ASD-TRANSFORMER: EFFICIENT ACTIVE SPEAKER DETECTION USING SELF AND MULTIMODAL TRANSFORMERS

Gourav Datta[*†1] Tyler Etchart[*2] Vivek Yadav[*2] Varsha Hedau[2] Pradeep Natarajan[2]
Shih-Fu Chang[2,3]

[1]University of Southern California, [2]Amazon Inc., [3]Carnegie Mellon University
gdatta@usc.edu {etchartt, ydvivek, hedauv, natarap, shihfu}@amazon.com

ABSTRACT

Multimodal active speaker detection (ASD) methods assign a speaking/not-speaking label per individual in a video clip. ASD is critical for applications such as natural human-computer interaction, speaker diarization, and video reframing. Recent work has shown the success of transformers in multimodal settings, thus we propose a novel framework that leverages modern transformer and concatenation mechanisms to efficiently capture the interaction between audio and video modalities for ASD. We achieve mAP similar to state-of-the-art (93.0% vs 93.5%) on the AVA-ActiveSpeaker dataset. Further, our model has $\sim 3\times$ smaller size (15.23MB vs 49.82MB), reduced FLOPs count (11.8 vs 14.3), and lower training time (15h vs 38h). To verify our model is making predictions from the right visual cues, we computed saliency maps over input images. We found that in addition to mouth regions, the nose, cheek, and area under the eye were helpful in identifying active speakers. Our ablation study reveals that the mouth region alone achieved lower mAP (91.9% vs 93.0%) compared to full face region, supporting our hypothesis that facial expressions in addition to mouth region are useful for ASD.

Index Terms— multimodal, active speaker detection, transformer, saliency maps, human-computer interaction

1. INTRODUCTION

In human to human interaction, there is a strong use of both audio and visual signals to enrich understanding and conversation. Building more conversational and natural AI requires improving techniques to understand and process information from both audio and video. The goal in ASD is to determine who among multiple individuals are speaking.

ASD is a deeply multimodal problem with an explicit need to learn alignment between audio and visual sources to confirm speaking. Recent work has applied transformers [1] in multimodal settings [2, 3] with success due to their ability to correlate features across modalities in long temporal context, which is crucial in ASD.

Our key contributions are applying multimodal (audio-video) transformers to ASD, leveraging bilinear pooling for audio-video fusion, and generating saliency maps to study the relative importance of face regions for ASD predictions. We also conduct ablation studies that illustrate the efficacy of our feature encoders as indicated by mAP gains. While our work is inspired by [4], we improve over that method's mAP performance (93.0% vs 92.3%) due to our full transformer modules, introduction of single modality self-transformers, efficient temporal feature extractors (Conformer for audio and 1D CNN for video) and richer modality fusion. Additionally, we near the SOTA performance of [5] (93.0% vs 93.5%) while significantly reducing our model size, FLOP count, and training time (see Table 3) as we partially avoid compute-intensive 3D CNNs.

2. RELATED WORKS

Early research in ASD includes approaches using visual-only inputs [6, 7, 8], or audio-only inputs [9]. Visual-only ASD is susceptible to errors from non-speech related face/mouth actions such as laughing or eating. Whereas audio-only ASD is susceptible to errors due to background speech and noise. More recently, with the introduction of the large-scale ASD benchmark AVA-ActiveSpeaker, there has been a plethora of research showing the benefits of multimodal inputs [4, 10, 5]. The first multimodal attempt at ASD in the wild was a joint audiovisual model by [11], a method trained end-to-end directly from image pixels and audio without the use of any pre-trained networks. Next, [12] proposed using a shallow 3D CNN based front-end and an ensemble of temporal convolution and LSTM classifiers to predict ASD, yielding significant improvements over earlier work. Concurrently, a novel representation that models pairwise relationships between multiple speakers over a large time window was proposed by [10]. This method achieved mean average precision (mAP) performance similar to [12].

Compared to these works, a significant leap in ASD performance and training efficiency was observed in [4], which explored long-term multimodal temporal feature extraction and served as the basis of our work. Later, [13] proposed a multi-objective learning scheme to leverage the best of each modality using a novel self-attention, uncertainty-based fusion mechanic.

Currently, the best performance on the AVA benchmark is achieved by [5], which introduces several architectural modifications to [10], including a Multilayer perceptron (MLP) for inter-speaker modeling, SincNet [14] for audio feature

[*]Equal Contribution. [†]Work done during Amazon internship.

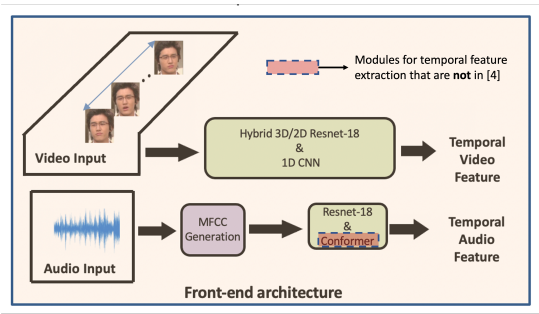


Fig. 1. Front-end ASD model architecture, consisting of audio-video feature extraction

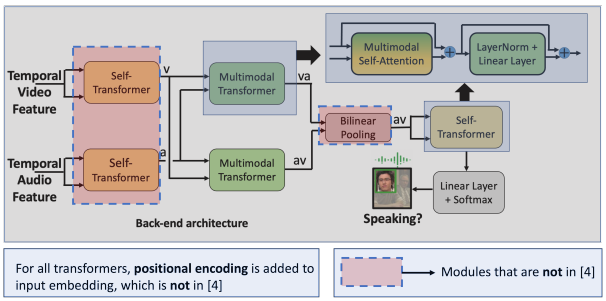


Fig. 2. Back-end ASD model architecture, consisting of cross-modal and self transformers

extraction, and deep 3D CNNs for video feature extraction. 3D CNNs are computationally expensive and MLPs don't allow for explicit interaction between the audio and visual modalities. Additionally, training the front-end (audio-visual integration) first, followed by feature extraction, then back-end training (inter-speaker and temporal modeling), leads to longer training times as opposed to training end-to-end.

Our method largely builds on the efforts of [4] to bring mAP performance closer to [5], while still maintaining lower compute costs and training times. We augment [4] by adding a self-transformer with multiple attention heads for each audio and video embedding, extending the cross-modal attention to be a full multimodal transformer with positional encoding, and employing bilinear pooling [15] for modality fusion. We describe our full architecture in Sections 3 and 4.

3. FRONT-END FEATURE EXTRACTION

3.1. Audio Encoder

The audio encoder ingests audio frames represented by a tensor of 13 Mel-frequency cepstral coefficients (MFCCs) across 25ms, then extracts the relevant audio features necessary for ASD, similar to [12, 4]. We use a 2D ResNet-18 network with dilated convolutions [4] such that the temporal dimension of the resultant audio features match that of the video features. We then perform temporal modeling of the extracted audio features using a conformer (convolution-augmented transformer) architecture, as shown in Fig. 1. We use a conformer because they have been shown to capture both the local and global dependencies of an audio sequence in a parameter-efficient way [16]. Note that temporal processing is crucial for accurate ASD, because it enables the refinement of the extracted audio features by attending to their temporal structure.

3.2. Video Encoder

Similar to audio, the video encoder extracts facial features which indicate speaking/not-speaking. For input, we used a pretrained, frozen face predictor network to find and crop the face regions offline. The video encoder uses a hybrid 3D-2D convolutional neural network, to extract the relevant facial features. In particular, similar to [4], it consists of one 3D convolutional layer, that first reduces each spatial dimension by $\sim 3\times$, followed by a 2D ResNet18 block for compute-efficiency. This front-end encoder is followed by a temporal processing block, which consists of 5 depthwise-separable 1D CNN layers, followed by a traditional 1D CNN layer for reducing the feature dimension [4]. We observe that simple, compute-efficient 1D CNN blocks are sufficient for video temporal modeling, given that global context is later captured by the video self-transformer.

4. BACK-END MULTI-MODAL PROCESSING

Our backend architecture, shown in Fig. 2, starts with separate audio and video self-transformers that model inter-speaker relational context. Next is two concurrent multimodal transformers [1], where the audio features attend to the video counterparts and vice-versa. The multimodal transformers perform audio-visual synchronization, which is particularly important given the noise introduced by AVA dubbed videos.

We combine our multimodal transformer outputs through bilinear pooling [15]. Bilinear pooling has been empirically shown to efficiently and expressively combine textual and visual information due its use of the outer product, without the typical increase in dimension size from standard outer product operations. We show in this work that it can be extended to audio-video concatenation. Finally, we apply a self-transformer, linear layer, and softmax to the fused embeddings to predict dense ASD labels.

Both the self and multimodal transformer architectures consist of an attention layer, followed by a feed-forward layer (with layer normalization and residual connections), similar to the encoder block in traditional transformer architectures [1]. The inputs are the projection vectors of query, key, and value from audio and visual embeddings, respectively. While self-transformers either ingest these from a particular modality (q_a, k_a, v_a or q_v, k_v, v_v) or from fused modalities (q_{av}, k_{av}, v_{av}), multimodal transformers require that key and value correspond to the same modality while query comes from a different modality, similar to [4]. The outputs of each of these transformers can be represented as follows:

$$F_{a-a}=SM(\frac{q_a k_a^T}{\sqrt{d}})v_a,\ F_{v-v}=SM(\frac{q_v k_v^T}{\sqrt{d}})v_v \quad (1)$$

$$F_{a-v}=SM(\frac{q_v k_a^T}{\sqrt{d}})v_a,\ F_{v-a}=SM(\frac{q_a k_v^T}{\sqrt{d}})v_v \quad (2)$$

$$F_{av-av}=SM(\frac{q_{av}k_{av}^{T}}{\sqrt{d}})v_{av} \quad (3)$$

Where SM denotes the softmax function, d denotes the dimensionality of q, k, v, and F_{av-av} denotes the fused output.

5. EXPERIMENTAL SETUP & DATA INGESTION

We use a dense sampling scheme to ensure each video frame has its corresponding ASD prediction in a particular batch, similar to [4], which increases the training efficiency by requiring less forward passes than sliding window-based approaches [5, 10]. We use cross-entropy loss to compare predicted labels with ground-truth.

Since it is challenging to accurately predict speaking/not-speaking in the presence of background speech/noise, [4] proposed a negative sampling technique, where the number of training samples are increased by randomly using a different audio clip from the same mini-batch as noise with the original video. We extend this technique by additionally shifting the audio track by a random number of frames. Our augmentation technique can thus help tolerate little audio-video desynchronization (up to $\sim$1 s) present in the dubbed AVA videos. Moreover, similar to [4], our approach does not require any external dataset to inject additional noise. For vision inputs, we apply the standard augmentation techniques of rotating, flipping, and cropping input images randomly.

Our architecture and training algorithm are implemented in PyTorch and all experiments are performed using a single NVIDIA V-100 GPU with 16 GB memory. We use the Adam optimizer for 50 epochs, with an initial learning rate of 0.0002, which is decreased by 5% for every epoch. The face crops are reshaped to 128×128. The number of MFCC vectors is set to 13. We set the dimensions of the audio and visual embeddings to 256. All attention blocks in the transformers used in our architecture have eight attention heads [4].

5.1. AVA-ActiveSpeaker Dataset

The AVA-ActiveSpeaker dataset is the current SOTA benchmark for large-scale ASD [11]. It consists of 262 15-minute video clips from Hollywood movies, 120 of which are used for training, and 33 for validation. There are a total of 3.65 million manually annotated face crops, each of which is assigned a binary ASD label. Contiguous face crops corresponding to a particular person are concatenated to create a face-track. The key challenges involved in this dataset are wide diversity in terms of languages and demographics, varying fps ($25-30$), large number of low-resolution (<100 pixels) face crops, noisy audio, and short utterance lengths ($\sim$1 s), which hinders the use of large sliding windows for ASD. We report mAP performance[1] as is customary for this dataset. We trim the first and last 2 frame predictions in reporting mAP, similar to sliding window approaches without padding, because we observe that the starting and ending frames of any face track perform worse due to lack of surrounding context.

[1]We use the validation set for evaluations because the test set evaluation server was closed during the submission of this paper.

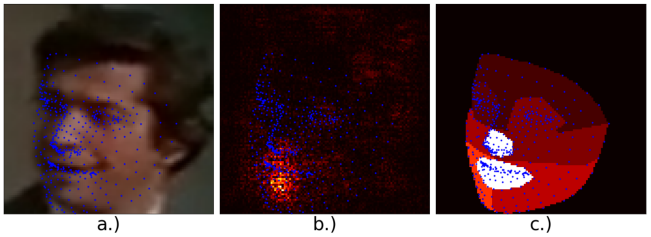


Fig. 3. From left to right: a.) face crop image from AVA-ActiveSpeaker dataset, b.) gradient-based saliency map representing visual attention, c.) average saliency value binned by face region

5.2. Ablation Studies

We performed numerous ablation studies to understand the effectiveness of our transformer based method. In Section 6.2, we show the learned behavior of the visual encoder by analyzing an aggregated facial region saliency map [17]. To generate the saliency map, we calculate gradients with respect to each input image in the validation set, compute facial keypoints, bin the saliency values based on facial region, and compute the average per facial region (process shown in Figure 3). In Section 6.3, we compare various temporal feature encoder architectures and explore different combinations of multimodal transformers and their relative placements. Finally, we compare model efficiency in Section 6.4.

6. RESULTS

6.1. mAP Results

In Table 1, we report a validation mAP of 93.0% on the AVA-ActiveSpeaker dataset. We near SOTA results [5] with significantly lower training time and model size, as shown in Section 6.4.

Approach	mAP (%)
ASD-Transformer	93.0
Kopuklu et.al. [5]	93.5
Talknet [4]	92.3
MAAS-TAN [18]	88.8
ASC [10]	87.1
Roth et. al. [11]	79.2

Table 1. Comparison with SOTA methods on the validation set of the AVA-ActiveSpeaker dataset.

6.2. Saliency Map Analysis

Figure 4 represents the average saliency value of each facial region normalized by the max region value (mouth). As expected, our method mostly attends to the mouth region for ASD predictions as seen by the nose region and the lower jaw left and right regions. It seems the further from the mouth region, the less attention the face region gets, with the notable exception that the eyes get nearly double the attention as the rest of the upper face. We further explore the attention on

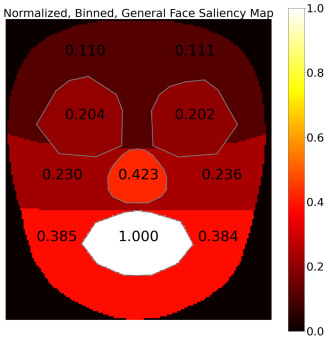


Fig. 4. Gradient-based saliency map for the canonical face generated from the AVA-ActiveSpeaker validation dataset

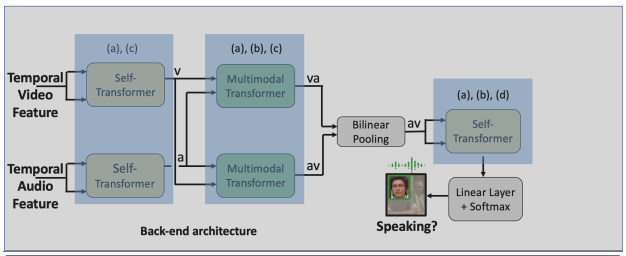
other facial regions by training our method first using mouth crop inputs and second using face crops with masked mouth regions as input. To generate mouth crop inputs offline, we predict facial keypoints on the face crop, then perform align and crop on the mouth such that the leftmost and rightmost mouth points lie on a horizontal line. We see mAP degradation in the mouth-only model versus the full face model (91.9% vs 93.0%) supporting the usefulness of facial expressions in ASD prediction. Additionally, we observe that our masked mouth model achieves 83.42% mAP, showing that facial expressions with audio is sufficient in our framework to achieve higher mAP than [11] (79.2% mAP). Even with no information from the mouth, the model is still able to pull out enough information to make reasonable predictions and beat another recent architecture that has access to mouth information.

Temporal Encoder Modeling		
Video	**Audio**	**mAP (%)**
1-D CNN	Conformer	93.0
GRU	Conformer	92.1
1-D CNN	GRU	92.4
GRU	GRU	91.9

Table 2. Comparison of mAP with different temporal encoder modeling techniques on the AVA-ActiveSpeaker dataset.

6.3. Temporal Feature Encoder Modeling and Multimodal Transformers

In Table 2, we compare various temporal feature encoder modeling architectures. The conformer and 1D CNN perform best for audio and video respectively. In Figure 5, we perform ablation studies on various combinations of multimodal transformers and their relative placements. Note that (b) in Figure 5 refers to removing the self-transformers applied to the extracted features for each modality, whereas (c) refers to removing the fused transformer after bilinear pooling. We also compare the ASD performance for early (immediately after front-end) fusion in (d), where we only apply the fused transformer, since the information corresponding to each modality is already lost. Our results indicate that the multimodal transformer gives the highest increase in mAP (2.6%, d to b), followed by the fused self-transformer (1.3%, c to a)



Approach	mAP
(a) ASD-AVTransformer	93.0
(b) Multimodal Transformer + Fused Self-Transformer	92.5
(c) Self-Transformer + Multimodal Transformer	91.7
(d) Early Fusion + Fused Self-Transformer	89.9

Fig. 5. Comparison of mAP with different transformer designs on the validation set of the AVA-ActiveSpeaker dataset.

then the feature self-transformers (.5%, b to a). Combining all of them gives the best mAP (93.0%, a).

6.4. Parameter, Compute, and Training Efficiency

Our approach has significant improvements in model size and FLOPs count over [5], as shown in Table 3, since we partially get rid of the expensive 3D convolutions, which enables the deployment of our models on low-power edge devices. Our approach can also be trained end-to-end, unlike [5], which needs to train feature extractors first, followed by extraction and back-end training thus resulting in $\sim$2.6$\times$ more training time. Also note that our method yields similar model size, FLOPs count, and training time compared to [4], while offering 0.7% higher mAP in the AVA benchmark. The simulation times are benchmarked on an Nvidia V-100 GPU with 15.6 GB of memory.

Method	Model Size (MB)	GFLOPs Count	Training time (hrs)
ASD-Transformer (Face)	15.02	11.8	15.0
Kopuklu et. al. [5]	48.75	14.25	38.7
TalkNet [4]	15.50	11.3	14.2

Table 3. Comparison of model size, compute, and training time of our proposed model with the top ASD models

7. CONCLUSION

In this work, we presented a novel framework composed of audio and video feature extractors, self and multimodal transformers, and bilinear pooling for audio-video modality fusion. The framework is trained end-to-end resulting in effective, efficient multimodal ASD. We showed that our framework nears the SOTA results against the mainstream ASD benchmark, namely AVA-ActiveSpeaker, while decreasing model size by 3x. Additionally, we validated our visual encoder's learned features and showed the usefulness of facial expressions in predicting ASD through a canonical face saliency map analysis. Finally, we performed ablation studies to show the efficacy of our multimodal transformer architecture and temporal feature encoders design.

8. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[2] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song, "Parameter efficient multimodal transformers for video representation learning," in *International Conference on Learning Representations*, 2021.

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, "VATT: transformers for multimodal self-supervised learning from raw video, audio and text," *CoRR*, vol. abs/2104.11178, 2021.

[4] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," *ACM Multimedia (MM)*, 2021.

[5] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1193–1203.

[6] Kalin Stefanov, Jonas Beskow, and Giampiero Salvi, "Vision-based Active Speaker Detection in Multiparty Interaction," in *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, 2017, pp. 47–51.

[7] Spyridon Siatras, Nikos Nikolaidis, Michail Krinidis, and Ioannis Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 133–137, 2009.

[8] J.M. Rehg, K.P. Murphy, and P.W. Fieguth, "Vision-based speaker detection using bayesian networks," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 1999, vol. 2, pp. 110–116.

[9] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. V. hamme, "Who's speaking?: Audio-supervised classification of active speakers in video," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015.

[10] J. L. Alcazar, F. Caba, L. Mai, F. Perazzi, J. Lee, P. Arbelaez, and B. Ghanem, "Active speakers in context," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru, "Ava active speaker: An audio-visual dataset for active speaker detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4492–4496.

[12] Joon Son Chung, "Naver at activitynet challenge 2019 – task b active speaker detection (ava)," *arXiv preprint arXiv:1906.10555*, 2019.

[13] Baptiste Pouthier, Laurent Pilati, Leela K. Gudupudi, Charles Bouveyron, and Frederic Precioso, "Active speaker detection as a multi-objective optimization with uncertainty-based multimodal fusion," *Interspeech 2021*, Aug 2021.

[14] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," *arXiv preprint arXiv:1808.00158*, 2019.

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, June 2016, pp. 317–326, IEEE Computer Society.

[16] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[17] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Workshop at International Conference on Learning Representations*, 2014.

[18] J.B. León-Alcázar, F.C. Heilbron, A. Thabet, and B. Ghanem, "Maas: Multi-modal assignation for active speaker detection," *arXiv preprint arXiv:2101.03682*, 2021.