PHASE DIFFUSION FOR THE SYNCHRONIZATION OF HETEROGENOUS SENSOR STREAMS

Steve Gu and Carlo Tomasi

Duke University
Department of Computer Science
450 Research Dr, Durham, NC 27708

ABSTRACT

The analysis of complex human activity typically requires multiple sensors: cameras that take videos from different directions and in different areas, microphones, proximity sensors, range finders, and more. Scenarios where it is not possible to associate reliable clocks to each of the sensors pose a synchronization problem between heterogeneous data streams. In this paper, we propose a new theoretical framework for measuring the synchrony between heterogenous sensor streams. The main idea is to model the phase disparity between two data streams explicitly as an Ornstein-Uhlenbeck random process. Based on this model, we derive a simple method for synchronizing of underlying sources. We illustrate the ideas with experiments on audio-visual synchronization and human motion categorization, and report promising results.

Index Terms— synchronization, phase diffusion, Ornstein-Uhlenbeck process

1. INTRODUCTION

It is often necessary to synchronize data streams recorded from the same scene but with different sensors, such as audio and video, or more. When no reliable clocks are associated to the various streams, synchronization has to be based on the sensor data itself. Even when time stamps are available, synchronization methods based on the sensed data will still be usefl. For example, in order to direct the camera to the person who is speaking during a teleconference, it is necessary to know who is talking at the moment and hence match that part of the image to the audio signal. The quality of audio-video synchronization in dubbed movies can be improved by synchronizing speech in the target language with the facial movements of the actor speaking in the source language.

The synchronization problem is much more general than these examples suggest. For instance, object classification, human motion categorization and image/video similarity comparisons can be thought of as the problem of synchronizing the underlying sources, rhythms and representation. In this paper, we develop a general theoretical framework for measuring the degree of synchronization between heterogenous sensor streams. We show that such a framework can address useful applications such as audio-visual synchronization and motion categorization.

Synchronization has been studied mainly for audio-video [5, 9, 2], through approaches that model the two signals either as different transformations of a hidden source signal, or as transformations of one another. These approaches start from measures of the similarity, correlation, or mutual information between the amplitude values of the sensor signals. While these methods can work well in practice, some ambiguity remains in the amplitude domain, because of the heterogeneous nature of the two streams.

Instead, we propose to model the two signals as separate dynamic systems, loosely coupled by occasional coincidences. To emphasize the temporal aspects of the two (or more) signals to be synchronized, we build our model in the phase domain through the Hilbert transform [3], a tool that is commonly used in electrical communications[7].

We show that the phase disparity between heterogenous signals is a process obeying the Langevin equation. This idea is germane to some work in computer vision[6] that uses Brownian motion to model stereo disparity. Through a first order Taylor approximation, we explicitly model the phase disparity as an Ornstein-Uhlenbeck process[4], and use it to develop a discriminative measure for synchrony. We report preliminary yet promising experiments on both audio-visual synchronization and motion categorization.

2. MODELING PHASE DISPARITY

We model the two signals to be synchronized as the states of a pair of weakly coupled dynamic systems[8]:

$$\frac{d\mathbf{x}^{(1)}}{dt} = \mathbf{f}^{(1)}(\mathbf{x}^{(1)}) + \varepsilon\mathbf{p}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \quad (1)$$

$$\frac{d\mathbf{x}^{(2)}}{dt} = \mathbf{f}^{(2)}(\mathbf{x}^{(2)}) + \varepsilon\mathbf{p}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \quad (2)$$

where $\mathbf{f}^{(i)}$ governs the dynamics of each individual signal, and $\varepsilon\mathbf{p}$ is a weak coupling force representing the interaction

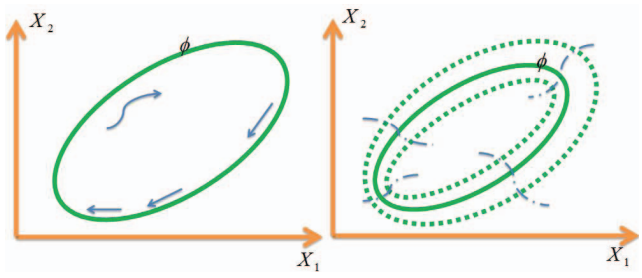

Fig. 1. A stable limit cycle and the perturbed version

between the two systems. Without coupling terms, each state vectors orbits around a closed, stable curve called the *limit cycle*. The weak coupling term will generally perturb the limit cycle away from its original trajectory (Figure 1). We introduce the phase ϕ as a coordinate along the limit cycle, such that it grows monotonically in the direction of the motion and gains 2π during each rotation. This phase signal disregards the amplitude of $\mathbf{x}^{(i)}$, and captures the durations of the intervals between its zero crossings. Thus, in a sense, the phase signal conveys pure temporal information about the input. The limit cycle can be re-parameterized so that the phase grows uniformly in time (a unit-speed curve) to obey the equation: $d\phi_i(\mathbf{x}^{(i)})/dt = \omega_i$ in the absence of perturbation. With weak coupling, we obtain the following equations:

$$\frac{d\phi_1(\mathbf{x}^{(1)})}{dt} = \omega_1 + \varepsilon \sum_k \frac{\partial \phi_1(\mathbf{x}^{(1)})}{\partial x_k^{(1)}} p_k^{(1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \quad (3)$$

$$\frac{d\phi_2(\mathbf{x}^{(2)})}{dt} = \omega_2 + \varepsilon \sum_k \frac{\partial \phi_2(\mathbf{x}^{(2)})}{\partial x_k^{(2)}} p_k^{(2)}(\mathbf{x}^{(2)}, \mathbf{x}^{(1)}) \quad (4)$$

We then define: $Q_i(\phi_1, \phi_2) = \sum_k \frac{\partial \phi_i(\mathbf{x}^{(i)})}{\partial x_k^{(i)}} p_k^{(i)}(\mathbf{x}^{(i)}, \mathbf{x}^{(3-i)})$ and further simplify the equation as: $\frac{d\phi_i(\mathbf{x}^{(i)})}{dt} = \omega_i + Q_i(\phi_1, \phi_2)$. Since Q_i is periodic in ϕ_1 and ϕ_2, we expand it using double Fourier series: $Q_i(\phi_1, \phi_2) = \sum_{k,l} a_i^{k,l} e^{\mathbf{i}k\phi_1 + \mathbf{i}l\phi_2}$. In the coupling, only low frequency components contribute steadily to the deviation of limit cycles, that is, only those terms which satisfy the resonance condition: $k\phi_1 + l\phi_2 \approx 0$ will be preserved. Assume that the two natural frequencies ω_1 and ω_2 are nearly in resonance: $\frac{\omega_1}{\omega_2} \approx \frac{\phi_1}{\phi_2} \approx \frac{m}{n}$, then all the terms in the Fourier series with $k = nj, l = -mj$ are resonant and contribute to the equations. Therefore, we have: $\frac{d\phi_i}{dt} = \omega_i + \varepsilon \sum_j a_1^{j(n-m)} e^{\mathbf{i}j(n\phi_i - m\phi_{3-i})} \triangleq \omega_i + \varepsilon q_i(n\phi_i - m\phi_{3-i})$ for $i = 1, 2$. Let the phase disparity $\Psi \triangleq n\phi_1 - m\phi_2$, $\Omega \triangleq m\omega_2 - n\omega_1$, $q(\Psi) \triangleq nq_1(\Psi) - mq_2(-\Psi)$. Then,

$$\frac{d\Psi}{dt} = -\Omega + \varepsilon q(\Psi) \quad (5)$$

In the presence of noise, we can further write (5) as:

$$\frac{d\Psi}{dt} = -\Omega + \varepsilon q(\Psi) + \sigma \xi(t) \quad (6)$$

where $\xi(t)$ is typically Gaussian. Equation (6) is called the Langevin equation.

3. ORNSTEIN-UHLENBECK PROCESS

The Langevin equation turns out to be a natural model for the phase disparity. However, the q function in (6) does not have a specific form, thus making the solution intractable in general cases. To seek an explicit solution, we need to specify q manually. Take the Taylor expansion of q as:

$$q(\Psi) = \sum_{n=0}^{\infty} \frac{q^{(n)}(0)}{n!} \Psi^n = q(0) + q'(0)\Psi + O(\Psi^2)... \quad (7)$$

If we only take a zero order approximation, that is, we replace $q(\Psi)$ with $q(0)$ in (6), it can be shown that Ψ follows a drifted Brownian motion: $\Psi(t) = [q(0) - \Omega]t + \sigma \int_0^t \xi(t)dt$. Although it is possible to substitute $q(\Psi)$ with an arbitrary higher-order approximation using Taylor series representation, it becomes harder to find an explicit solution for this stochastic differential equation. Instead, we only take the first-order Taylor approximation, that is, we substitute $q(\Psi)$ with $q(0) + q'(0)\Psi$ and derive the equation:

$$d\Psi(t) = -\lambda \left[\Psi(t) - \mu\right] dt + \sigma \xi(t) dt \quad (8)$$

where $\lambda = -\varepsilon q'(0)$ and $\mu = \frac{\Omega - \varepsilon q(0)}{\varepsilon q'(0)}$.Equation (8) defines a stochastic process called Ornstein-Uhlenbeck process, and admits an explicit solution:

$$\Psi(t) = \Psi(0)e^{-\lambda t} + \mu(1 - e^{-\lambda t}) + \sigma e^{-\lambda t} \int_0^t e^{\lambda s} \xi(s) ds \quad (9)$$

4. MEASURE FOR SYNCHRONY

Consider the physical intuitions behind the parameters of Ornstein-Uhlenbeck process. The term $|\lambda|$ is proportional to ε, which measures the strength of the coupling. The bigger ε, the stronger the coupling, and the more likely synchronization is to occur. Also, $|\mu| \approx \left|\frac{\Omega}{\varepsilon q'(0)}\right| \propto |\Omega| = |m\omega_2 - n\omega_1|$. The smaller $|\Omega|$ is, the more likely resonance will be, so that a small $|\Omega|$ is a condition for synchrony. Also notice that σ measures the strength of noise. The bigger σ, the more unreliable the synchrony measure will be, so σ can be used to deemphasize poor segments of the signal. We combine all these three considerations into the following, single *measure of synchrony*:

$$\kappa = \left|\frac{\lambda}{\sigma\mu}\right| \quad (10)$$

5. PARAMETER ESTIMATION

It can be proven that the Ornstein-Uhlenbeck process is Markovian and Gaussian with parameters $\mathbb{E}[\Psi_t] = \Psi_0 e^{-\lambda t} + \mu(1-e^{-\lambda t}) \xrightarrow[t\to\infty]{} \mu$ and $\text{Var}[\Psi_t] = \frac{\sigma^2}{2\lambda}(1-e^{-2\lambda t}) \xrightarrow[t\to\infty]{} \frac{\sigma^2}{2\lambda}$. In order to estimate the parameters λ, μ, σ, consider a set of $n+1$ samples from $\Psi(t)$: $\Psi_0, \Psi_1, ..., \Psi_n$ where the sampling interval is ϵ. We can write the conditional probability as a Gaussian distribution: $p(\Psi_{i+1}|\Psi_i) = \mathcal{N}(\Psi_{i+1}|\hat{\mu}_i, \hat{\sigma})$ where $\hat{\mu}_i = \Psi_i e^{-\lambda\epsilon} + \mu(1 - e^{-\lambda\epsilon})$ and the variance $\hat{\sigma} = \sigma\sqrt{\frac{(1-e^{-2\lambda\epsilon})}{2\lambda}}$. By the Markov property the joint probability distribution can be written as:

$$p(\Psi_0, \Psi_1, ..., \Psi_n) = p(\Psi_0) \prod_{i=1}^{n} p(\Psi_i|\Psi_{i-1}) \tag{11}$$

$$\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp^{-\frac{(\psi_i - \hat{\mu}_{i-1})^2}{2\hat{\sigma}^2}} \tag{12}$$

By taking the logarithm of the right side of (12) and setting the derivatives w.r.t λ, μ and σ separately to zero, we derive the closed-form expression for κ defined in equation (10). Details are omitted for brevity.

6. EXPERIMENTAL RESULTS

Our audio-visual synchronization experiments show that our theoretical analysis can capture the synchrony between heterogenous sensor streams, *i.e.*, audio and video signals. To show generality, we also apply our analysis to the human motion categorization task of distinguishing between running and walking gaits in an unsupervised way. Although both sequences are recorded by video cameras, the appearance and moving directions of a person are very different, making such a job non-trivial.

6.1. Audio Visual Synchronization

The video is extracted from an online CNN news broadcast where one single person is talking in front of the camera. The sequence consists of 584 frames of size 320×240, collected at 30fps. The corresponding audio signal contains 623556 samples, at a sampling frequency of 32 kHz. To process the visual part, we first use the Lucas-Kanade-Tomasi [10] tracker to extract the regions containing the moving lips. After that, each snapshot of the mouth is decomposed into a set of Haar wavelet coefficients. We measure the strength of the movement of the mouth by summing the square of all horizontal wavelet coefficients.

To process the audio signal, we first take the envelope by using a full-wave rectifier and a low pass filter. Both video and audio signals are further smoothed by using a Bessel IIR band pass filter with a common narrow-band window. It is important to use the same band pass filter for both signals so

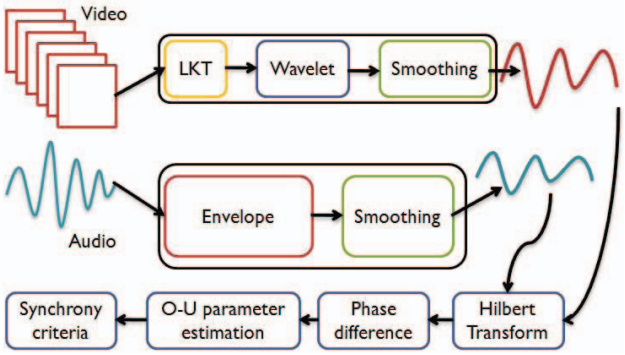

Fig. 2. A system for measuring the audio-visual synchrony

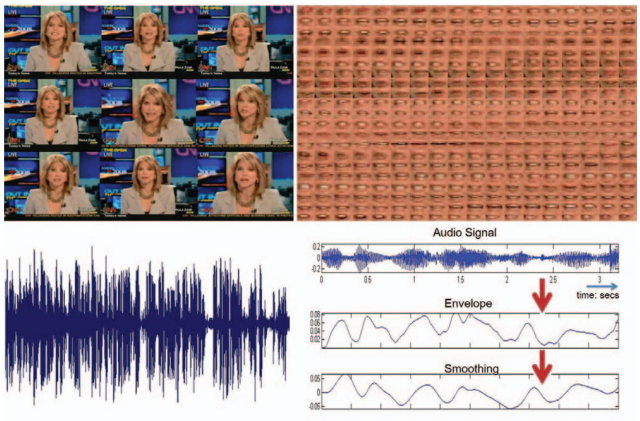

Fig. 3. Video and audio samples with pre-processing

that the resonance condition is automatically imposed. This is because $\Delta\phi = \Delta\omega\Delta t$ and small $\Delta\omega$ leads to small phase disparity, which can be closely approximated by the Ornstein-Uhlenbeck process. Fig.2 describes our system for measuring the synchrony between video and audio signals. The top row of Fig.3 displays several snapshots of the video and the extracted region of mouth by using the Lucas-Kanade tracker while the bottom row shows a plot of the audio signal along with the steps for pre-processing (envelope + smoothing).

In order to measure the synchrony between audio and video parts, we extract the central 400 frames from the video and shift this sequence in time from -50 frames to $+50$ frames. For each shift, we calculate the phase using Hilbert transform for both segments of signals and re-estimate the parameters for κ. Fig.4 shows the result using our synchrony measure where the peak is almost in its central location. Although there is a noticeable deviation of +3 frames shift against the ground-truth, the deviation is considerably small relative to the duration of the video.

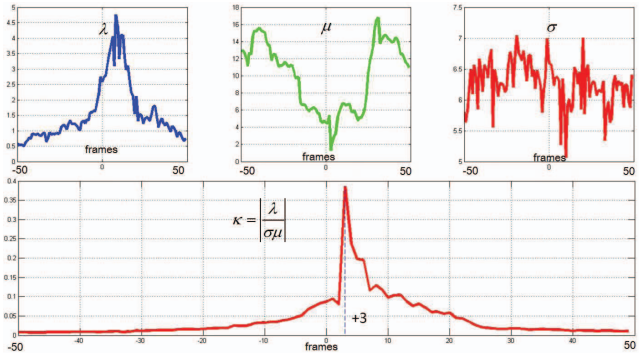

Fig. 4. Measure for audio-visual synchrony using κ

6.2. Human Motion Categorization

We also apply our synchrony measure to distinguish between walking and running gaits using the video from [1]. Fig.5 shows four snapshots of the videos we used in the experiment. To measure motion, we sum up the absolute values of the temporal gradients for each frame in both videos. The left image of Fig.6 plots the measure for two walking sequences, where an obvious periodic motion is being detected. The right image shows the phase disparity between two videos: Each curve corresponds to the phase disparity between two video streams shifted by the frame index. We take the lower envelope of the all the measures and calculate κ for synchrony. In the experiment, we find that κ is much higher when either two walking sequences or two running sequences are compared with each other, and drops dramatically when videos from different categories (walking vs. running) are compared. This indicates the effectiveness of our synchrony measure.

7. CONCLUSIONS

In this paper, we develop a new theoretical framework for measuring the synchrony between heterogenous data streams, which we model as two interacting dynamic systems. We use the Ornstein-Uhlenbeck process, *i.e.*, the first-order Taylor approximation to the Langevin equation, to model the phase disparity. Based on that, we develop a simple synchrony measure, namely, $\kappa = \left|\frac{\lambda}{\sigma\mu}\right|$. We report promising experiment

Fig. 5. Snapshots of video sequences

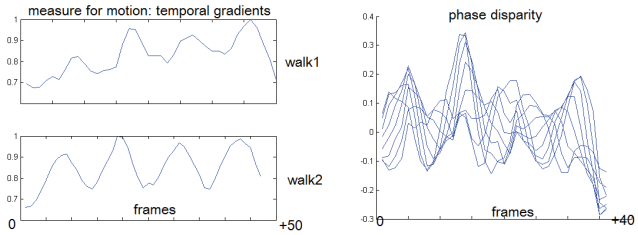

Fig. 6. Measure for videos and phase disparity

results on both audio-visual synchronization and human motion categorization using our theoretical analysis. We believe that the concept of synchronization and the theoretical model based on Ornstein-Uhlenbeck process can be further generalized and adapted to many applications in future work.

8. REFERENCES

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[2] R. Cutler and L. S. Davis. Look who's talking: Speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo (III)*, pages 1589–1592, 2000.

[3] D. Gabor. *Theory of Communication*. J. Inst. Elec. Eng. (London), 93:429–457, 1946.

[4] G.E.Uhlenbeck and L.S.Ornstein. *On the theory of Brownian Motion*. Phys.Rev. 36:823C41, 1930.

[5] J. Hershey and J. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 1999.

[6] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *PAMI.*, 16(9):920–932, 1994.

[7] P. Kovesi. *Invariant Measures of Image Features From Phase Information*. PhD thesis, University of Western Australia, 1996.

[8] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge Nonlinear Science Series, 2001.

[9] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *NIPS*, 2001.

[10] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Pittsburgh, PA, April 1991.