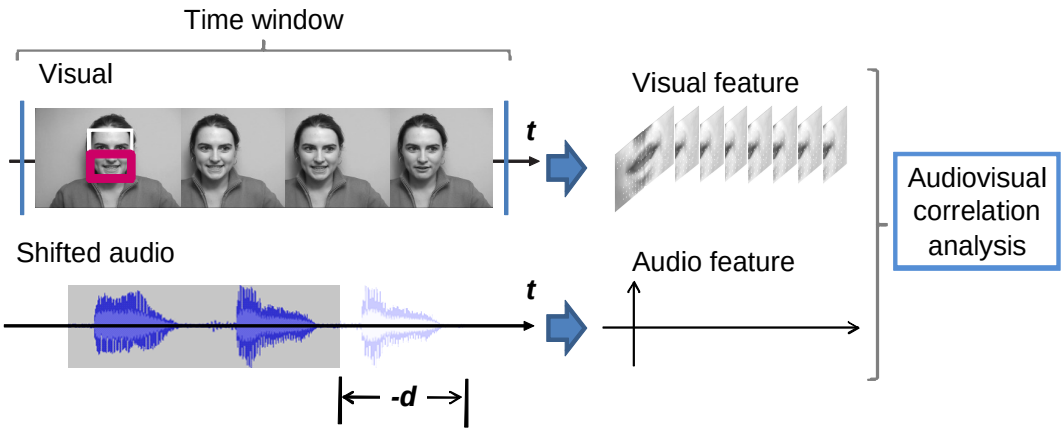



Figure 5.1. Process to detect drift.

in the audiovisual scale. The key to our method is using kernel density estimation with adaptive bandwidth and quadratic mutual information.

## 5.2 AV-sync recovery by analysis of audiovisual correlations

The first thing in recovering drifted AV-sync is to find how much it has drifted. We assumed that this drift value would be constant in video clips in recovering AV-sync. Taking into consideration the reasons for drift, we found that this assumption held for most situations. For cases where this assumption did not hold, such as where drift was caused by sudden network-transfer delays, they were considered to be piecewise constants. The assumption still held if we divided these video clips into segments and processed them separately.

The process for finding drift is given in Figure 5.1. We first ask a user to specify the time window where there is a stationary speaker. The speaker's face must be photographed in the video with his or her speech recorded in the audio. Only the data within this time window are used.

We then detect the speaker region using a face detection technique (Viola and Jones, 2004). To concentrate on mouth movements, we take the lower half of the detected face as the speaker region. When multiple faces are detected, we select the one that has the highest audiovisual correlation. If face detection fails, we ask the user to manually specify the speaker region. As the speaker is supposed to be stationary, the speaker region is determined using only the first image frame in the time window.

We search the optimum drift based on analysis of audiovisual correlation. First, we quantize the drift value into integral multiples of the video frame duration $T_v$, i.e., $d = \{-L, \cdots, -1, 0, 1, \cdots M\}$. $[-L, M]$ represents a pre-defined search range. Positive drift values mean how much the audio lags behind the video. Negative ones indicate how much the audio precedes the video. For each $d$, we shift the audio data temporally by $-d$ and compute the average audiovisual correlation inside the speaker region. The optimum value, $d^*$, which has the maximum average correlation is regarded as the coarse drift found, i.e.,

$$d^* = \arg\min_d C(d) \tag{5.1}$$

where $C(\cdot)$ represents the average audiovisual correlation analyzed with respect to $d$.

Second, we refine $d^*$ to sub-frame accuracy. We fit a parabola to the analyzed correlation values around $d^*$ and take its peak as the final detected audiovisual drift, $d^*_{av}$, which is computed by

$$d^*_{av} = T_v \cdot (d^* + \frac{0.5 \cdot (C(d^* - 1) - C(d^* + 1))}{C(d^* - 1) - 2C(d^*) + C(d^* + 1)}. \tag{5.2}$$

Since we assumed the drift would be constant in the current video clip, temporally shifting the audio by $-d^*_{av}$ will set the clip back to the state of AV-sync.

## 5.3 Analysis of audiovisual correlations

This section describes the computation of $C(d)$ in detail. We shift audio by $-d$ and extract audio feature $a_t$ and visual feature $v_t(x, y)$ from the $N$-frame audiovisual data. Note that visual feature $v$ also changes spatially for different $(x, y)$. The number of

frames $N$ should not exceed the maximum possible number of frames within the time window specified by the user. Based on $a_t$ and $v_t(x, y)$, we compute the audiovisual correlations at all positions $(x, y)$ in the speaker region using quadratic mutual information with kernel density estimation. Finally $C(d)$ is computed by averaging the computed correlation inside the speaker region. Below, we first introduce the audiovisual features and then correlation analyses.

### 5.3.1 Audiovisual feature

We adopt the same method as introduced in Chapter 3 to extract the audio and visual feature. To make this chapter self-containable, we briefly explain the extraction of audio and visual features below.

**Audio features.** As audio is usually sampled at much higher frequencies than video, we first divided the audio samples into frames. The frame duration, $T_a$, is set to be the same as the visual frame duration, $T_v$. An overlap of duration of $T_a/2$ between each pair of two successive frames is set. We also multiply a Hamming window to the audio signal in each frame to reduce side effects.

The audio feature is defined to be the differential energy between the current and next frames, i.e., $a_t = e(t+1) - e(t)$. The energy, $e(t)$, of all audio frames is computed by

$$e(t) = \log\left(\frac{1}{M}\sum_{m=1}^{M}(w(m)s(t,m))^2\right), \tag{5.3}$$

where $s(t, m)$ represents the audio samples in frame $t$ and the two overlapping components from the neighboring frames. $w(m)$ is the weight of the Hamming window. $M$ is the number of audio samples in the duration of $2T_a$.

We ensure there is speech in all frames. This is done by checking whether or not audio energy $e(t)$ is larger than a pre-defined threshold. Frames failing this test are regarded as silent and dropped, together with their corresponding visual frames. Only frames passing this test are buffered till the frame number reached a pre-defined value. If we discuss $N$ audiovisual frames in this work, it refers to frames that are buffered.

**Visual features.** Optical flow is used as the visual feature in our method. In particular, we only take the vertical element of optical flow, considering that most speaking actions moved vertically. Visual feature $v_t(x, y)$ is thus the vertical optical flow extracted at $(x, y)$ between frames $t$ and $t + 1$, which is computed by the Lucas-Kanade method in (Lucas and Kanade, 1981). Since optical flows cannot be estimated stably in areas with less texture, we verify the variation of pixel intensities inside each window where we compute optical flows. If these are below a threshold, we set the flow value to zero.

### 5.3.2 Computation of correlations

Correlations is computed independently at all image coordinates $(x, y)$ inside the speaker region. We first estimate the joint probability density function (pdf) between audio and visual features using kernel density estimation, and compute audiovisual correlation using quadratic mutual information. Yet, the two steps are merged in the implementation, i.e., the correlation is computed directly from the audiovisual-feature samples, because the computation can be done analytically. The method we use to analyze the audiovisual correlation is in fact same as the one we introduced in Chapter 3. Below we briefly introduce this analysis process.

We adopt quadratic mutual information to compute their correlations, which was first proposed in (Xu et al., 1998). The same as MI, quadratic mutual information indicates the amount of information that one random variable conveys about another. For audio feature $a$ and visual feature $v$ at each position $(x, y)$, quadratic mutual information is computed by definition as

$$QMI(a; v) = \log \frac{\iint p^2(a, v)dadv \iint p^2(a)p^2(v)dadv}{(\iint p(a, v)p(a)p(v)dadv)^2}. \tag{5.4}$$

Probability density function $p(\mathbf{z})$ of a $n$-dimensional random variable $\mathbf{z}$ is estimated by kernel density estimation (Parzen, 1962). Given $N$ data points $\mathbf{z}_i$ $i = 1, \cdots, N$ of $\mathbf{z}$, $p(\mathbf{z})$ is given as

$$p(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{z} - \mathbf{z}_i), \tag{5.5}$$

where $K_{\mathbf{H}}(\cdot)$ is the specified kernel function. $\mathbf{H}$ is a symmetric positive definite $n \times n$ bandwidth matrix. We adopt a Gaussian kernel with a diagonal bandwidth matrix, i.e., $\mathbf{H} = diag(\sigma_1^2, \cdots, \sigma_d^2)$.

The bandwidth is estimated from the variance of the samples by using the method proposed in (Turlach, 1993), which is computed as

$$\sigma = 1.06\hat{\sigma}N^{-\frac{1}{5}}, \tag{5.6}$$

where $\hat{\sigma}^2$ is the variance in $N$ samples.

Substituting Equation (3.4) and Equation (3.5) into Equation (3.10) and integrating out the integrals, quadratic mutual information $QMI(a; v)$ between the audio and visual feature can be analytically computed as

$$\begin{cases} QMI(a; v|\{a_t, v_t\}) = \log \frac{V_c(\{a_t,v_t\})V_m(\{a_t\})V_m(\{v_t\})}{V_{nc}^2(\{a_t,v_t\})} \\ V_c(\{a_t, v_t\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} K_{2\sigma_a^2}(a_i - a_j)K_{2\sigma_v^2}(v_i - v_j) \\ V_s(a_j, \{a_i\}) = \frac{1}{N}\sum_{i=1}^{N} K_{2\sigma_a^2}(a_j - a_i) \\ V_s(v_j, \{v_i\}) = \frac{1}{N}\sum_{i=1}^{N} K_{2\sigma_v^2}(v_j - v_i) \\ V_m(\{a_i\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(a_j, \{a_i\}) \\ V_m(\{v_i\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(v_j, \{v_i\}) \\ V_{nc}(\{a_t, v_t\}) = \frac{1}{N}\sum_{j=1}^{N} V_s(a_j, \{a_i\}V_s(v_j, \{v_i\}), \end{cases} \tag{5.7}$$

where $\sigma_a$ and $\sigma_v$ are the estimated bandwidths of the audio and visual features.

As shown in Appendix A, this correlation analysis is invariant to the change of the scale of both the audio and visual features. The invariance of analysis of correlations to changes in scale makes our method robust against audiovisual changes in scale.

Finally, the average audiovisual correlation inside the speaker region is computed as

$$C(d) = \frac{1}{D}\sum_{x}\sum_{y} QMI(a(d); v(x, y)), \tag{5.8}$$

where $D$ is the number of pixels inside the speaker region. $a(d)$ means the audio shifted by $-d$ for drift hypothesis $d$.

Table 5.1. Detected drift *vs.* ground truth

| Ground truth (ms) | $d_{av}^*$ (ms) |
|---|---|
| -540 | -538 |
| -170 | -170 |
| 230 | 230 |

## 5.4 Experiments

As most off-the-shelf video cameras supply video data at 30 fps, we adopted these kinds of data in our experiments. The search range was set to $[-1, 1]$ s, i.e., $L = M = 30$. The window size to compute optical flow was 9×9 in all our experiments. We also assumed that the user specified the time span to be 0–5 s, within which we adopted 60 video frames to compute audiovisual correlation, i.e., $N = 60$.

To test the accuracy of our method, we synthesized a ground truth video that simulated a speaking action by vertically shaking a random dot image (320×240). The movement was computed by multiplying two sine functions with a long (3s) and a short (0.3s) period. Audio was synthesized by modulating a 2 KHz sine wave with the same movement curve. Audiovisual data are shown in Figure 5.2. We first applied our method to the data in the state of AV-sync. The detected $d_{av}^*$ was 1 ms, which was very close to the real value 0 ms. We also produced ground truth data whose AV-sync had drifted and applied these to our method. The drift values were selected so that they were not integral multiples of $T_v$ to test the accuracy of detecting sub-frames. The results are listed in Table 5.1. Our method successfully detected drifts with a maximum error below 2 ms. Experiments with the ground truth demonstrated that our method could correctly detect the drift in AV-sync.

We applied our method to real data using the CUAVE database (Patterson et al., 2002), in which people spoke English numbers in front of a green background individually or in groups. We converted images into gray in the experiments and downsampled the resolution from 720×480 to 240×160. The experimental results for the three CUAVE

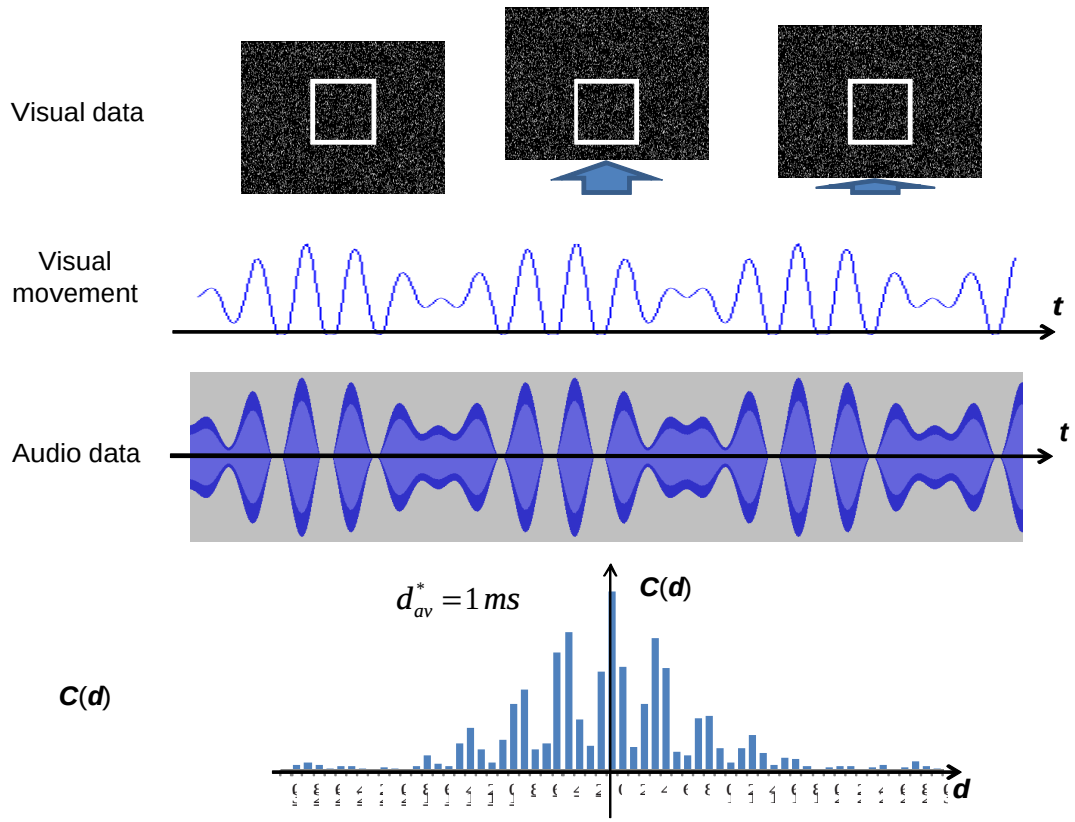



Figure 5.2. Experimental results for ground truth data. Both visual data (dotted pattern and its temporal movements) and audio data are shown, where white rectangles indicate assumed speaker regions. Bottom figure plots change in $C(d)$, with $d^*_{av}$ shown at top left.

clips are shown in Figure 5.3. Our method detected minor drifts for all three clips. Since all the drifts were below 45 ms that humans can perceive as has been suggested in (ATSC group, 2003), we could not establish their accuracy. However, the results conformed to our perception that the clips were in a state of AV-sync. Additionally, for clip (c) where there were multiple people, our method successfully located the speaker based on a comparison of audiovisual correlations as demonstrated in Figure 5.3 (c).

To test the accuracy of our method when drift occurred, we intentionally added the same drifts as the ones in Table 5.1 to the three clips in Figure 5.3. The experimental results are listed in Table 5.2. Our method successfully detected the added drifts.

As mentioned in Section 5.3, our method is robust against different visual scales and

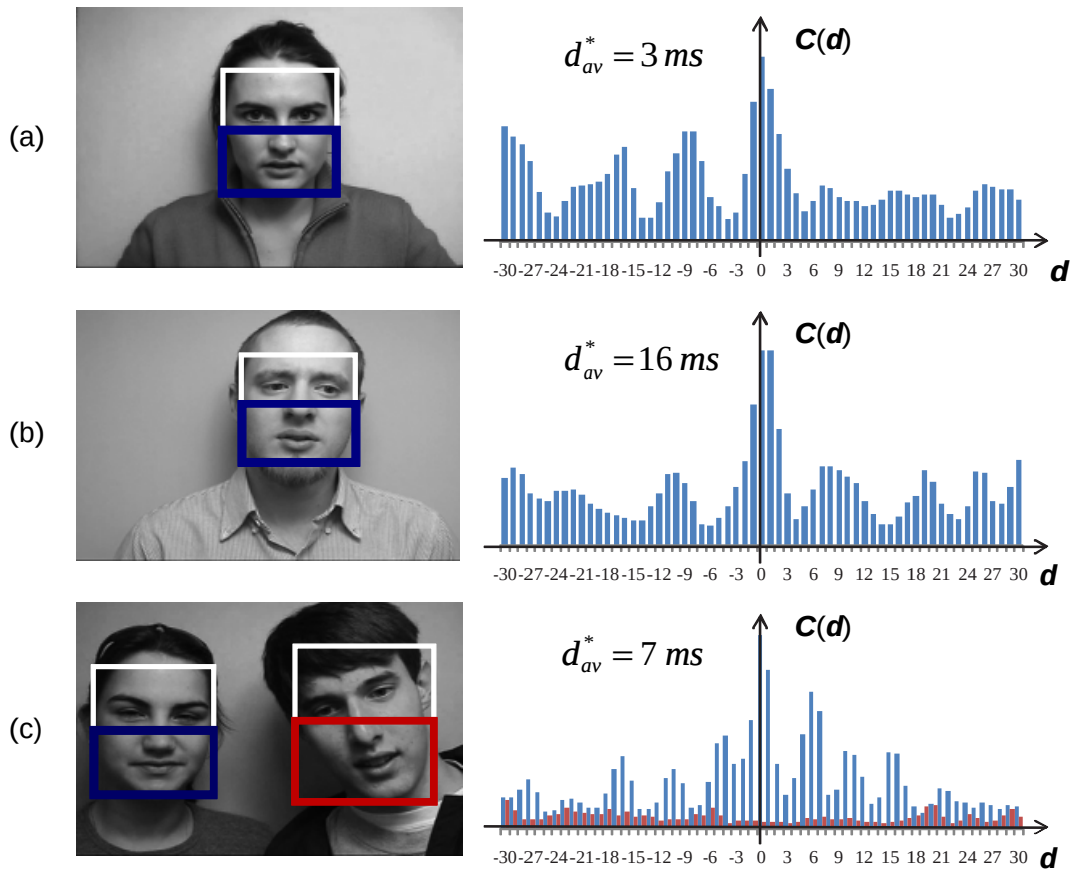



Figure 5.3. Experimental results on real data. (a), (b), and (c) are experimental results corresponding to woman, man, and two persons including only one speaker. First column shows video images and detected speaker regions. Second column shows change in $C(d)$, with $d^*_{av}$ shown at top left.

audio gains. Using clip (a) in Figure 5.3, we changed the visual scales and audio gains by 1.5 and 2 times and applied them to our method. The detected drifts are listed in Table 5.3. The visual scale was changed by increasing the downsampling rate from $1/3$ to $1/2$ and $2/3$. The audio gain was changed by multiplying audio-sample magnitudes with gains of 1.5 and 2. Considering a face detection classifier can only tolerate changes in the visual scale of 1.1–1.2 times (Viola and Jones, 2004), we applied rather large scale changes to the data. However, our method detected similar drifts for the data with audiovisual-scale changes, whose error was within 10 ms. The results in Table 5.3 also demonstrate that audio gain changes had more influence on our method because spatially

Table 5.2. Added drifts *vs.* computed values.

| Drift added (ms) | $d^*_{av}$ of (a)(ms) | $d^*_{av}$ of (b)(ms) | $d^*_{av}$ of (c)(ms) |
|---|---|---|---|
| -540 | -537 | -514 | -526 |
| -170 | -166 | -155 | -156 |
| 230 | 234 | 251 | 242 |

Table 5.3. Detection results with audiovisual-scale changes.

| | Original | Visual scale | | Audio gain | | Audiovisual | |
|---|---|---|---|---|---|---|---|
| | | 1.5 | 2 | 1.5 | 2 | 1.5 | 2 |
| Drift (ms) | 2.8 | 2.0 | 0.5 | 2.2 | 6.7 | 4.6 | 12.9 |

averaging audiovisual correlation improved the robustness of our method against changes in the visual scale.

We used two clips to test our method for languages other than English. The results for video frames and detection are shown in Figure 5.4. Clip (a) was of a native speaker who was reading Chinese news. Clip (b) was of a native speaker reading Japanese news. As we used a cheap video camera, our method detected larger drifts compared to the CUAVE data. Yet, the drifts were still below 45 s and fitted our perceptions.

The computation time with our method depends on the size of the face region and the number of frames $N$ adopted. It generally took about 190 s to do the computations for the experiments discussed in this section on our desktop, which had an Intel Core2 Quad 2.6 Ghz CPU, 3Gb of memory, and a Windows XP OS. Only one core was used for the computation. The code was written in C++ without any special optimization.

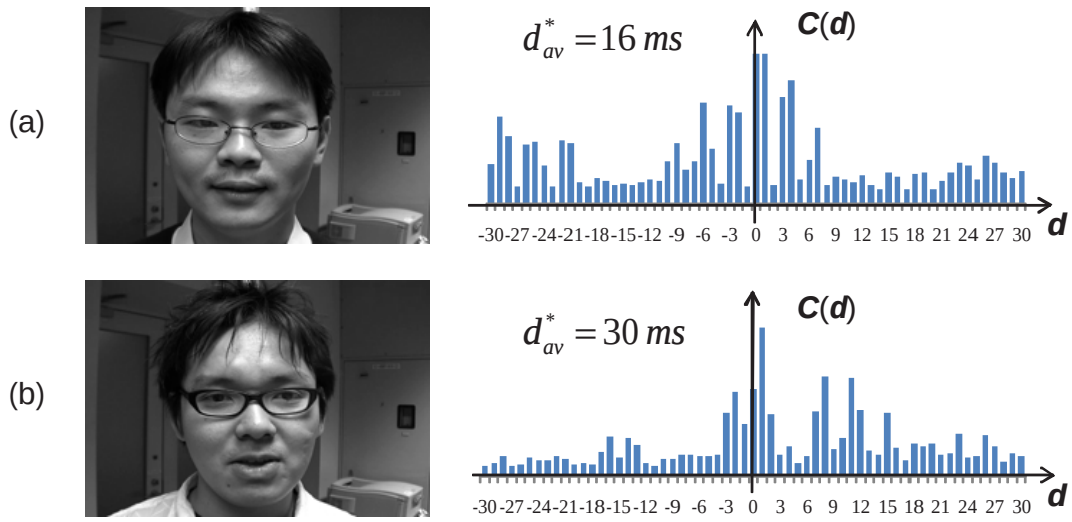



Figure 5.4. Experimental results for different languages. (a) photograph of native speakers of Chinese and (b) that of Japanese. First column shows video images. Second column shows change in $C(d)$, with $d^*_{av}$ shown at top left.

## 5.5 Conclusions

We have developed a method of recovering drifted AV-sync by analyzing audiovisual correlations. Users only need to specify the time window within which there is a stationary speaker. Our method could detect drift and recover AV-sync based on analysis of audiovisual correlations by using quadratic mutual information with kernel density estimation. Our method was not only robust against audiovisual changes in scale but was also independent of different languages.

Our current system did not take into consideration solutions against noise. Consequently, our method may break down when presented with low quality visual data or speech with background noise. In future work, we intend to investigate other audiovisual features that are robust. Although our current method requires the speaker to be stationary within a time window, we are planning to relax this constraint.

# Chapter 6

# Conclusions

This thesis has presented a framework to analyze auditory and visual information based on the evaluation of synchrony. We have given our answers to the two questions of this framework: How to and where to apply this synchrony-based audiovisual analysis.

In Chapter 2, we have reviewed all the existing techniques to analyze the audiovisual correlation. We then classified all the existing techniques to analyze the audiovisual correlation according to the feature and measure which they adopted. An experimental comparison for these features and measures has been performed to supply objective evidence on designing methods to analyze the audiovisual correlation.

In Chapter 3, we developed a method to segment the face region of a stationary speaker. To overcome the fragmental problem in the existing techniques, we initially incorporated audiovisual correlation analysis into an image segmentation framework to compute the speaker's face region by a global optimization with the similarity between pixels considered. We also developed a new method to analyze the audiovisual correlation, including the development of our feature and our measure. This correlation analysis has an advantage of being invariant to the change of both visual scale and audio gain.

In Chapter 4, we extended the ability of localization to non-stationary sound sources. To correctly analyze the audiovisual correlation for a moving sound source, we searched its movement with an optimization framework whose objective function was set as the

maximization of the correlation between locally extracted audio and visual features. The feature was developed by us and named as inconsistency, which describes the acceleration in the change of audio and visual signals. We also introduced the incremental computation of mutual information to speed up the search.

In Chapter 5, we developed a method with audiovisual correlation analysis to solve an old problem that used to be a human effort intensive work — the recovery of drifted audio-to-video synchronization. We analyzed the average audiovisual correlation in a speaker region for different drift hypotheses and took the one that maximized the average correlation to recover synchronization. The developed method is also invariance to the change of both visual scale and audio gain.

# Contributions

- Initially supplied a classification and an experimental comparison on the existing techniques to analyze the audiovisual correlation, which not only helped the research of this thesis, but also contribute to this community for future researches.
- Introduced expectation maximization learning and image segmentation framework to solve the problems of classification threshold and fragmental localization.
- Introduced the framework of audiovisual correlation maximization to correctly analyze the audiovisual correlation for non-stationary sound source, which used to be impossible.
- Used audiovisual correlation analysis to solve the old problem —recovery of the drifted audio-to-video synchronization, which used to require both special device and dedicated human effort.
- Developed new features like optical flow and inconsistency, and new measures like quadratic mutual information and incremental mutual information to analyze the audiovisual correlation.

# Directions for future work

## Sound separation

A challenging problem of analyzing the audiovisual correlation is the possible impureness of audio signal. Since microphone is omni-directional and sums the audio signals from all directions, audio signal may be a mixture of the sounds from multiple sources. In such a situation, analyzing the audiovisual correlation becomes much more difficult.

However, according to the characteristics of the mixed sound sources, the difficulty to analyze the audiovisual correlation is different. We classify all the possibilities into three situations.

First, all the sound sources emit a same sound. An example of this situation is a chorus in which all the persons are singing a same song. This situation brings marginal difficulty to the current frameworks of analyzing the audiovisual correlation since the mixed audio is equivalent to an amplified single sound. The spatial distribution of the analyzed audiovisual correlation should have multiple peaks that correspond to the positions of the sound sources. This situation in fact has been addressed in Chapter 4 of this thesis and (Kidron et al., 2007).

Second, sound sources emit different sounds, but all of them are visible to the camera. An example is an orchestra, where different instruments are played to form a symphony. Analyzing this audiovisual correlation becomes much more difficult in this case because, before we can analyze the degree of synchrony, we have to first know whether or not a set of audio events belongs to a same sound source. To first separate the audio into the sounds from different sources seems to be a solution. Yet this separation by only one audio channel is theoretically impossible. Therefore the solution of this problem relies on a joint optimization of the audiovisual correlation and the sound separation. Some works have tried to address this problem, such as (Barzelay and Schechner, 2007) and (Casanovas, 2006). However, they require a strong constraint on the combination of the sound sources, which is that the sounds from different sources cannot have overlaps

in the frequency domain. This constraint considerably limits the applicability of these techniques. The general solution of this problem is still under development.

Third, sound sources emit different sounds, and some of them are not visible. Examples include a singer with background music, or a speaker in a noisy party. This is a common case in our daily lives, but also the most difficult situation for analyzing the audiovisual correlation. Currently no technique has tried to address this problem.

Being difficult, it is attractive to solve the problems belonging to the second and third situations. With this ability, we can extract the sound emitted by a designated visual sound source, and can thus clearly hear what we want to hear in any noisy environment, such as a party, a factory, and so on. With the progresses in this field, we do hope that this can come true in the future.

## Other audiovisual applications

In this thesis, we have focused on answering the two questions: how to perform and where to apply this audiovisual analysis with synchrony evaluation. However, we believe that the answers to the second question are far more than the ones we have developed. On one hand, many old problems in computer vision and audio processing can be better solved with the integration of auditory and visual information by this framework, such as surveillance, authentication, video-teleconferencing, expression recognition, and so on. On the other hand, the problems that used to be solved by other ways may be solved by using this new framework. An example is the method we developed to recover audio-to-video synchronization, which has been introduced in Chapter 5. The method to separate the sounds from different sources is another example. We do hope that more novel applications like these two examples can be discovered to make machines more intelligent and helpful to our lives.

# Appendix A

# Scale invariance of our audiovisual correlation analysis

In this appendix, we show that our method to analyze the audiovisual correlation is invariant to the change of scale.

First, we show that the change of the scale of a one-dimensional random variable $z$ leads to the multiplication of a coefficient to the original pdf only. Suppose that $z$ is multiplied by a scale coefficient of $s$, i.e., $z_s = sz$. The new bandwidth $\sigma_s$ estimated by Equation (3.5) becomes

$$\sigma_s = 1.06 s \hat{\sigma} n^{-\frac{1}{5}} = s\sigma. \tag{A.1}$$

Subsituting Equation (A.1) into Equation (3.4), $p(sz)$ can be represented by $p(z)$ as

$$\begin{aligned} p(sz) &= \frac{1}{N}\sum_{i=1}^{N} K_{\sigma_s}\left((sz - sz_i)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}s\sigma} \exp\left(-\frac{s^2(z-z_i)^2}{2s^2\sigma^2}\right) \\ &= \frac{1}{s}p(\mathbf{z}). \end{aligned} \tag{A.2}$$

This conclusion can be easily extended to the $n$-dimensional case. Since the band-

width matrix $\mathbf{H}$ is supposed to be diagonal, i.e., $\mathbf{H} = diag(\sigma_1, \cdots, \sigma_n)$, we have

$$\begin{aligned} p(\mathbf{sz}) &= \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{n} K_{\sigma_{s_j}}(s_j z - s_j z_{ij}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j=1}^{n}\frac{1}{s_j}\right) K_{\mathbf{H}}(\mathbf{z}-\mathbf{z}_i) \\ &= \left(\prod_{j=1}^{n}\frac{1}{s_j}\right) p(\mathbf{z}), \end{aligned} \tag{A.3}$$

where $\mathbf{s}$ is a diagonal scale matrix, $\mathbf{s} = diag(s_1, \cdots, s_n)$.

Then, we show that quadratic mutual information is invariant to the change of the scale of audiovisual features. Suppose that the scale of the audio feature $a$ is $s_a$, and the scale of the visual feature $v$ is $s_v$. Substituting Equation (A.2) into Equation (3.10), we have

$$\begin{aligned} & QMI(s_a a; s_v v) \\ = \; & \log \frac{\iint \frac{1}{s_a^2 s_v^2} p^2(a,v) da dv \iint \frac{1}{s_a^2 s_v^2} p^2(a) p^2(v) da dv}{(\iint \frac{1}{s_a^2 s_v^2} p(a,v) p(a) p(v) da dv)^2} \\ = \; & QMI(a; v). \end{aligned} \tag{A.4}$$

Consequently, our method to analyze the audiovisual correlation is invariant to the change of scale.

# Appendix B

# Incremental computation of entropy

Here we show the correctness of Equation (4.16) such that entropy can be computed incrementally. In frame $k+1$, entropy can be computed by definition as

$$H^{(k+1)}(z) = -\sum_i \frac{n_i^{(k+1)}}{k+1} \log \frac{n_i^{(k+1)}}{k+1}, \tag{B.1}$$

where $n_i^{(k+1)}$ is the number of the samples in the histogram bin $i$. Suppose that the index of the sample that comes in frame $k+1$ is $c$. Since only one sample is added, we have

$$n_i^{(k+1)} = \begin{cases} n_i^{(k)} & i \neq c \\ n_i^{(k)} + 1 & i = c \end{cases}. \tag{B.2}$$

Based on this consideration, the computation of $H^{(k+1)}(z)$ can be extended as

$$\begin{aligned}
H^{(k+1)}(z) &= -\sum_i \frac{n_i^{(k+1)}}{k+1}(\log \frac{n_i^{(k+1)}}{k} + \log \frac{k}{k+1}) \\
&= -\sum_i \frac{n_i^{(k+1)}}{k+1} \log \frac{n_i^{(k+1)}}{k+1} - \sum_i \frac{n_i^{(k+1)}}{k+1} \log \frac{k}{k+1} \\
&= -\sum_{i,i\neq c} \frac{n_i^{(k+1)}}{k+1} \log \frac{n_i^{(k+1)}}{k+1} - \frac{n_c^{(k+1)}}{k+1} \log \frac{n_c^{(k+1)}}{k} - \log \frac{k}{k+1} \\
&= -\sum_{i,i\neq c} \frac{n_i^{(k)}}{k+1} \log \frac{n_i^{(k)}}{k+1} - \frac{n_c^{(k)}+1}{k+1} \log \frac{n_c^{(k)}+1}{k} - \log \frac{k}{k+1} \\
&= -\frac{k}{k+1} \sum_i \frac{n_i^{(k)}}{k} \log \frac{n_i^{(k)}}{k} + \frac{n_c^{(k)}}{k+1} \log \frac{n_c^{(k)}}{k} - \frac{n_c^{(k)}+1}{k+1} \log \frac{n_c^{(k)+1}}{k} - \log \frac{k}{k+1} \\
&= \frac{k}{k+1} H^{(k)}(z) + \frac{1}{k+1} \left( n_c^{(k)} \log \frac{n_c^{(k)}}{k} - (n_c^{(k)}+1) \log \frac{n_c^{(k)}+1}{k} \right) - \log \frac{k}{k+1} \\
&= \frac{1}{k+1} \left( kH^{(k)}(z) - \log \frac{(1+n_c^{(k)})^{1+n_c^{(k)}}}{{n_c^{(k)}}^{n_c^{(k)}}} \right) + \frac{k}{k+1} \log k - \log \frac{k}{k+1}. \quad \text{(B.3)}
\end{aligned}$$

Setting that $n = n_c^{(k)}$, and $C(k) = \frac{k}{k+1} \log k - \log \frac{k}{k+1}$, we have Equation (4.16).

# Appendix C

# Limit of Equation (4.17)

In this part, we show the limit in Equation (4.17). Given that $n$ is a continuous variable and approaching zero, Equation (4.17) can be extended as

$$\begin{aligned}
\lim_{n\to 0}\frac{(1+n)^{1+n}}{n^n} &= \lim_{n\to 0}(1+n)(1+\frac{1}{n})^n \\
&= \lim_{n\to 0}(1+n)\lim_{n\to 0}(1+\frac{1}{n})^n \\
&= \lim_{n\to 0} e^{n\ln(1+\frac{1}{n})} \\
&= e^{\lim_{n\to 0} n\ln(1+\frac{1}{n})}. \qquad \text{(C.1)}
\end{aligned}$$

Below we show the limit of the exponent in Equation (C.1) by L'Hospital's rule.

$$\begin{aligned}
\lim_{n\to 0} n\ln(1+\frac{1}{n}) &= \lim_{n\to 0}\frac{\ln(1+\frac{1}{n})}{\frac{1}{n}} \\
&= \lim_{n\to 0}\frac{\frac{1}{1+\frac{1}{n}}\frac{-1}{n^2}}{\frac{-1}{n^2}} \\
&= 0. \qquad \text{(C.2)}
\end{aligned}$$

Substituting Equation (C.2) into Equation (C.1), we have

$$\lim_{n\to 0}\frac{(1+n)^{1+n}}{n^n} = e^0 = 1. \qquad \text{(C.3)}$$

# Bibliography

ATSC group. “Relative Timing of Sound and Vision for Broadcast Operations”. *Advanced Television Systems Committee report*, IS-191, 2003.

Barzelay, Z. and Schechner, Y.. “Harmony in Motion”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–8, 2007.

Bertelson, P., Vroomen, J., Wiegeraad, G., and Gelder, B.. “Exploring the Relation between McGurk Interference and Ventriloquism”. In *Proceedings of the International Conference on Spoken Language Processing*, 2:559–562, 1994.

Berthouze, L., Kozima, H., Prince, C., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C.. “Taking Synchrony Seriously: A Perceptual-Level Model of Infant Synchrony Detection”. *Lund University Cognitive Studies*, 117:89–96, 2004.

Boykov, Y., Veksler, O., and Zabih, R.. “Fast Approximate Energy Minimization via Graph Cuts”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

Boykov, Y. and Kolmogorov, V.. “An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

Boykov, Y. and Funka-Lea, G.. “Graph Cuts and Efficient N-D Image Segmentation”. *International Journal of Computer Vision*, 70(2):109–131, 2006.

Casanovas, A.. “Blind Audiovisual Source Separation using Sparse Redundant Representations”. *Master thesis*, Signal Processing Institute, EPFL, 2006.

Chen, J. and Tang, C.. "Spatio-Temporal Markov Random Field for Video Denoising". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–8, 2007.

Dalal, N. and Triggs, B.. "Histograms of Oriented Gradients for Human Detection". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:886–893, 2005.

Doornik, J. and Hansen, H.. "An Omnibus Test for Univariate and Multivariate Normality (Working paper)", NuOEeld College, Oxford, 1994.

Driver, J.. "Enhancement of Selective Listening by Illusory Mislocation of Speech Sounds due to Lip-Reading", *Nature*, 381:66–68, 1996.

Fisher, J. and Darrell, T.. "Probabalistic Models and Informative Subspaces for Audiovisual Correspondence". In *Proceedings of the European Conference on Computer Vision*, pp.592–603, 2002.

Fisher, J. and Darrell, T.. "Speaker Association with Signal-Level Audiovisual Fusion". *IEEE Transactions on Multimedia*, 6(3):406–413, 2004.

Hershey, J. and Movellan, J.. "Audio Vision: Using Audiovisual Synchrony to Locate Sounds". In *Proceedings of the Neural Information Processing Systems Conference*, pp.813–819, 1999.

Huang, C., Ai, H., Li, Y., and Lao, S.. "Vector Boosting for Rotation Invariant Multi-View Face Detection". In *Proceedings of the IEEE International Conference on Computer Vision*, 1:446–453, 2005.

Jaimes, A., Nagamine, T., Liu, J., Omura, K., and Sebe, N.. "Affective Meeting Video Analysis". In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp.1412–1415, 2005.

Kass, M., Witkin, A., and Terzolpoulos, D.. "Snakes: Active Contour Models". *International Journal of Computer Vision*, 1(4):321–331, 1988.

Kidron, E., Schechner, Y., and Elad, M.. "Pixels that Sound". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.88–95, 2005.

Kidron, E., Schechner, Y., and Elad, M.. "Cross-Modal Localization via Sparsity", *IEEE Transactions on Signal Processing*, 55(4):1390–1404, 2007.

Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., and Rother, C.. " Bi-Layer Segmentation of Binocular Stereo Video". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1186–1194, 2005.

Kumano, S., Otsuka, K., Yamato, J., Maeda, J., and Sato, Y.. "Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates". In *Proceedings of the Asian Conference on Computer Vision*, pp.324–334, 2007.

Li, X., Sun, L., Tao, L., Xu, G., and Jia, Y.. "A Speaker Tracking Algorithm Based on Audio and Visual Information Fusion Using Particle Filter". In *Proceedings of the International Conference on Image Analysis and Recognition*, II:572–580, 2004.

Lucas, B. and Kanade, T.. "An Iterative Image Registration Technique with an Application to Stereo Vision". In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp.674–679, 1981.

Luettin, J., Thacker, N., and Beet, S.. "Speaker Identification by Lipreading". In *Proceedings of the International Conference on Spoken Language*, 1:62–65, 1996.

Maison, B., Neti, C., and Senior, A.. "Audio-Visual Speaker Recognition for Video Broadcast News: Some Fusion Techniques". In *Proceedings of the IEEE Multimedia Signal Processing*, pp.1–7, 1999.

Matsumoto, M. and Nishimura, T.. "Mersenne Twister: a 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator". *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30, 1998.

McGurk, H. and MacDonald, J.. "Hearing Lips and Seeing Voices". *Nature*, 264(5588):746–748, 1976.

Monaci, G., Escoda, O. and Vandergheynst, P.. "Analysis of Multimodal Signals using Redundant Representations". In *Proceedings of the IEEE International Conference on Image Processing*, pp. 145–148, 2005.

Monaci, G. and Vandergheynst, P.. "Audiovisual Gestalts". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, pp.1–8, 2006.

O'Donovan, A., Duraiswami, R., and Neumann, J.. "Microphone Arrays as Generalized Cameras for Integrated Audio Visual Processing". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–8, 2007.

Parzen, E.. "On the Estimation of Probability Density Function and the Mode". *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Patterson, E., Gurbuz, S., Tufekci, Z., and Gowdy, J.. "Moving-Talker, Speaker-Independent Feature Study and Baseline Results using the Cuave Multimodal Speech Corpus". *EURASIP Journal on Applied Signal Processing*, 2002(11):1189-1201, 2002.

Petajan, E.. "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)". *Ph.D thesis*, University of Illinois at Urbana-Champaign, 1984.

Poh, N. and Korczak, J.. "Hybrid Biometric Person Authentication Using Face and Voice Features". In *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*, pp.348–348, 2001.

Rabiner, L. and Juang, B.. "Fundamentals of Speech Recognition". *Prentice Hall*, 1993.

Ravulapalli, S. and Sarkar, S.. "Association of Sound to Motion in Video using Perceptual Organization". In *Proceedings of the International Conference on Pattern Recognition*, 1:1216–1219, 2006.

Reeves, B. and Voelker, D.. "Effects of Audio-Video Asynchrony on Viewer's Memory, Evaluation of Content and Detection Ability". *Research report*, Standford University, 1993.

Renyi, A.. "On Measures of Entropy and Information". In *Proceedings of the Fourth Berkeley Symposium on Mathimatical Statistics and Probability*, 1:547–561, 1961.

Rivet, B., Girin, L., and Jutten, C.. "Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures". *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):96–108, 2006.

Rodgers, J. and Nicewander, W.. "Thirteen Ways to Look at the Correlation Coefficient". The American Statistician, 42:59–66, 1988.

Schoenemann, T. and Cremers, D.. "High Resolution Motion Layer Decomposition using Dual-Space Graph Cuts", In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1–7, 2008.

Shannon, C.. "Prediction and Entropy of Printed English". *The Bell System Technical Journal*, 30:50-64, 1951.

Shechtman, E. and Irani, M.. "Space-Time Behaviour-Based Correlation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):2045–2056, 2007.

Shi, J. and Tomasi, C.. "Good Features to Track". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.593–600, 1994.

Silva, D.. "Audiovisual emotion recognition". In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 1:649–654, 2004.

Smaragdis, P. and Casey, M.. "Audio/visual Independent Components". In *Proceedings of the International Symposium on Independent Component Analysis and Blind Source Separation*, 709–714, 2003.

Song, M., Bu, J., Chen, C., and Li, N.. "Audio-Visual based Emotion Recognition - A New Approach". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1020–1025, 2004.

Torkkola, K.. "Feature Extraction by Non-Parametric Mutual Information Maximization". *Journal of Machine Learning Research*, 3:1415–1438, 2003.

Turlach, B.. "Bandwidth Selection in Kernel Density Estimation: A Review". In *CORE and Institut de Statistique*, pp.23–493, 1993.

Vandergheynst, P. and Frossard, P.. "Efficient Image Representation by Anisotropic Refinement in Matching Pursuit". In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:1757–1760, 2001.

Viola, P. and Jones, M.. "Robust Real-Time Face Detection". *International Journal of Computer Vision*, 57(2):137–154, 2004.

Xu, D., Principe, J., and Fisher, J.. "A Novel Measure for Independent Component Analysis (ICA)". In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:1161–164, 1998.

Yu, T., Zhang, C., Cohen, M., Rui, Y., and Wu, Y.. "Monocular Video Foreground/Background Segmentation by Tracking Spatial-Color Gaussian Mixture Models". In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pp.55–63, 2007.

Waters, K. and Levergood, T.. "An Automatic Lip-Synchronization Algorithm for Synthetic Faces". In *Proceedings of the ACM international conference on Multimedia*, pp.149–156, 1994.

Zhang, Y.. "Elements of Operational Research". *Tsinghua University Press*, 1994.

Zhu, S. and Yuille, A.. "Region Competition: Unifying Snakes, Region Growing, and Bayes/Mdl for Multiband Image Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

# Publications

## Journals

[1] 劉玉宇，佐藤洋一.
“音と映像の相関を用いた画像分割による話者領域の切り出し”.
**情報処理学会論文誌コンピュータビジョンとイメージメディア**,
1(2):32–40, 2008 年 7 月.

[2] Liu, Yuyu and Sato, Yoichi.
“Recovery of Audio-to-Video Synchronization through Analysis of Cross-Modality Correlation”.
Submitted to *Pattern Recognition Letters*.

[3] Liu, Yuyu and Sato, Yoichi.
“Segmentation of the Speaker’s Face Region with Audiovisual Correlation”.
Submitted to *IEICE Transactions on Information and Systems*.

## International conference

[4] Liu, Yuyu and Sato, Yoichi.
“Finding Speaker Face Region by Audiovisual Correlation”.
In *Proceedings of the European Conference on Computer Vision Workshop on* $M^2SFA^2$,
pp.1–12, October, 2008.

[5] Liu, Yuyu and Sato, Yoichi.

"Recovering Audio-to-Video Synchronization by Audiovisual Correlation Analysis".
In *Proceedings of the International Conference on Pattern Recognition*,
pp.1–4, December, 2008.
(Best Industry-Related Paper Award)

[6] Liu, Yuyu and Sato, Yoichi.
"Visual Localization of Non-stationary Sound Sources".
to appear in ACM Multimedia 2009,
October, 2009.

## Domestic conference with peer review

[7] Liu, Yuyu and Sato, Yoichi.
"Talking Speaker Segmentation using Audio-Visual Correlation".
**画像の認識・理解シンポジウム** *(MIRU2007)*,
pp.1–8, 2007 年 7 月.

[8] Liu, Yuyu and Sato, Yoichi.
"Recovering Audio-to-Video Synchronization by Audiovisual Correlation Analysis".
**画像の認識・理解シンポジウム** *(MIRU2008)*,
pp.1–6, 2008 年 7 月.

[9] 劉玉宇，佐藤洋一.
"音と映像の相関分析に基づく移動音源特定".
**画像の認識・理解シンポジウム** *(MIRU2009)*,
pp.1–8，2009 年 7 月.
(Best Student Paper Award)