AUDIO-VISUAL SYNCHRONIZATION RECOVERY IN MULTIMEDIA CONTENT

Jong-Seok Lee and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{jong-seok.lee, touradj.ebrahimi}@epfl.ch

ABSTRACT

This paper proposes a method recovering audio-visual synchronization of multimedia content. It exploits the correlation between the acoustic and the visual signals in order to estimate the audio-visual drift existing in the content. By shifting the audio signal relative to the visual signal, the estimation of the drift is obtained by searching for the shift producing the maximal audio-visual correlation. We consider two correlation measures, namely, mutual information and canonical correlation, and compare their performance. Experimental results demonstrate that the method using the canonical correlation is effective in recovering the audio-visual synchronization for both speech and non-speech sequences.

Index Terms— Audio-visual synchronization, mutual information, canonical correlation, multimedia

1. INTRODUCTION

The synchronization of the acoustic and the visual signals is one of the most important factors affecting the quality of experience of multimedia content. It has been shown that poor audio-visual synchronization causes significant degradation in perceived quality by human observers and even deteriorates intelligibility of the content.

However, errors in audio-visual synchronization sometimes occur and corrupt the content in multimedia applications. It may be caused during acquisition, editing, processing or network transfer of the content. For example, when a microphone is placed far from the sound source during recording, the difference of the speeds of the sound and the light may cause audio-visual desynchronization in the recorded data. Different processing time and different network transfer delay of the two signals may also cause desynchronization between them. Such audio-visual drifts may be even accumulated during various stages to produce the final content from the recorded material. Therefore, in order to enhance the quality of experience of the multimedia content, it is desirable to detect the temporal misalignment between the two signals and recover correct synchronization between them. While there are efforts to avoid such temporal misalignments by specifications that preserve synchronization during multimedia processing, e.g. the presentation time stamp in MPEG-2, recovering synchronization of the desynchronized content has been rarely attempted.

Recently, a method has been proposed to estimate the audio-visual drift in a semi-automatic way [1]. This approach has been shown to perform successfully for some selected drift conditions. However, it has a limitation that it works only for video clips containing talking heads and the mouth region needs to be located prior to the drift estimation.

This paper proposes a novel method for recovering the audio-visual synchronization of a given desynchronized multimedia content. The method analyzes the correlation between the visual motion information in the scene and the dynamics of the audio signal that is temporally shifted. Two measures of the correlation are considered, namely, mutual information and canonical correlation, which are compared experimentally. It is shown that the method using the canonical correlation produces successful drift estimation results without necessity of adjustment of any content-dependent algorithm parameter. The proposed method does not have any assumption on the sound-emitting object and thus, is applicable to both speech and non-speech contents.

2. PROPOSED METHOD

Given an audio-visual sequence having an unknown drift, the proposed method begins with extracting features from the acoustic and the visual signals. The two features must have the same frame rate for further analysis described below. Typically, the audio features are extracted at the rate of the visual frame rate. Then, the correlation analysis is performed for the audio feature stream shifted by t frames and the visual feature stream. The correlation measures between them are calculated by varying the value of t within a pre-defined search range $[-T, T]$, among which the maximum is selected.

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia), and the Swiss NCCR Interactive Multimodal Information Management (IM2).

For a fixed t, the whole sequence is divided into N_B (possibly overlapping) temporal blocks. It is assumed that the spatial location of the sound source is stationary within a block. In order to be robust to the motion of the sound source, the length of the blocks needs to be kept reasonably short. On the other hand, every block should contain a sufficient number of samples in order to properly perform correlation analysis of the samples. For each block, the correlation analysis is performed as follows. First, each image frame is divided into N_T small tiles in order to reduce the computational complexity of the algorithm. Then, we measure the correlation between the mean-normalized acoustic feature sequence for the current temporal block ($\mathbf{x}$) and the mean-normalized visual feature sequence for each tile ($\mathbf{y}$), where the mean normalization is performed in order to make the mean of each feature over time zero. It is expected to observe the maximum correlation when the two signals are time-synchronous. Two different measures are suggested to obtain the correlation, denoted by $C(t,i,j)$ for the i-th block and the j-th tile: mutual information (MI) and canonical correlation (CC).

The MI of two random variables is a quantity measuring the mutual dependence between them. In particular, we employ the quadratic mutual information (QMI) measure based on the Renyi's quadratic entropy combined with the Parzen's nonparametric probability distribution function (pdf) estimation, which allows us to easily examine the dependence of the two information sources directly from the given samples. The QMI measure of $\mathbf{x}$ and $\mathbf{y}$ is given by [2]

$$C_{MI}(t,i,j) = \log \frac{\int\int f_{XY}(\mathbf{x},\mathbf{y})^2 d\mathbf{x}d\mathbf{y} \int\int f_X(\mathbf{x})^2 f_Y(\mathbf{y})^2 d\mathbf{x}d\mathbf{y}}{(\int\int f_{XY}(\mathbf{x},\mathbf{y}) f_X(\mathbf{x}) f_Y(\mathbf{y}) d\mathbf{x}d\mathbf{y})^2}, \quad (1)$$

where $f_X(\mathbf{x})$ and $f_Y(\mathbf{y})$ are the marginal pdfs of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $f_{XY}(\mathbf{x},\mathbf{y})$ is their joint pdf. $C_{MI}(t,i,j)$ is nonnegative and becomes zero when $\mathbf{x}$ and $\mathbf{y}$ are independent. Here, the Parzen's pdf estimator with the spherical Gaussian kernel is used. Then, the pdfs in (1) are obtained by

$$f_X(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} G(\mathbf{x};\mathbf{x}_m,\sigma_x^2), \quad (2)$$

$$f_Y(\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} G(\mathbf{y};\mathbf{y}_m,\sigma_y^2), \quad (3)$$

$$f_{XY}(\mathbf{x},\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} G(\mathbf{x};\mathbf{x}_m,\sigma_x^2)G(\mathbf{y};\mathbf{y}_m,\sigma_y^2), \quad (4)$$

where M is the number of samples and $G(\cdot;\mathbf{z},\sigma^2)$ the Gaussian kernel having mean $\mathbf{z}$ and variance σ^2.

The canonical correlation analysis aims at finding the projection vectors by which the correlation of the projected data becomes maximal. Thus, the CC between $\mathbf{x}$ and $\mathbf{y}$ is obtained by

$$C_{CC}(t,i,j) = \max_{\mathbf{w}_x,\mathbf{w}_y} \frac{E\{(\mathbf{w}_x^T\mathbf{x})(\mathbf{w}_y^T\mathbf{y})\}}{\sqrt{E\{\mathbf{w}_x^T\mathbf{x}\}^2 E\{\mathbf{w}_y^T\mathbf{y}\}^2}}, \quad (5)$$

where $\mathbf{w}_x$ and $\mathbf{w}_y$ are the projection vectors. This maximization problem can be resolved by solving an eigenvalue problem [3]. It is necessary to keep the length of the feature sequences (i.e., the length of the temporal block) larger than any of the two feature dimensions so that the problem does not become an underdetermined equation but has a unique solution.

After obtaining the correlation measures for all tiles, the maximum value is stored, which is repeated for all temporal blocks. The collection of the maximum correlation values are averaged to obtain the final correlation measure for the current temporal shift being examined:

$$D(t) = \frac{1}{N_B}\sum_{i=1}^{N_B} \max_{1\le j\le N_T} C(t,i,j). \quad (6)$$

If the two signals become synchronized by a shift, the tile showing the maximum correlation is expected to be located at the sound-emitting region and the corresponding correlation value to be the largest among all t's. Therefore, the estimated audio-visual drift is obtained by

$$t^* = \arg \max_{-T\le t\le T} D(t). \quad (7)$$

Note that the above equation gives the drift estimation at the resolution of the visual frame rate. In order to refine the estimation at a finer level, the quadratic interpolation is used:

$$t_f^* = t^* + \frac{D(t^*-1) - D(t^*+1)}{2\{D(t^*-1) - 2D(t^*) + D(t^*+1)\}}, \quad (8)$$

which is finally converted to the dimension of time:

$$\tau_f^* = t_f^*/F_y, \quad (9)$$

where F_y is the visual frame rate. Therefore, in order to recover audio-visual synchronization of the given sequence, the audio signal is temporally shifted by $-\tau_f^*$.

It is worth mentioning that the proposed method using QMI can be considered as an extension of the method presented in [1]. While both methods use QMI for measuring audio-visual correlation, the proposed method can be used for both speech and non-speech data with reduced complexity, whereas the previous method is applicable only to speech data. In addition, the previous method requires the user's intervention of indicating the time interval to be examined, while the proposed method is fully automatic.

(a) Data #1 (b) Data #2 (c) Data #3

Fig. 1. Example frames of the test data.

3. EXPERIMENTS

3.1. Setup

In order to evaluate the proposed algorithm, we used three 10 second long audio-visual sequences shown in Fig. 1. Data #1 and Data #2 are from the CUAVE database [4]. In Data #1, a person speaks English digits singly. Data #2 contains a person pronouncing digits next to a silent person. The two data were recorded at the visual frame rate of 29.97 Hz with a resolution of 720×480 pixels and the acoustic frequency of 44.1 kHz. Data #2 is more challenging than Data #1 because in Data #2, while a person is speaking, the silent person also makes motions with his head, lips and blinking eyes; the corresponding regions in the image frames will produce nonzero correlation values that will compete with those from the region containing the speaker's mouth. Finally, we recorded Data #3 where a hand holding a pen beats a desk continually to make bumping sound. The visual component of the data has a resolution of 720×408 pixels and the frame rate of 25 Hz. The audio signal was captured at the rate of 48 kHz. This sequence was chosen to examine effectiveness of our method for non-speech data and test its general applicability. For each content, we generated asynchronous audio-visual test sequences with drift values of 0 ms, ±200 ms, ±400 ms, ±600 ms, ±800 ms and ±1000 ms.

For the visual features, the difference of the luminance component of two consecutive image frames was used. The acoustic features were obtained by calculating the temporal derivative of the log-normalized energy for the samples within a moving window (i.e. $\mathbf{x}$ is one-dimensional). In our case, they were extracted at the rate of 100 Hz by using a 25 ms-long window as in many applications of acoustic signal analysis [5], and then downsampled to match the visual frame rate. Each set of features was normalized so that the feature values range from 0 to 1.

In order to reduce the computational complexity, each image frame was resized to 1/16 of its original resolution. The image frames were divided into 4×4 tiles for analysis (i.e. the dimension of $\mathbf{y}$ is 16). We used temporal blocks containing 50 visual frames (i.e. $M = 50$) and the overlaps of 50% with the precedent and the subsequent ones. T was set to 1100 ms. To compute QMI, we chose σ_x=0.2 and σ_y=0.5 experimentally.

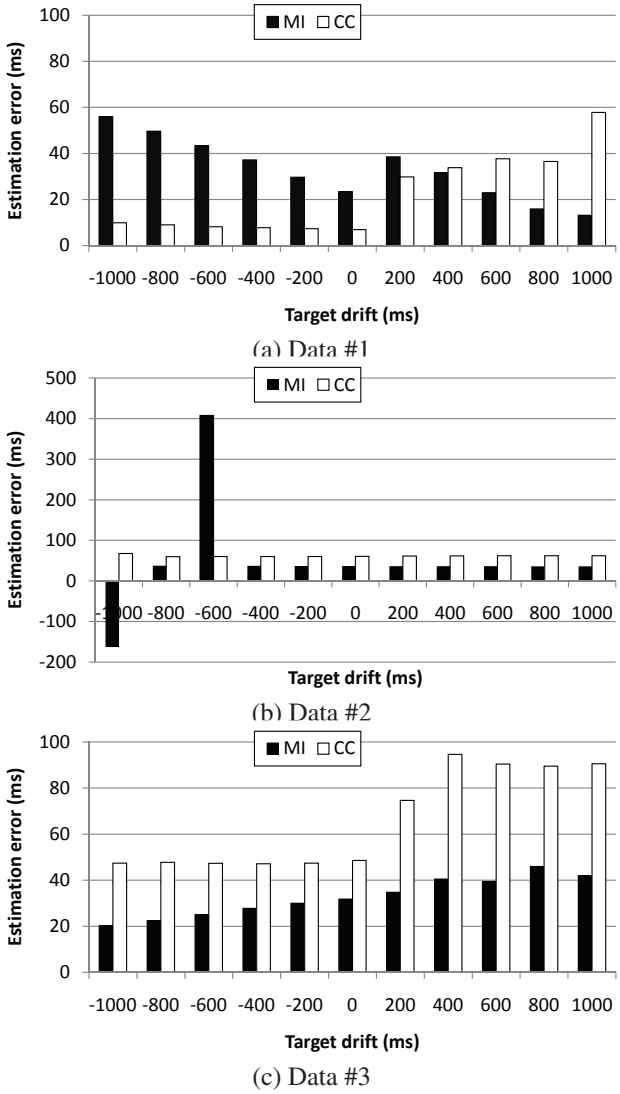

Fig. 2. Errors in drift estimation by the proposed method when MI or CC is used.

3.2. Results

Fig. 2 shows the drift estimation error for the three datasets. For Data #1 and Data #3, the two correlation measures perform well similarly. However, MI shows two cases with large errors in Data #2, whereas CC still produces good results. This is mainly due to difficulty in selecting the values of σ_x and σ_y. We observed that the performance varies with their values. Especially, Data #2, which is the most difficult for drift estimation due to the distracting motion by the silent person, was the most sensitive to their values among the three datasets. On the other hand, there is no such parameter to be carefully tuned in the proposed method using CC.

One can observe that the estimation errors when CC is used are less than 100 ms in Fig. 2, which we consider as

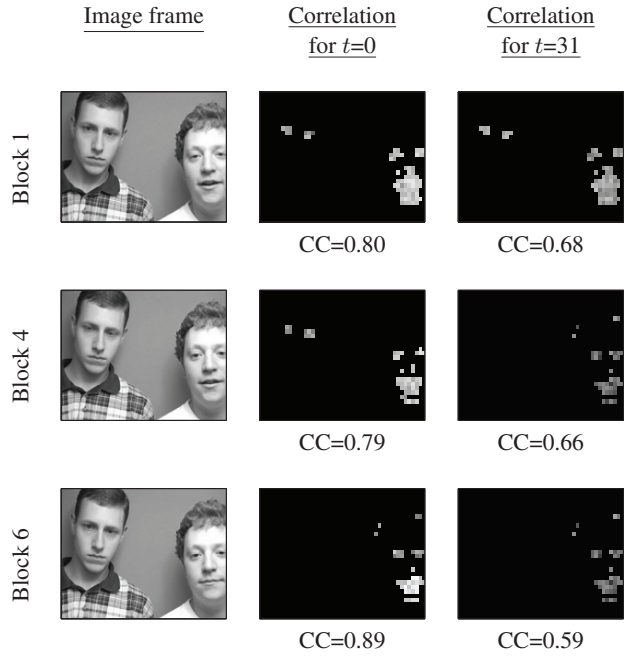

Fig. 3. Comparison of the results with correct and incorrect hypotheses (t=0 and t=31, respectively) for the perfectly synchronized version of Data #2. A bright pixel indicates a high correlation value for the corresponding location. The tiles that are not considered for correlation calculation due to little motion are marked with black color. The calculated CC values for the blocks, $\max_j C(t, i, j)$, are also shown.

successful drift estimation. It has been shown that there exists an "intersensory synchrony window" during which the performance of the human perception is not affected for desynchronized audio-visual events [6]. This window of asynchrony tolerance typically ranges up to 200 ms [7]. Based on such observations, the standard [8] concluded that the acceptability thresholds of audio-visual synchronization errors are +90 ms and -125 ms. Therefore, we conclude that the estimation errors obtained by our method are acceptable.

Fig. 3 compares the results of the correctly hypothesized drift and an incorrectly hypothesized one for Data #2 when CC is used. It is observed that the correlation values for the speaking person's mouth region are larger for the correct hypothesis than for the incorrect one consistently over different temporal blocks. In addition, the difference between the correlations of the mouth region of the speaker and the eye region of the silent person is small when the hypothesis is wrong, whereas the difference is notably large for the correct hypothesis.

4. CONCLUSION

We have proposed an automatic audio-visual drift estimation method that is applicable to both speech and non-speech sequences. We have designed the method to find the optimal shift of the audio signal that maximizes the audio-visual correlation in terms of MI or CC. The experimental results show that the method using CC can successfully recover the audio-visual synchronization within an acceptable error bound without need of adjusting algorithm parameters nor manual interaction.

In our future work, we will test the proposed method for sequences containing global motion and more complex local audio-visual activities, where a global motion compensation technique and more elaborate acoustic and visual features may be required, respectively.

5. REFERENCES

[1] Y. Liu and Y. Sato, "Recovering audio-to-video synchronization by audiovisual correlation analysis," in *Proc. ICPR*, 2008, pp. 1–4.

[2] R. A. Morejon and J. C. Principe, "Advanced search algorithms for information-theoretic learning with kernel-based estimators," *IEEE Trans. Neural Netw.*, vol. 15, no. 4, pp. 874–884, 2004.

[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Comput.*, vol. 16, pp. 2639–2664, 2004.

[4] E. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: a new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP*, 2002, pp. 2017–2020.

[5] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, New Jersey: Prentice Hall, 2001.

[6] B. Conrey and D. B. Pisoni, "Auditory-visual speech perception and synchrony detection for speech and non-speech signals," *J. Acoust. Soc. Amer.*, vol. 119, no. 6, pp. 4065–4073, 2006.

[7] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, pp. 598–607, 2007.

[8] Recommendation ITU-R BT.1359-1, "Relative timing of sound and vision for broadcasting," 1998.