LoCoNet: Long-Short Context Network for Active Speaker Detection

Xizi Wang¹  Feng Cheng²  Gedas Bertasius²
¹Indiana University  ²UNC Chapel Hill
xiziwang@iu.edu  {fengchan,gedas}@cs.unc.edu

Abstract

Active Speaker Detection (ASD) aims to identify who is speaking in each frame of a video. Solving ASD involves using audio and visual information in two complementary contexts: long-term intra-speaker context models the temporal dependencies of the same speaker, while short-term inter-speaker context models the interactions of speakers in the same scene. Motivated by these observations, we propose LoCoNet, a simple but effective Long-Short Context Network that leverages Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM) in an interleaved manner. LIM employs self-attention for long-range temporal dependencies modeling and cross-attention for audio-visual interactions modeling. SIM incorporates convolutional blocks that capture local patterns for short-term inter-speaker context. Experiments show that LoCoNet achieves state-of-the-art performance on multiple datasets, with 95.2% (+0.3%) mAP on AVA-ActiveSpeaker, 97.2% (+2.7%) mAP on Talkies, and 68.4% (+7.7%) mAP on Ego4D. Moreover, in challenging cases where multiple speakers are present, LoCoNet outperforms previous state-of-the-art methods by 3.0% mAP on AVA-ActiveSpeaker. The code is available at https://github.com/SJTUwxz/LoCoNet_ASD.

1. Introduction

Real-world interactive computer vision systems need to recognize not only the physical properties of a scene, such as objects and people, but also the social properties, including how people interact with each other. One fundamental task is identifying, at any moment, who is speaking in a complex scene with multiple interacting individuals. This Active Speaker Detection (ASD) problem [2, 3, 11, 18, 24, 25, 29, 32, 35, 39, 40, 42, 53, 58, 59, 62, 65, 70] is crucial for many real-world applications like human-robot interaction [33, 61, 66], speech diarization [14, 16, 20, 21, 68, 73], video re-targeting [6, 44, 74], multimodal learning [4, 7, 17, 19, 28, 30, 34, 34, 46, 49, 50, 50, 54, 55, 55, 60, 63], etc.

How can we tell whether someone is speaking? Visual cues such as movements of the mouth and eyes, when correlated with audio signals, often offer direct and primary evidence. The inter-modality synchronization over a longer audio-visual segment also provides complementary information [2, 39, 48, 65]. The first row of Fig 2 shows how Long-term Intra-speaker Modeling helps discern this primary indicator by observing one person for a long time span.

However, in a complex video, a person’s face is often occluded, turned away, off-frame, or very small, posing challenges for directly inferring speaking activity from visual cues. Fortunately, valuable evidence about the target speaker can be gleaned from the behaviors of others in the scene [61]. The second row of Fig 2 shows an example: from the left m-frame video segment, even with a partially visible face of the man on the right, it is easy to see that he is speaking at \( T_{i+m} \), because the woman in the middle turns her head (\( T_{i+1} \rightarrow T_{i+m} \)) and neither of the other two people open their mouths at \( T_{i+m} \). Notably, the woman’s gaze towards the man on the right at \( T_{i+m} \) does not provide substantial information on whether the man is speaking at a distant time \( T_{i+1} \). Therefore, we argue that Inter-speaker Modeling is sufficient in short-term temporal windows, since activities
Long-term Intra-speaker Modeling (LIM), Short-term Inter-speaker Modeling (SIM), and comparison of LoCoNet with existing long-term parallel-inference ASD methods. Red boxes show inactive speakers and green boxes show active speakers. LIM uses the features of a single speaker across all frames to capture long-term relationships. SIM models the relationships of speakers within a short $m$-frame segment to capture the conversation pattern. The speaker context modeling of the existing long-term parallel-inference ASD methods [15, 45, 65] only focuses on LIM, while LoCoNet models LIM and SIM to learn both contexts.

To solve this per-frame video classification task, existing ASD methods can be put into two categories: 1) parallel-inference methods [15, 40, 45, 65, 75] take all frames as input and predict the results for all frames in one pass. 2) sequential-inference methods [2, 3, 39, 42, 48] take a short clip as input and only give prediction result of the center frame. Thus, a sliding window strategy is often adopted to get results for all the frames. As depicted in Fig. 1, the average-FLOPs of parallel-inference methods [15, 45, 65] to predict the speaking activity of one face crop are often much lower than sequential-inference methods [2, 3, 39, 42, 48]. However, most parallel-inference methods [15, 45, 65, 72] that take long video clips as input do not consider multi-speaker context, which could cause performance degradation as speakers’ interaction is crucial for ASD task.

With the above issues in mind, we propose LoCoNet, an end-to-end Long-Short Context Network. Long-term Intra-speaker Modeling (LIM) employs a self-attention mechanism [67] for long-range dependencies modeling and a cross-attention mechanism for audio-visual interactions. Short-term Inter-speaker Modeling (SIM) incorporates convolutional blocks to capture local conversational patterns. Moreover, while most ASD methods use vision backbones for audio encoding [2, 3, 42, 65] due to high temporal downsampling in most audio backbones [12, 23, 27, 38, 56], we propose VGGFrame to leverage pretrained AudioSet [22] weights to extract per-frame audio features. We also use a parallel inference strategy for more efficient video processing.

Our extensive experiments validate the effectiveness of our approach. On the AVA-ActiveSpeaker dataset [58], LoCoNet achieves 95.2% mAP, outperforming previous state-of-the-art method SPELL+ [48] by 0.3% with 38× less computational cost. Furthermore, LoCoNet achieves 97.2% (+2.7%) mAP on Talkies [42] and 68.4% (+7.7%) mAP on Ego4D’s Audio-Visual benchmark [24]. LoCoNet works especially well in challenging scenarios such as with multiple speakers or small speaker faces.

2. Related Work

Most recent techniques for ASD can be characterized in terms of three salient dimensions: frame-level processing strategy that determines the inference speed of the method, the extracted context information to enhance the feature representations for prediction, and the training mechanisms.

Frame-level processing strategy. Given a long video, existing ASD methods employ two main strategies for generating per-frame predictions. 1) Parallel-inference [15, 40, 45, 65]
takes all frames of the video as input and predict the per-frame results in one pass. Such methods are often fast, as depicted in Fig. 1. However, they typically do not consider interactions among multiple speakers. 2) Sequential-inference [2, 3, 39, 42, 48] predicts one frame by sampling a short clip centered around that frame. They need a sliding window strategy to produce predictions for all frames, resulting in slower inference speed. Our proposed LoCoNet adopts a parallel inference strategy, combining fast inference speed with effective consideration of speakers’ interactions.


In our proposed LoCoNet, we introduce Long-Short Context Modeling (LSCM), composed of Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM) in an interleaved manner. LIM captures long-term temporal dependencies with self-attention, and audio-visual interactions with cross-attention. SIM incorporates inter-speaker convolutional blocks to learn speakers’ interactions.

**Training mechanisms.** Training on long videos can be memory-intensive, prompting some prior work [2, 39, 42, 48] to adopt a multi-stage training mechanism. In this process, a short-term feature extractor is initially trained, and subsequently, a long-term context modeling network is trained on the features extracted by the pre-trained feature extractor. TalkNet [65], EASEE [3], UniCon [75], light-ASD [45], ASD-transformer [15], and ADENet [72] use end-to-end training, fully leveraging the learning capabilities of the model. Similarly, our proposed LoCoNet is also trained end-to-end, enabling joint optimization of audio-visual feature representation learning and context modeling.

### 3. LoCoNet

Given the stacked visual face track \( V \in \mathbb{R}^{S \times T \times H \times W \times 1} \) and the audio Mel-spectrograms \( A \in \mathbb{R}^{4T \times M} \), LoCoNet aims to predict the speaking activity \( \hat{R} \in \mathbb{R}^{T} \) of the target person in each frame. \( S \) is the number of speakers including the target speaker and \( S - 1 \) context speakers in the same scene. \( T \) is the temporal length of the face track. \( H \) and \( W \) are the height and width of each visual face crop. \( M \) is the number of frequency bins of the audio Mel-spectrograms.

As shown in Fig 3, LoCoNet consists of a visual encoder, an audio encoder, and a Long-Short Context Modeling (LSCM) module with \( N \) LSCM blocks. We explain each module in more detail below.

#### 3.1. Encoders

**Visual encoder.** Given the face crop track \( V_{i} \in \mathbb{R}^{T \times H \times W \times 1} \) of speaker \( v_{i} \), the visual encoder yields a time sequence of visual embeddings \( f_{v_{i}} \in \mathbb{R}^{T \times C}, i = 1, ..., S \). The stacked visual embeddings of the target speaker and all the sampled context speakers \( f_{v} \in \mathbb{R}^{S \times T \times C} \) represent temporal context
of each speaker independently.

**Audio Encoder.** The audio encoder takes audio Mel-spectrograms $a \in \mathbb{R}^{4T \times M}$ as input. We need frame-level audio features for per-frame classification, but most pretrained audio encoders [12, 23, 27, 56] are for audio classification and thus have a high degree of temporal downsampling. To solve this problem, we propose VGGFrame as the audio encoder, which can fully utilize the pretrained VGGish [27]. The architecture of VGGFrame is illustrated in Fig. 4. We remove the temporal downsampling layer after block-4 and add a deconvolutional layer to upsample the temporal dimension. We concatenate the intermediate features with the upsampled features to extract a hierarchy of representations. VGGFrame outputs the audio embeddings $f_a \in \mathbb{R}^{T \times C}$. To align with $S$ speakers, we repeat $f_a, S$ times to produce the audio embedding $f_a \in \mathbb{R}^{S \times T \times C}$.

### 3.2. Long-Short Context Modeling

The visual and audio embeddings, derived independently by the visual and audio encoders, lack consideration of intra/inter-speaker context. Our Long-Short Context Modeling (LSCM) is designed to enhance these embeddings by learning long-term intra-speaker and short-term inter-speaker context in an interleaved manner. As shown in Fig. 3, LSCM consists of $N$ blocks, each incorporating a Long-term Intra-speaker Modeling (LIM) module and a Short-term Inter-speaker Modeling (SIM) module consecutively. LIM limits the model to look at the same speaker across all frames, encouraging it to learn speaker-independent patterns from audio and visual interactions. In contrast, SIM constrains the model to examine all speakers in nearby frames and capture local interactions. Such inductive bias is instrumental in enhancing the model’s capacity to glean valuable insights from these contextual dimensions.

LSCM inputs audio embeddings $f_a$ and visual embeddings $f_v$, and produces context-aware embeddings $u_a^N, u_v^N \in \mathbb{R}^{S \times T \times C}$. These embeddings are concatenated to yield the final embeddings $u^N = concat(u_a^N, u_v^N)$. Via a linear layer, $u^N$ is used to predict the speaking activities $R \in \mathbb{R}^T$ of the target person. The computation process at each LSCM block $l$ is detailed below.

#### 3.2.1 Long-term Intra-speaker Modeling (LIM)

LIM consists of two submodules: (i) **Audio/Visual Self-Attention** models an individual person’s behavior over a longer time period, and (ii) **Audio/Visual Cross-Attention** learns the interaction between audio and visual embeddings.

**Audio/Visual Self-Attention.** Since the model needs large capacity and to learn long-term dependencies, we employ the attention mechanism [67] with a Transformer layer applied on the temporal dimension to achieve long-term modeling,

$$
\tilde{u}_v^l = \text{LN}(\text{MHA}(u_v^{l-1}, u_v^{l-1}, u_v^{l-1}) + u_v^{l-1}),
$$

$$
\tilde{u}_a^l = \text{LN}(\text{MLP}(\tilde{u}_v^l) + \tilde{u}_v^l),
$$

where $u_v^0 = f_v, u_v^0 = f_a, u_v^{l-1} \in \mathbb{R}^{S \times T \times C}$ are the output visual embeddings of the previous LSCM block, LN($\cdot$) denotes Layer Normalization [5], MHA($q, k, v$) is multi-head attention [67] with query $q$, key $k$, value $v$, and MLP is a multi-layer perceptron. Audio Self-Attention is applied in the same way to $u_a^{l-1}$ to obtain the audio output $\tilde{u}_a^l$.

**Audio-Visual Cross-Attention.** The visual and audio streams are so far processed separately. To enhance visual features with audio and vice versa, we use an Audio-Visual dual Cross-Attention,

$$
\tilde{u}_v^l = \text{LN}(\text{MHA}(\tilde{u}_v^l, \tilde{u}_a^l, \tilde{u}_a^l) + \tilde{u}_a^l),
$$

$$
\tilde{u}_a^l = \text{LN}(\text{MLP}(\tilde{u}_v^l) + \tilde{u}_v^l),
$$

where $\tilde{u}_v^l$ are audio-enhanced visual embeddings. Visual-enhanced audio embeddings $\tilde{u}_a^l$ are obtained in the same way via Audio-Visual Cross-Attention given $\tilde{u}_a^l$ and $\tilde{u}_v^l$.

#### 3.2.2 Short-term Inter-speaker Modeling (SIM)

For a given moment in the video, the speaking activity of a target person is more coordinated with other speakers in closer frames [57], so the model should capture local temporal inter-speaker relationships. To do this, we employ a small **Inter-speaker Convolutional Network.**

$$
u_a^l = \text{MLP}(\text{LN}(\text{Conv}_{s \times k}(\tilde{u}_a^l))) + \tilde{u}_a^l,
$$

$$
u_v^l = \text{MLP}(\text{LN}(\text{Conv}_{s \times k}(\tilde{u}_v^l))) + \tilde{u}_v^l,
$$

where visual embeddings $u_v^l \in \mathbb{R}^{S \times T \times C}$ and audio embeddings $u_a^l \in \mathbb{R}^{S \times T \times C}$ are the output of the $l$-th LSCM block that will be passed to the next block. $k$ is the temporal length of the receptive field and $s$ is the number of speakers considered. Explicitly modeling inter-speaker context in nearby frames enables cross-frame inter-speaker information exchange. Our SIM module with a short temporal receptive field can help capture local dynamic patterns in interactions.
3.3. Training and Inference
Following [1, 10, 64], we train our model with multiple supervisions utilizing \( u^N \in \mathbb{R}^{S \times T \times 2C} \) (Sec. 3.2) and \( u^i \) derived from each intermediate block of LSCM. For each \( u^i \), a fully-connected (FC) layer is applied, yielding prediction results \( R^i \in \mathbb{R}^7 \) corresponding to the target speaker for each frame. All FC layers share their parameters. The overall loss function is \( L = \sum_{i=1}^{N} \text{CrossEntropy}(R^i, R) \) where \( R \) is the ground-truth. During inference, we can reduce the computation by a factor of \( S \) by reusing speakers’ features extracted by the visual encoder, which uses the most FLOPs.

3.4. Implementation details
Following [65], our visual encoder consists of a 3D convolutional layer, ResNet-18 [26], and a visual temporal convolution network (V-TCN) [41]. Our audio encoder is the proposed VGGFrame initialized with VGGish [12] weights pretrained on AudioSet [22]. We sample \( S = 3 \) speakers and \( T = 200 \) frames. In SIM, \( s \) is set to 3, which is the same as \( S \), and \( k \) is set to 7 frames. The face crops are resized to \( 112 \times 112 \). Visual augmentation includes randomly resized cropping, horizontal flipping, and rotation. For audio augmentation, another audio signal is randomly chosen from the rest of the training set and added as noise to the target audio. We train LoCoNet with Adam [36] for 25 epochs on 4 RTX6000 GPUs with batch size 4 using PyTorch [51]. The learning rate is \( 5 \times 10^{-5} \) and reduced by 5% each epoch.

4. Experimental Setup
4.1. Datasets
AVA-ActiveSpeaker [58] is a standard benchmark for ASD, consisting of 262 videos from Hollywood movies with 3.65 million frames and 5.3 million face crops. Following [3, 15, 31, 45, 47, 48], we evaluate on the validation set.

Talkies [42] is an in-the-wild ASD dataset with 23,507 face tracks extracted from 421,997 labeled frames. It focuses on challenging cases with more speakers, diverse actors and scenes, and more off-screen speech.

Ego4D [24]’s Audio-Visual benchmark has 3,670 hours of egocentric video of unscripted daily activities in many environments. It includes many challenges that complement egocentric benchmarks, including unusual viewpoints and speakers who are off-screen.

4.2. Evaluation Metric
Following [2, 3, 39, 42, 65], we use the official ActivityNet [9] evaluation tool to compute mean average precision (mAP) and evaluate on the AVA-ActiveSpeaker [58]. We also compute AUC [8] as another evaluation metric using Sklearn [52]. We use mAP to evaluate on Talkies [42] and Ego4D [24].

5. Results and Analysis
We first compare the proposed method LoCoNet with previous state-of-the-art methods on multiple datasets and challenging scenarios. Then we validate our hypotheses of long-term intra-speaker and short-term inter-speaker modeling. Finally, following [2, 39, 45, 65, 75], we conduct extensive ablations on each component of LoCoNet (on AVA-ActiveSpeaker [58] unless otherwise noted).

Table 1. Comparison with SOTAs on AVA-ActiveSpeaker. 3DRNext denotes 3D ResNext [71]. \( R \) denotes 2D ResNet [26]. average-FLOPs represents the averaged FLOPs needed to process a single face crop. \( \dagger \) denotes our estimates based on their visual encoders. Most methods incur higher costs by extracting features for each frame through stacking multiple adjacent frames (i.e., 11 in SPELL). LoCoNet achieves the highest mAP with modest FLOPs.

5.1. Comparison with State-of-the-Art
In this section, we compare our approach with state-of-the-art methods on the three datasets.

AVA-ActiveSpeaker. From Table 1, end-to-end methods exhibit fewer FLOPs while maintaining competitive mAP compared to multi-stage methods. The higher FLOPs of multi-stage methods stem from their sequential-inference processing strategy (Sec. 2) of stacking multiple neighboring frames centered at time \( t \). LoCoNet achieves a 95.2% mAP, surpassing the best-performing end-to-end ASD method Light-ASD [45] by 1.1% while using modest average-FLOPs. Moreover, LoCoNet outperforms previous state-of-the-art multi-stage method SPELL+ [48] by 0.3% with about 32% fewer parameters and over 38× fewer average-FLOPs.

Talkies set. We evaluate LoCoNet with other methods under three training settings: (i) AVA-ActiveSpeaker, (ii) Talkies, and (iii) pretrained on AVA-ActiveSpeaker and finetuned on Talkies. As shown in Table 2, LoCoNet outperforms EASEE by 1.7%, 2.5%, and 2.7% in these three settings, respectively.
### Table 2. Comparison on Talkies dataset under three training settings: train on AVA-ActiveSpeaker alone, train on Talkies alone, or train on AVA-ActiveSpeaker and finetune on Talkies.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train Set</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVA Talkies</td>
<td></td>
</tr>
<tr>
<td>MAAS [42]</td>
<td>✓</td>
<td>79.7</td>
</tr>
<tr>
<td>EASEE [3]</td>
<td>✓</td>
<td>86.7</td>
</tr>
<tr>
<td>LoCoNet</td>
<td>✓</td>
<td>88.4</td>
</tr>
<tr>
<td>EASEE [3]</td>
<td>✓</td>
<td>93.6</td>
</tr>
<tr>
<td>light-ASD[45]</td>
<td>✓</td>
<td>96.1</td>
</tr>
<tr>
<td>LoCoNet</td>
<td>✓</td>
<td>97.2</td>
</tr>
<tr>
<td>EASEE [3]</td>
<td>✓ ✓</td>
<td>94.5</td>
</tr>
</tbody>
</table>

### Table 3. Comparison on Ego4D dataset. The challenge winner is not specifically optimized for ASD but the large improvement (+7.7%) still shows the strong generalizability of LoCoNet. The result of TalkNet was obtained by us with their released code.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TalkNet [65]</td>
<td>51.7</td>
</tr>
<tr>
<td>Challenge Winner [47]</td>
<td>60.7</td>
</tr>
<tr>
<td>LoCoNet</td>
<td>68.4</td>
</tr>
</tbody>
</table>

It also outperforms light-ASD by 2.2% when both models are trained on Talkies only.

**Ego4D dataset.** We evaluate our method on Ego4D Audio-Visual benchmark [24]. LoCoNet achieves 68.4% mAP, outperforming TalkNet and Challenge Winner [47] (SPELL [48]-based) by 16.7% and 7.7% respectively. Egocentric videos, characterized by constant camera motion, lower clarity, and more complex scenes compared to exocentric videos, demonstrate the potential of the proposed Long-Short Context Modeling in real-life scenarios. Our approach benefits from capturing both multi-speaker interaction and single-speaker behavior.

### 5.2. Challenging Scenario Evaluation

#### Quantitative analysis.

We report the performance of LoCoNet on AVA-ActiveSpeaker under different face sizes: (i) Small: faces with width less than 64 pixels; (ii) Medium: faces with width between 64 and 128 pixels; (iii) Large: faces with width larger than 128 pixels. We also study the effect of the number of visible faces in a video frame (1, 2, or 3).

The results, along with the portions of each category, are shown in Tables 4 and 5 respectively. LoCoNet consistently performs the best across all scenarios, exhibiting the most significant improvement in the challenging multi-speaker case: +3.0% for 3 faces. This suggests that our method more effectively models both the target speaker’s speaking pattern and the interactions of context speakers, allowing accurate inference of the speaking activity of the target person.

#### Qualitative analysis.

Fig 5 visualizes the results of LoCoNet and TalkNet [65] on AVA-ActiveSpeaker [58] with the groundtruth labels. The video on the left shows four visible speakers talking, posing a challenge in distinguishing the active speaker amid multiple discussions in the same scene. LoCoNet accurately locates the active speakers in this case, whereas TalkNet fails to recognize some of them. The first two columns of the video on the right shows a woman with a very small visible face as the active speaker, while the man with a large visible face is not speaking. TalkNet fails to locate the active speaker while LoCoNet succeeds.

Fig 5 also visualizes the results of LoCoNet and light-ASD [45] on AVA-ActiveSpeaker [58] with the groundtruth labels. The video on the left shows four visible speakers talking, posing a challenge in distinguishing the active speaker amid multiple discussions in the same scene. LoCoNet accurately locates the active speakers in this case, whereas TalkNet fails to recognize some of them. The first two columns of the video on the right shows a woman with a very small visible face as the active speaker, while the man with a large visible face is not speaking. TalkNet fails to locate the active speaker while LoCoNet succeeds.

By combining long-term intra-speaker context to compare the speaking pattern of each individual and short-term inter-speaker context to examine the conversations, our approach better overcomes this challenging speaking scenarios. However, in the last column, both methods fail to recognize the active speaker at the back. This scenario is especially challenging as the two active speakers are in separate conversations with one being far less salient than the other, making it difficult to infer the speaking activity of the less salient speaker.

### Table 4. Results as a function of face size. LoCoNet achieves the highest mAP among all face sizes.

<table>
<thead>
<tr>
<th>Method</th>
<th>Face Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small (18%)</td>
</tr>
<tr>
<td>ASC [2]</td>
<td>56.2</td>
</tr>
<tr>
<td>MAAS [42]</td>
<td>55.2</td>
</tr>
<tr>
<td>TalkNet [65]</td>
<td>63.7</td>
</tr>
<tr>
<td>ASDNet [39]</td>
<td>74.3</td>
</tr>
<tr>
<td>EASEE [3]</td>
<td>75.9</td>
</tr>
<tr>
<td>light-ASD [45]</td>
<td>77.5</td>
</tr>
<tr>
<td>LoCoNet</td>
<td>77.8</td>
</tr>
</tbody>
</table>

### Table 5. Results as a function of visible faces in the scene. Larger improvements are observed on more challenging cases (i.e., 3 faces).

<table>
<thead>
<tr>
<th>Method</th>
<th># Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 (45%)</td>
</tr>
<tr>
<td>ASC [2]</td>
<td>91.8</td>
</tr>
<tr>
<td>MAAS [42]</td>
<td>93.3</td>
</tr>
<tr>
<td>TalkNet [65]</td>
<td>95.4</td>
</tr>
<tr>
<td>ASDNet [39]</td>
<td>95.7</td>
</tr>
<tr>
<td>EASEE [3]</td>
<td>96.5</td>
</tr>
<tr>
<td>light-ASD [45]</td>
<td>96.2</td>
</tr>
<tr>
<td>LoCoNet</td>
<td>97.0</td>
</tr>
</tbody>
</table>
5.3. Attention visualizations of LIM and SIM

We next visualize the effectiveness of Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM). The left part of Fig. 6 visualizes the attention weights across different frames of a single speaker in LIM. It is evident that speaking and non-speaking activities are distinctly separated, with clear boundaries when speaking activities change. This verifies that LIM contributes to accurate speaking activity detection. The right part of Fig. 6 shows the portions of information (measured by L2-norm, as convolutions are used for SIM) drawn by the target speaker at target frame from all speakers at nearby frames (and information flows with very small portions are not shown). The woman on the left gradually turns her face to the target speaker, which is the most indicative sign that the target speaker has started speaking. The distribution of information flow reveals that SIM infers from the behaviors of context speakers, assigning more attention to the woman on the left.

5.4. Context Modeling Analysis

Does Intra-speaker Modeling require long-term? We train LoCoNet by keeping the number of speakers $S$ as 1, and varying the temporal length of input frames $T$ from 20 to 400. Table 6 indicates that the network performs worst when trained with the shortest video segments of 20 frames (0.8 sec). Performance improves as video segments become longer, with a 5.0% mAP increase at 100 frames (4 sec) and an additional 0.4% at 200 frames. This underscores the importance of long-term temporal context in intra-speaker modeling, aligning with findings from TalkNet [65] and ASC [2] that long-term temporal context provides better evidence of a speaking episode. We found 200 frames to be a good balance between performance and memory cost.

Is Short-term Inter-speaker Modeling Sufficient? We first validate the importance of inter-speaker context. Keeping the temporal length at 200 frames, we vary the number of speakers $S$ from 1 to 3. Larger $S$ values were not considered because > 99% of videos in AVA-ActiveSpeaker [58] have at most 3 speakers in the same scene. Table 7 demonstrates that performance increases with more speakers included in the model.
Table 8. **Temporal Receptive Field** $k$ in SIM, indicating how many neighboring frames each speaker can explicitly query on in the SIM module. $S$ is set as 3 and $T$ as 200.

<table>
<thead>
<tr>
<th>Receptive Field</th>
<th>1</th>
<th>7</th>
<th>15</th>
<th>31</th>
<th>101</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFLOPs</td>
<td>0.28</td>
<td>0.98</td>
<td>1.93</td>
<td>3.82</td>
<td>12.1</td>
</tr>
<tr>
<td>mAP(%)</td>
<td>94.5</td>
<td>95.2</td>
<td>95.2</td>
<td>95.0</td>
<td>94.9</td>
</tr>
</tbody>
</table>

Table 9. **Ablations on components of LoCoNet.** We study the efficacy of audio encoder (ResNet-34 or VGGFrame), Long-term Intra-speaker Modeling (LIM) and Short-term Inter-speaker Modeling (SIM). LIM and SIM are stacked 3 times if applied.

<table>
<thead>
<tr>
<th>Ablation Settings</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-34 VGGFrame LIM SIM</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>91.2</td>
</tr>
<tr>
<td>✓</td>
<td>92.8</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>94.1</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>94.0</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓</td>
<td>95.2</td>
</tr>
</tbody>
</table>

Table 10. **Ablation on designs of Short-term Inter-speaker Modeling.** We implement SIM using Convolution or Window Self-Attention (Eqn. 5). We keep the reception field the same (i.e., 7 frames) and stack it three times ($N = 3$).

<table>
<thead>
<tr>
<th>Design</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution</td>
<td>95.2</td>
</tr>
<tr>
<td>Window Self-Attention</td>
<td>94.4</td>
</tr>
</tbody>
</table>

Table 11. **Ablation on the number of blocks in LSCM ($N$).**

<table>
<thead>
<tr>
<th>N</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>92.8</td>
</tr>
<tr>
<td>1</td>
<td>94.3</td>
</tr>
<tr>
<td>2</td>
<td>95.0</td>
</tr>
<tr>
<td>3</td>
<td>95.2</td>
</tr>
<tr>
<td>4</td>
<td>95.0</td>
</tr>
<tr>
<td>5</td>
<td>95.0</td>
</tr>
</tbody>
</table>

5.5. Other Ablations

**Does each component of LoCoNet help?** In Table 9, replacing ResNet-34 [26] with VGGFrame as the audio encoder enhances mAP by 1.6%, showing the effectiveness of an audio encoder pretrained on audio datasets compared to a common vision encoder. With VGGFrame as audio encoder, adding 3 Long-term Intra-speaker Modeling (LIM) modules increases the performance by 1.3%, while 3 Short-term Inter-speaker Modeling (SIM) modules adds 1.2%. Both together improve mAP by 2.4%.

**Convolution versus Window Self-Attention in SIM.** Besides Convolution, we also try Window Self-Attention to capture local patterns for SIM (Eqn. 5). In Table 10, Convolution outperforms Window Self-Attention by 0.8% mAP, highlighting the superiority of Convolution in modeling local patterns of speakers’ interaction.

**Number of blocks in LSCM.** We vary the number of blocks $N$ in Long-Short Context Modeling (Sec. 3.2). Adding 1 block of LSCM yields a performance gain of 1.5%, showing the effectiveness of LIM and SIM. Adding two more blocks further increases by 0.7%, but results saturate at $N = 3$.

6. Conclusion

In this work, we observe that speaker activity can be more efficiently inferred from long-term intra-speaker context and short-term inter-speaker context. We thus design an end-to-end long-short context ASD framework that uses self-attention and cross-attention mechanisms to model long-term intra-speaker context and a convolutional network to model short-term inter-speaker context. With a simple backbone network, our method achieves state-of-the-art performance on 3 mainstream ASD benchmarks and significantly outperforms previous state of the art methods by 7.7% on Ego4D. We also show that in challenging scenarios where multiple speakers are in the same scene or speakers have small faces, our proposed method also outperforms previous methods. All of these results show the robustness and effectiveness of our method. Similar to existing long-term ASD methods, our method utilizes 8 sec of context. Future work should study how to implement larger contexts. Additionally, enhancing ASD in egocentric datasets could include adding other modalities, such as eye gaze.

Acknowledgements

The authors gratefully acknowledge Prof. David Crandall’s guidance and feedback on earlier versions of this work. This work was supported in part by the National Science Foundation under award DRL-2112635 to the AI Institute for Engaged Learning, Sony Faculty Innovation Award, Laboratory for Analytic Sciences via NC State University, and ONR Award N00014-23-1-2356. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
References


