Mutual Alignment between Audiovisual Features for End-to-End Audiovisual Speech Recognition

Hong Liu, Yawei Wang, Bing Yang

Key Laboratory of Machine Perception
Shenzhen Graduate School, Peking University
Shenzhen, China

hongliu@pku.edu.cn, Wongyawei@pku.edu.cn, bingyang@sz.pku.edu.cn

Abstract—Asynchronization issue caused by different types of modalities is one of the major problems in audio visual speech recognition (AVSR) research. However, most AVSR systems merely rely on up sampling of video or down sampling of audio to align audio and visual features, assuming that the feature sequences are aligned frame-by-frame. These pre-processing steps oversimplify the asynchrony relation between acoustic signal and lip motion, lacking flexibility and impairing the performance of the system. Although there are systems modeling the asynchrony between the modalities, sometimes they fail to align speech and video precisely over some even all noisy conditions. In this paper, we propose a mutual feature alignment method for AVSR which can make full use of cross modality information to address the asynchrony issue by introducing Mutual Iterative Attention (MIA) mechanism. Our method can automatically learn an alignment in a mutual way by performing mutual attention iteratively between the audio and visual features, relying on the modified encoder structure of Transformer. Experimental results show that our proposed method obtains absolute improvements up to 20.42% over the audio modality alone depending upon the signal-to-noise-ratio (SNR) level. Better recognition performance can also be achieved comparing with the traditional feature concatenation method under both clean and noisy conditions. It is expectable that our proposed mutual feature alignment method can be easily generalized to other multimodal tasks with semantically correlated information.

Index Terms—multimodal alignment, audio visual speech recognition, mutual iterative attention

I. INTRODUCTION

With the rise of the Internet and the popularity of smartphones, artificial intelligence attracts a lot of attention in recent years. Consequently, intelligent human machine interaction interface becomes more and more important for smart devices. As speech is the most common and convenient way to communicate with others, automatic speech recognition (ASR) plays an indispensable role in effective human machine interaction. Although recent ASR systems, e.g., [1], [2], [3], can achieve the state-of-the-art recognition accuracy with the help of deep neural networks (DNNs) and its variations, there still remains a challenge for ASR robustness particularly under noisy conditions. Motivated by the fact that human speech perception is bimodal, i.e., humans understand speech not only by listening but also by considering visual cues of lips and face [4], audio visual speech recognition (AVSR) has gained significant development by many researchers over the past few decades. In AVSR research, visual modality is incorporated as the supplementary information to benefit speech intelligibility in noisy environment.

One of the major open problems that is often ignored in AVSR is to deal with asynchronization issue caused by different types of modalities. It is recognized that the visual and acoustic signals of speech do not necessarily occur at the same time and there is a natural asynchrony between audio and visual modalities in audio visual speech recognition [5]. Back to 1930s, anticipatory coarticulation, a naturally linguistic phenomenon of audio visual speech asynchrony, has been studied by Keating [6], the research showed that lip rounding during articulation of the consonant may be influenced by neighboring phones. In the speech recognition literature, Bergler and König [7] quantified the cross-modal mutual information between the acoustic feature and the visual feature under the assumption that visual speech precede acoustic signals with various temporal offsets, the results showed that the acoustic features are maximal correlated with visual features of 120 milliseconds on average.

Actually, the asynchrony relation between audio and visual modalities is uncertain. For some phones the modalities are synchronized, for other acoustic signals precede visual speech, and for other visual speech precedes acoustic signals [8], [9], [10]. The fact reveals the necessity of alignment between audio and visual modalities. However, the audio and visual features are assumed to be synchronized in most studies. Since their sampling rates are usually different, up sampling of video or down sampling of audio is performed to make the feature lengths identical as an alignment process. This pre-processing step is vital to align audio visual features but this approach oversimplify the synchronization issue and does not provide enough flexibility to capture the asynchrony relation between modalities for certain phones.

There are systems modeling the asynchrony between the modalities in AVSR research, such as coupled hidden Markov model (CHMM) [11], AV Align [12] and AliNN [13]. In CHMM, the state transitions in each modality depend on the state of the other modality. Nevertheless, the asynchrony modeling is limited only within the boundaries of each phone or viseme while observed asynchrony often crosses multiple phonemes. Both AV Align and AliNN model adopt attention mechanism within sequence-to-sequence architecture to learn an alignment between audio and visual modalities. In AV
Align model, sometimes speech can not align precisely to video when it is corrupted by noise so that the performance of AV Align model degrades over most even all noisy conditions comparing to single modality system [12], [14]. AliNN model can also see a decline when applied to Gaussian mixture model with hidden Markov model (GMM-HMM) back-end system in tablet conditions with noise. Recently, the Mutual Iterative Attention (MIA) [15] has been very successful in solving the fine-grained representations of the image task in vision-and-language grounding problems. The MIA is built to align the visual features and textual concepts of an image with their relevant counterparts in each domain. The idea of using MIA as an alignment model is to iteratively perform mutual attention between two domains to link up a feature in one domain by the correlated features in the other domain. By aligning two modalities, the Mutual Iterative Attention can capture the semantic-grounded image representations and show great power on the image captioning and visual question answering tasks.

In this work, we will address the asynchronization issue in AVSR by introducing Mutual Iterative Attention mechanism, which aims to align the audiovisual features by iteratively performing mutual attention between audio and visual modalities. The key contribution of our approach is that we adopt a mutual feature alignment method where the features from the other modality are utilized as the guide for aligning the features of current modality iteratively to make full use of cross modality information in the process of alignment. The framework relies on the modified encoder structure of Transformer [16]. The MIA is incorporated to generate new aligned audio and visual sequences which can be taken as input features directly for back-end AVSR systems. We evaluate the use of Mutual Iterative Attention mechanism in the feature concatenation based model architecture proposed in [17]. Experimental results demonstrate that our Mutual Iterative Attention equipped feature concatenation based end-to-end AVSR system can outperform the audio-only system up to 20.42% absolute improvement in -5dB SNR, showing the effectiveness of our proposal.

II. MUTUAL ITERATIVE ATTENTION

The Mutual Iterative Attention acts on plainly audio and visual features extracted from the front-end AVSR systems. It performs mutual attention iteratively between audio and visual features and generates new aligned audiovisual sequences. Fig. 1 gives an overview of MIA. In the following, the input audio and visual features are denoted as \( A \) and \( V \), respectively.

A. Positional Encoding

Since both input audio and visual features contain abundant sequential information, in order for the module to make use of the order of the sequence, some information about the relative or absolute position of the input features must be injected into the sequence. Thus, we add "positional encodings" to the input features at the bottoms of the MIA. This is a difference from the original MIA whose inputs are visual features and textual concepts of a single image which have no sequential information. The positional encodings have the same dimension \( d_{\text{input}} \) as the input features, so that the two can be added. In this work, we use sine and cosine functions of different frequencies as positional encodings:

\[
PE_{\text{pos}(2i)} = \sin \left( \frac{\text{pos}/10000^{2i}/d_{\text{input}}}{k} \right),
\]

\[
PE_{\text{pos}(2i+1)} = \cos \left( \frac{\text{pos}/10000^{2i}/d_{\text{input}}}{k} \right)
\]

(1)

where \( \text{pos} \) is the position and \( i \) is the dimension.

B. Mutual Attention

Mutual attention contains two sub-layers. The first sub-layer makes use of multi-head attention to learn the alignment in one modality by querying the features from the other modality. The second sub-layer uses a fully connected feed-forward network to add sufficient expressive power.

The multi-head attention is composed of \( k \) heads in parallel. Each head is formulated as a scaled dot-product attention:

\[
\text{Att}_i(Q, S) = \text{softmax} \left( \frac{QW^Q(SW^K_i)^T}{\sqrt{d_k}} \right) SW^V_i,
\]

(2)

where \( Q \in \mathbb{R}^{m \times d_{\text{input}}} \) and \( S \in \mathbb{R}^{n \times d_{\text{input}}} \) denotes \( m \) querying features and \( n \) source features; \( W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{input}} \times d_k} \) are parameters that require to be learned of linear transformations; \( d_k = d_h/k \) is the size of the output features for each attention.
head. Results from each head are concatenated and passed through a linear transformation to generate the output:

$$\text{MultiHead}(Q, S) = [\text{Att}_1(Q, S), \ldots, \text{Att}_k(Q, S)] W^O$$  \hspace{1cm} (3)

where $W^O \in \mathbb{R}^{d_{aux} \times d_{appr}}$ is a learnable parameter. The multi-head attention integrates $n$ source features into $m$ output features in the order of querying features. Here we keep $m$ the same as $n$ for simplified computation.

The fully-connected feed-forward network following the multi-head attention is defined as:

$$\text{FFN}(X) = \max(0, XW_1 + b_1) W_2 + b_2$$  \hspace{1cm} (4)

where $\max(0, \cdot)$ is the ReLU activation function; $W_1$ and $W_2$ are matrices for linear transformation; $b_1$ and $b_2$ are corresponding bias terms. Each sub-layer is followed by an operation sequence of dropout [18], shortcut connection [19], and layer normalization [20].

Finally, the mutual attention is formulated as:

$$A' = \text{FFN}(\text{MultiHead}(V, A)),$$
$$V' = \text{FFN}(\text{MultiHead}(A', V))$$  \hspace{1cm} (5)

Since the information from the other modality only serves as the attentive weight, the output features only contain homogeneous information. The information from either modality can serve as the guide for aligning the two kinds of features, because for the same position in the two feature matrices, the integrated audio features and the integrated visual features are co-referential and represent the same high-level speech semantics.

C. Mutual Iterative Attention

Mutual attention is performed iteratively in order to better refine both the audio and visual features. The process in (5) which takes the original features as input is considered to be the first round:

$$A_1 = \text{FFN}(\text{MultiHead}(V_0, A_0)),$$
$$V_1 = \text{FFN}(\text{MultiHead}(A_1, V_0))$$  \hspace{1cm} (6)

where $A_0$, $V_0$, $A_1$ and $V_1$ stand for the original audio features, the original visual features, the macro audio features, and the macro visual features, respectively. By repeating the same process for $N$ times, the final outputs of the two stacks are obtained as:

$$A_N = \text{FFN}(\text{MultiHead}(V_{N-1}, A_{N-1})),$$
$$V_N = \text{FFN}(\text{MultiHead}(A_N, V_{N-1}))$$  \hspace{1cm} (7)

To avoid potential over-smoothing problem that all features contribute the same to the attentive weight if iterating too many times, an extra shortcut connection from the input of each layer (not the sub-layer) is applied to the output. It is also worth emphasizing that the parameters of the mutual attention are shared in each iteration.

III. END-TO-END MUTUAL ITERATIVE ATTENTION AVSR

In this section, a detailed introduction of the end-to-end AVSR model with Mutual Iterative Attention is given. The proposed audio visual speech recognition system is shown in Fig. 2. It consists of audio and visual front ends which attend over raw image sequence and original audio signal respectively in order to extract corresponding features. Each front end is composed of a residual network (ResNet) [19] which learns to extract features automatically and a 2-layer BGRU which models the temporal dynamics of features in each domain. MIA module is applied to align the audiovisual features by performing mutual attention iteratively between audio and visual modalities. Finally, another 2-layer BGRU is used to fuse the information of new aligned audio and visual sequences.

A. Audio Front End

The audio front end consists of a temporal convolutional layer followed by a 18-layer ResNet and a 2-layer BGRU. Due to the audio signal is one-dimensional, both the temporal convolutional layer and the ResNet-18 apply 1D kernels instead of 2D. The temporal convolutional layer uses a 5ms temporal kernel with a stride of 0.25ms to extract fine-scale spectral information. The 18-layer ResNet, which consists of kernels of size 3 by 1, is allowed to extract long-term speech characteristics. The average pooling is applied in ResNet-18 to keep audio features have the same length as visual features.
which contain 29 frames. Finally, a 2-layer BGRU which consists of 1024 cells in each layer is followed to model the audio temporal dynamics.

B. Visual Front End

The visual front end is composed of a spatiotemporal convolution, a 34-layer ResNet and a 2-layer BGRU. The spatiotemporal convolutional layer is utilized to capture the short-term dynamics of the lip region. It consists of a 3D convolutional layer with 64 kernels whose size is 5 by 7 by 7, followed by batch normalization, rectified linear units and max pooling. The 34-layer ResNet extracts visual features further and compress the visual information into a single dimensional tensor per time step. Finally, one same 2-layer BGRU as the audio front end is followed to model the visual temporal dynamics.

C. Mutual Iterative Attention Module

The detailed structure of Mutual Iterative Attention is shown as Fig. 1. MIA takes the audio and visual features from previous front ends as input. Positional encodings are firstly added to the input features to make use of the order information of the sequences. The $d_{seq}$ in positional encodings have the same dimension 1024 as the input audio and visual features. Following mutual attention contains a multi-head attention layer and a fully connected feed-forward layer. The multi-head attention is composed of 8 parallel heads. We use multi-head attention to learn the alignment in a certain domain by querying the features from the other domain. The fully connected feed-forward network is applied to add sufficient expressive power. Considering that a single iteration does not suffice to align audio and visual features and over-smoothing problem may occur if iterating too many times, we perform mutual attention twice where we set $N = 2$ to better refine the audio and visual features. In this work, audio features are first aligned according to visual features, and then visual features are aligned according to integrated audio features. It is worth noticing that it is also possible to reverse the order by first aligning visual features by querying audio features. Here we followed the same assumption proposed in [12] that acoustic encodings corrupted by noise can be partially corrected or even reconstructed from visual encodings.

D. Classification Layers

New aligned audio and visual features generated from Mutual Iterative Attention module are concatenated per frame for fusing the information from the audio and visual streams in the feature level. Concatenated features are then fed to a 2-layer BGRU to model their temporal dynamics jointly. Finally a softmax layer is followed to provide a label.

IV. EXPERIMENTS

A. Dataset

We apply our audio visual speech recognition with Mutual Iterative Attention method in the context of Lip Reading in the Wild (LRW) dataset [21] which is a publicly available audio visual speech recognition dataset collected from in-the-wild videos. All videos are extracted from British television programs like news and current affairs debate which have a change set of talking head, and each video contains 29 frames (1.16 seconds) in length. While existing audio visual datasets for word recognition merely contain 10 to 50 words, the LRW dataset is much higher than them since it consists of up to 1000 utterances of 500 different words, spoken by more than 1000 different speakers. It is also challenging due to its large variation in head pose and illumination condition.

It should be emphasized that the target word occurs in the middle of video. Although visual modality has a lot more to benefit from sentence-level context instead of isolated words as longer contexts are critical to disambiguate homophones (e.g. ‘bad’, ’bat’, ‘pat’, ‘mat’, etc. are all visually identical), there may be co-articulation of the lips from preceding and subsequent words. Another challenge of the dataset is that there are still several words which are visually similar despite sentence-level context. For example, there are words presenting in their singular and plural forms or simply different forms of the same word, e.g., thing and things, America and American.

B. Preprocessing

Zero-mean normalization is applied in every original audio signal before being fed to audio front end, in other words, each audio segment has zero mean and standard deviation one to account for variations in different levels of loudness between the speakers. The visual front-end processing contains two steps. The first step is the extraction of the lip region of interest (ROI). Since the centre of the lip region is already located, a fixed bounding box of 96 by 96 is used for all image sequences. The second step is lip ROI post-processing where the lip ROIs are transformed to grayscale and are normalized with respect to the overall mean and variance.

C. Data argumentation

Babble noise at different levels is added into the original audio signals during training. The SNR levels range from -5dB to 20dB with an interval of 5dB. For the sake of enhancing robustness to different noise levels, one of the noise levels or the clean signal is selected using a uniform distribution. During training, two data augmentation methodologies are performed in raw image sequence, random cropping and horizontal flipping. Specifically, each lip ROI is randomly cropped to a size of 88 by 88. During testing, the central patch is cropped. Horizontal flipping with a probability of 50% is used to increase the variation on training samples.

D. Training

We train several uni- and bimodal end-to-end deep learning systems. The unimodal networks process either the audio input or the visual input. The audio front end in Fig. 2 is one unimodal network which attend over original audio signal both in clean form and additively corrupted by six SNR types. Another unimodal network is the visual front end which processes only row image sequence. The bimodal networks take audio signal
and raw images as input at the same time. One bimodal system is the feature concatenation based AVSR model [17], the structure of which is the framework shown in Fig. 2 without Mutual Iterative Attention module. In feature concatenation based model, the outputs of the audio front end and visual front end are assumed to be aligned frame-by-frame, so the two feature sequences are concatenated directly and then fed into classification layers. Our proposed Mutual Iterative Attention AVSR system is another bimodal network. All systems train end-to-end directly without a 3-step procedure as in [17]. We compare our proposed method (AV_MIA) with the audio-only baseline, the visual-only system and the feature concatenation based AVSR system (AV_baseline).

We perform classification of 500 words from the LRW dataset. The videos in the dataset are already partitioned into training, validation and test sets. For each word there are between 800 and 1000 sequences in the training set, 50 sequences in the validation and 50 in the test sets. Totally there are 488766, 25000, and 25000 examples in the training, validation and test sets, respectively. We measure the speech recognition performance in terms of the word classification rate (CR). In training, we directly optimize the cross entropy loss between ground-truth label and predicted label via the Adam optimizer [22].

We first train audio front end and visual front end for single stream training using a softmax output layer from scratch. An initial learning rate of 0.0003 and a mini-batch of 36 are used for each front end training. As for AV_baseline model, the Adam training algorithm is used for end-to-end training with a mini-batch size of 18 sequences and an initial learning rate of 0.0003. Finally, our proposed AV_MIA model is trained end-to-end with a mini-batch size of 18 sequences and an initial learning rate of 0.0000375. All systems are trained until there is no improvement in the classification rate on the validation set for more than 5 epochs.

E. Results and Discussion

Results are shown in Table I. We report the word recognition performance of the end-to-end audio-only, visual-only, feature concatenation based audiovisual (AV_baseline) and Mutual Iterative Attention audiovisual (AV_MIA) systems. In order to investigate the robustness to audio noise of audiovisual systems, experiments are performed under different noise levels. We add babble noise from the NOISEX database [23] to original audio waveforms with one of the 6 SNR types randomly. Thus with this multi-condition data, we train the audio-only, AV_baseline as well as AV_MIA system, and test them on different SNR test sets.

Since the visual information is not influenced by the addiction of the acoustic noise, the performance of visual-only system maintains a constant over all noisy conditions. On the other hand, the performance of audio-only system drops significantly along with the decent of SNR as expected. Although visual-only system performs much worse than audio-only system in most conditions, it is helpful to integrate visual information into audiovisual systems. At first glance, visual information brings improvement over both AV_baseline and AV_MIA. The contribution of video depends on the SNR of audio signal, for example, AV_baseline improves audio-only system from 68.90% CR to 87.17% CR in -5dB condition which is a absolute 18.27% performance gain. Obviously, supplementary information from the visual modality is more helpful in low SNR conditions, we need to pay more attention to audio streams at high SNR.

Our proposed AV_MIA system performs best under both clean and noisy conditions. We notice absolute improvements starting at 0.86% at 20db SNR (96.68% CR upwards to 97.54%), up to 20.42% at -5db SNR (68.90% CR upwards to 89.32%) using our AV_MIA model over the audio-only model. When increasing noise level, it shows an increased advantage and achieves a absolute performance improvement up to 2.15% compared with AV_baseline system where the audio and visual sequences generated from previous front ends are assumed to be aligned frame-by-frame, showing the effectiveness of our mutual feature alignment method for audio visual speech recognition. Interestingly, we find that AV_MIA system achieves higher CR in clean condition than AV_baseline system although audio and visual features are still equally treated. We have to attribute the success to our proposed mutual feature alignment method because the only difference between the two systems is whether modelling an alignment or not.

AV_MIA model is able to pay attention to integrated features instead of the separate audio features or visual features. Mutual attention serves as a way to integrate correlated features by aligning audio and visual modalities, which is our main purpose. As expected the Mutual Iterative Attention based audiovisual model has a higher classification rate compared with audio-only, visual-only and audiovisual baseline across all noise levels. It is worth pointing out again that when an alignment between audiovisual modalities is properly modelled then a better performance is achieved.

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>90.74</td>
</tr>
<tr>
<td>20dB</td>
<td>96.68</td>
</tr>
<tr>
<td>15dB</td>
<td>96.48</td>
</tr>
<tr>
<td>10dB</td>
<td>95.85</td>
</tr>
<tr>
<td>5dB</td>
<td>94.01</td>
</tr>
<tr>
<td>0dB</td>
<td>88.07</td>
</tr>
<tr>
<td>-5dB</td>
<td>68.90</td>
</tr>
<tr>
<td>AVG</td>
<td>89.32</td>
</tr>
</tbody>
</table>

5352
V. CONCLUSION

In this work, a mutual feature alignment method for audio visual speech recognition is proposed. We introduce Mutual Iterative Attention (MIA) mechanism, which relies on the modified encoder structure of Transformer, to automatically learn an alignment between audio and visual modalities in a mutual way without mixing in the heterogeneous information. The MIA module aligns the audio and visual features by performing mutual attention over the two modalities in an iterative way and generates new aligned audio and visual sequences for back-end AVSR system. Experimental results show that our proposed method obtains a 20.42% absolute improvement comparing with audio-only system in -5dB SNR and outperforms the feature concatenation based AVSR system over all noisy conditions. Since our method is able to capture asynchrony relation between audio and visual modalities, we anticipate that it is easy to generalize the method to other multimodal tasks where the input modalities are semantically correlated.

There are many directions to extend this work. A natural next step would be able to recognize sentences instead of isolated words. Moreover, it would be interesting to explore different fusion mechanisms for audio visual speech recognition system in future works.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No.61673030, U1613209), National Natural Science Foundation of Shenzhen (No.JCYJ20190808182209321)

REFERENCES