WATCH, LISTEN ONCE, AND SYNC
audio-visual synchronization with multi-modal regression CNN

Toshiki Kikuchi and Yuko Ozasa
Keio University, Japan
Background

- The audio-visual synchronization drifts in some videos uploaded to video hosting/social networking services.
- The viewers evaluate the video contents more negatively when the synchronization drifts [Reeves & Voelker, 1993].
- Recovering audio-visual synchronization is an important task in the field of visual speech processing.
- Our work focuses on recovering audio-visual synchronization of single-person speech videos.

Previous Work

Recovering AV-sync using QMI [Liu&Sato, 2010]
computes QMI between audio and visual features, and uses it as the correlation value to determine whether the audio and video are synchronized correctly or not.

Weakness of the Previous Work

a. Feature Extraction
They use the vertical optical flows of speaking lip image sequences as visual feature.
→ Lack of Robustness

b. Searching Approach
They have to shift by all possible synchronized position using the sliding window to find the correct shift position.
→ Worse Computational Complexity
Our Contributions

1. We propose a novel architecture for recovering audio-visual synchronization using a CNN → Addressing Feature Extraction problem

2. We show the benefit of treating audio-visual synchronization as a regression problem → Addressing Searching Approach problem
AV-sync Approaches

• The sliding window approach

Video Seq.  
Audio Seq.  
Corr.

• The regression approach

Video Seq.  
Audio Seq.

Predict the correct position and sync!
Summary of Our Work

• We propose a multi-modal regression CNN for audio-visual synchronization for single-person speech videos. We call the proposed model WLOS (Watch, Listen Once, and Sync).

\[
\begin{align*}
n\text{-frame Audio Feature} & \quad \rightarrow \quad \text{WLOS} \\
n\text{-frame Visual Feature} & \quad \rightarrow \quad \text{drifted frame number}
\end{align*}
\]

• We also show experimental results that demonstrate the proposed method outperforms baseline methods.
Visual Input Representation

① Detect facial landmarks by Kazemi and Sullivan’s method [3] for each video frame
② Facial alignment with Affine Transform
③ Extract lip area of spatial resolution $32 \times 32$
④ Compute optical flows with Gunnar Farnebäck method [4]

Audio Input Representation

We use Mel-Frequency Cepstrum Coefficients (MFCCs) [Davis & Mermelstein, 1990] as the audio feature

① Apply the hamming window whose size is 256
② Compute the 13 MFCCs and use 12 MFCCs except the very first MFCC which is not informative about the actual spectral content

Network Architecture of WLOS

Drifted frame number $d$

Fusion Network

Visual Network

Audio Network

Optical Flow

MFCC
Fine-tuning

We use pre-trained weights as the initial weights of the visual network and the audio network.

The network for pre-training

Train to predict whether the audio and visual information are correlated or not (binary classification problem)

We call this “classification correlation CNN (C³)”

\[ n\text{-frame Audio Feature} \xrightarrow{\text{C³}} \text{Correlated or not} \]

\[ n\text{-frame Visual Feature} \xrightarrow{\text{C³}} \text{Correlated or not} \]
Network Architecture of C³

Correlated or not

Fusion Network:
- FC / Softmax (2)
- FC / ReLU (1024)

Visual Network:
- FC / ReLU (1024)
- 3D Spatial Dropout
- 3D MaxPool
- 3D Conv / BN / ReLU
- 3D Spatial Dropout
- 3D MaxPool
- 3D Conv / ReLU
- Visual Input (10×32×32×1)

Audio Network:
- FC / ReLU (1024)
- 2D MaxPool
- 2D Conv / BN / ReLU
- 2D MaxPool
- 2D Conv / ReLU
- Audio Input (39×12×1)

Optical flow

MFCC
Experiment – Dataset

Overview

• Intentionally drifted audio-visual dataset from GRID Corpus [Cooke et al., 2006]
• The audio-visual pairs including −9- to +9-frame drifts

Details

• 63739 pairs for S1, 64000 pairs for S2
  63992 pairs for S4, 63997 pairs for S7
• 80%:20% random split for train and test

Experiment – Baseline Methods

a. **QMI**
   Sliding Window Search with QMI
   (based on [Liu&Sato, 2010])

Experiment – Baseline Methods

a. QMI
   Sliding Window Search with QMI
   (based on [Liu&Sato, 2010])

b. C³
   Sliding Window Search with C³
Experiment – Baseline Methods

a. **QMI**
   Sliding Window Search with QMI (based on [Liu&Sato, 2010])

b. **C³**
   Sliding Window Search with C³

c. **WLOS (scratch)**
   WLOS without pre-training.

Experiment – Baseline Methods

a. **QMI**
   Sliding Window Search with QMI (based on [Liu & Sato, 2010])

b. **C³**
   Sliding Window Search with C³

c. **WLOS (scratch)**
   WLOS without pre-training.

---

Result – **Accuracy Evaluation**

- CNN-based method outperform QMI baseline
- Positive effect of pre-training

**Table 1.** Mean absolute error on Test Data (frame).

<table>
<thead>
<tr>
<th>Method</th>
<th>S1</th>
<th>S2</th>
<th>S4</th>
<th>S7</th>
</tr>
</thead>
<tbody>
<tr>
<td>C^3</td>
<td>1.352</td>
<td>2.134</td>
<td>2.983</td>
<td>1.019</td>
</tr>
<tr>
<td>WLOS (scratch)</td>
<td>0.937</td>
<td>1.003</td>
<td>1.116</td>
<td>0.848</td>
</tr>
<tr>
<td>WLOS (fine-tune)</td>
<td><strong>0.907</strong></td>
<td><strong>0.916</strong></td>
<td><strong>1.038</strong></td>
<td><strong>0.799</strong></td>
</tr>
</tbody>
</table>

Result – Computational Time

WLOS is approximately 19 times faster than $C^3$ because WLOS does not use a sliding window.

<table>
<thead>
<tr>
<th>Method</th>
<th>Approach</th>
<th>Time [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>QMI (based on [Liu&amp;Sato, 2010])</td>
<td>Sliding Window</td>
<td>46.07</td>
</tr>
<tr>
<td>$C^3$</td>
<td>Sliding Window</td>
<td>34.00</td>
</tr>
<tr>
<td>WLOS</td>
<td>Regression</td>
<td><strong>1.80</strong></td>
</tr>
</tbody>
</table>

Table 2. Processing time to synchronize a frame

Conclusion

• We proposed a multi-modal regression CNN for recovering audio-visual synchronization

• The proposed approach enables us to recover errors without searching with a sliding window which would increase computational cost

• Experimental results show that the proposed method performs better than the baseline methods

• WLOS is more accurate and faster!

• In future work, we will make it possible to correct audiovisual synchronization errors of general videos instead of speech videos