IVUL ActivityNet Challenge 2022 Submission - Active Speaker Detection (AVA)

Juan León Alcázar[1], Moritz Cordes[2], Chen Zhao[1]

[1] King Abdullah University of Science and Technology (KAUST), [2]Leuphana University of Lüneburg

jc.leon@uniandes.edu.co, moritz.cordes@stud.leuphana.de, chen.zhao@kaust.edu.sa

Our method is an alternative to the traditional two-stage Active Speaker Detection (ASD) methods. EASEE (End-to-end Active Speaker dEtEction) is able to learn multi-modal features from multiple visual tracklets, while simultaneously modeling their spatio-temporal relations in an end-to-end manner. Our end-to-end architecture relies on a spatio-temporal module for context aggregation.

EASEE Architecture EASEE has three main components: (i) audio Encoder, (ii) visual Encoder, and a (iii) spatio-temporal Module. The visual encoder (f_v) performs multiple forward passes (one for each available tracklet), and the audio encoder (f_a) performs a single forward pass on the shared audio clip. These features are arranged according to their temporal order and (potential) spatial overlap, creating an intermediate feature embedding (Φ) that enables spatio-temporal reasoning. Unlike other methods, we construct Φ such that it can be optimized end-to-end. Figure 1 contains an overview of our proposed approach.

Temporal Endpoints We define a set of temporal endpoints (L) where the original video data (visual and audio) is densely sampled. At every temporal endpoint, we collect visual information from the available face tracklets and sample the associated audio signal. To further limit the memory usage, we define a fixed number of tracklets (i) to sample at every endpoint.

Spatio-Temporal Embedding. We build the embedding Φ over the endpoint set L. We define the spatio-temporal embedding e at time t for speaker s as $e_{t,s} = \{f_a(t), f_v(s,t)\}$. Since there may be multiple visible persons at this endpoint (*i.e.* $|s| \geq 1$), we define the embedding for an endpoint at time t with up to i speakers as $E_{t,i} = \{e_{t,0}, e_{t,1}, e_{t,2}, ..., e_{t,i}\}$. The full spatio-temporal embedding $\Phi_{i,k,l,t}$ is created by sampling audio and visual features over the endpoint set L, thus $\Phi_{i,k,l,t} = \{E_{t,i}, ..., E_{t+k,i}, ..., E_{t+lk,i}\}$. As $\Phi_{i,k,l,t}$ is assembled from independent forward passes of the f_a and f_v encoders, we share weights for forward passes in the same modality, thus each forward/backward pass accumulates gradients over the same weights. This shared weight scheme largely simplifies the complexity of the proposed network, and keeps the total number of parameters stable regardless of the

Method	mAP Val	mAP Test
Extended UniCon [13]	N/A	93.3
EASEE-50 (Ours)	94.1	93.0
ASDNet [7]	93.5	91.7
EASEE-18 (Ours)	93.3	N/A
TalkNet [11]	92.3	90.8
UniCon [13]	92.0	90.7
EASEE-2D (Ours)	91.1	N/A
Active Speakers in Context [1]	87.1	86.7
Zhang *et al.* [14]	84.0	N/A

Table 1. **State-of-the-art Comparison on AVA-ActiveSpeaker.** Our best network (EASEE-50) outperforms any other method by at least 0.6 mAP even approaches that build upon much deeper networks. Our smaller network (EASEE-18) remains competitive with the previous state-of-the-art. In the 2D scenario EASEE-2D only lags behind UniCon [13], improving the closest method by at least 0.9 mAP.

Graph Neural Network We propose an interleaved Graph Neural Network (iGNN) block to model the relationships between speakers in adjacent timestamps. iGNN performs two message passing steps: first a spatial message passing that models local interactions between speakers visible at the same timestamp, and then a temporal message passing that effectively aggregates long-term temporal information:

$$\Phi^s = M^s(A^s\Phi^J;\theta^s), \Phi^t = M^t(A^t\Phi^J;\theta^t)$$
$$iGNN(\Phi^J) = (M^t \circ M^s)(\Phi^J) + \Phi^J$$

Here, M^s is a GCN layer that performs spatial message passing using the spatial adjacency matrix A^s over an initial feature embedding (Φ^J), thus producing an intermediate representation with aggregated local features (Φ^{J+1}). Afterwards, the GCN layer M^t performs a temporal message passing using the temporal adjacency matrix A^t. θ^s and θ^t are the parameter set of their respective layers. The residual connection favors gradient propagation.

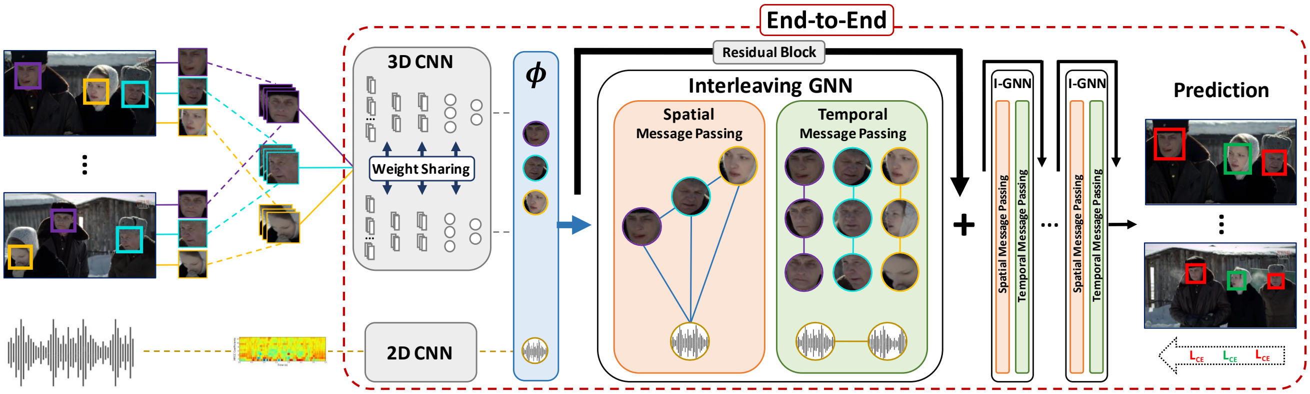

Figure 1. **Overview of the EASEE architecture.** We fuse information from multiple visual tracklets, and their associated audio track. We rely on a 3D CNN to encode individual face tracklets, and a 2D CNN to encode the audio stream (Grey Encoders). These embeddings are assembled into an initial multi-modal embedding (Φ) containing audiovisual information from multiple persons in a scene. We map this embedding into a graph structure that performs message passing steps over spatial (light orange) and temporal dimensions (light green). Our layer arrangement favors independent massage passing steps along the temporal and spatial dimensions.

Network	End-to-End	iGNN	Residual Connections	mAP
EASEE-50	✗	✗	✗	91.9
EASEE-50	✓	✗	✗	93.5
EASEE-50	✓	✗	✓	93.7
EASEE-50	✓	✓	✗	93.8
EASEE-50	✓	✓	✓	**94.1**

Table 2. **AVA-ActiveSpeaker Ablation.** We assess the empirical contribution of the most relevant components in EASEE. Residual connections contribute about 0.3 mAP and the proposed iGNN block 0.4 mAP. Overall the most relevant design choice is the end-to-end trainable nature of EASEE contributing 1.6 mAP.

Implementation Details We implement the audio encoder f_a with the Resnet18 convolutional encoder [4] pre-trained on ImageNet [2]. We adapt the raw 1D audio signal to fit the input of a 2D encoder by generating Mel-frequency cepstral coefficients (MFCCs) of the original audio clip, and then averaging the filters of the network's first convolutional layer to adapt for a single channel input [9]. We create the MFCCs with a sampling rate of 16 kHz and an analysis window of 0.025 ms. Our filter bank consists of 26 filters and a fast Fourier transform of size 256 is applied, resulting in 13 cepstrums. The visual encoder f_v is based on the R3D architecture, pre-trained on Kinetics-400 dataset [6]. For fair comparison with other methods, we also implement f_v as a 2D encoder by stacking the temporal and channel dimensions into a single one, then we replicate the filters on the encoder's first layer to accommodate for the input of dimension (B, CT, H, W) [10, 9]. We also rely on ImageNet pre-training [2] for this encoder.

We assemble Φ on-the-fly with multiple forward passes of f_a, f_v, and then map Φ into nodes of the Graph Convolutional Network and continue with the GCN in a single forward pass. We design the GCN module using the pytorch-geometric library [3] and use the EdgeConvolution operator [12] with filters of size 128. Each GCN layer contains a single iGNN block. EdgeConvolution allows to build a sub-network that performs the message passing between nodes, where every layer (spatial or temporal) in the iGNN is built by a sub-network of two linear layers with ReLu [8] and batch normalization [5]. Therefore, a single iGNN block contains 4 linear layers in total.

Results We compare EASEE against state-of-the-art ASD methods. The results for EASEE are obtained with $l = 7$ temporal endpoints, $i = 2$ tracklets per endpoint, and a stride of $k = 5$. This configuration allows for a sampling window of about 2.41 seconds regardless of the selected backbone. For fair comparison with other methods, we report results of three EASEE variants: 'EASEE-50' that uses a 3D backbone based on the ResNet50 architecture, 'EASEE-18' that uses a 3D model based on the much smaller Resnet18 architecture, and 'EASEE-2D' that uses a 2D Resnet18 backbone. Results are summarized in Table 1.

Ablation We ablate our best model (EASEE-50) to assess the individual contributions of our design choices, namely end-to-end training, the iGNN block, and the residual connections between the iGNN blocks. Table 2 contains the individual assessment of each component. The most important architectural design is the end-to-end training, which contributes 1.6 mAP. The proposed iGNN brings about 0.4 mAP when compared against a baseline network where spatial and temporal message passing is performed in the same layer. Finally, residual connections between iGNN blocks contribute with an improved performance of 0.3 mAP.

References

[1] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[3] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 2

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[7] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. *arXiv preprint arXiv:2106.03932*, 2021. 1

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2

[9] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. 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)*, pages 4492–4496. IEEE, 2020. 2

[10] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 2

[11] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 1

[12] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 2

[13] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen. Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3964–3972, 2021. 1

[14] Yuan-Hang Zhang, Jingyun Xiao, Shuang Yang, and Shiguang Shan. Multi-task learning for audio-visual active speaker detection. 1