

MAAS: Multi-modal Assignment for Active Speaker Detection

Juan León Alcázar¹, Fabian Caba Heilbron², Ali K. Thabet¹ & Bernard Ghanem¹

¹ King Abdullah University of Science and Technology (KAUST), ²Adobe Research

jc.leon@uniandes.edu.co, caba@adobe.com, ali.thabet@kaust.edu.sa, bernard.ghanem@kaust.edu.sa

Abstract

Active speaker detection requires a mindful integration of multi-modal cues. Current methods focus on modeling and fusing short-term audiovisual features for individual speakers, often at frame level. We present a novel approach to active speaker detection that directly addresses the multi-modal nature of the problem and provides a straightforward strategy, where independent visual features (speakers) in the scene are assigned to a previously detected speech event. Our experiments show that a small graph data structure built from local information can approximate an instantaneous audio-visual assignment problem. Moreover, the temporal extension of this initial graph achieves a new state-of-the-art performance on the AVA-ActiveSpeaker dataset with a mAP of 88.8%.

1. Introduction

Active speaker detection aims at identifying the current speaker (if any) from a set of candidate face detections in an arbitrary video. This research problem is an inherently multi-modal task that requires the integration of subtle facial motion patterns and the characteristic waveform of speech. Despite its multiple applications such as speaker diarization [3, 44, 46, 48], human-computer interaction [16, 58] and bio-metrics [34, 40], the detection of active speakers in-the-wild remains an open problem.

Current approaches for active speaker detection are based on recurrent neural networks [41, 43] or 3D convolutional models [1, 6, 60]. Their main focus is to jointly model the audio and visual streams to maximize the performance of single speaker prediction over short sequences. Such an approach is suitable for single speaker scenarios, but is overly simplified for the general (multi-speaker) case.

The general (multi-speaker) scenario has two major challenges. First, the presence of multiple speakers allows for incorrect face-voice assignments. For instance, false positives emerge when facial gestures closely resemble the motion patterns observed while speaking (*e.g.* laughing, grinning). Second, it must enforce temporal consistency over multi-modal data, which quickly evolves over time, *e.g.*, when active speakers switch during a fluid conversation.

In this paper, we address the general multi-speaker prob-

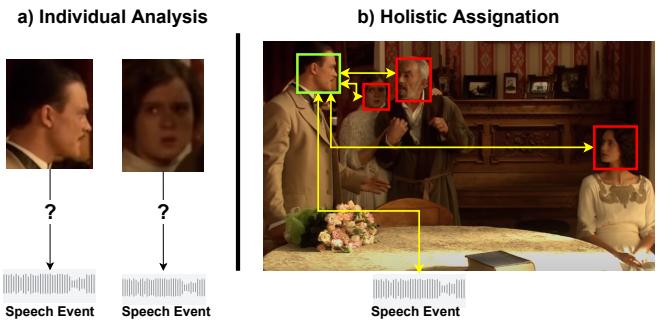


Figure 1. **Audiovisual assignment for active speaker detection.** Active speaker detection is highly ambiguous. Even if we analyze joint audiovisual information, unrelated facial gestures can easily resemble the natural motion of lips while speaking. In **a)** we show two face crops from a sequence, where a speech event was detected. The gestures, illumination, and capture angle make it hard to assess which face (if any) is the active speaker. Our strategy **b)** focuses on the attribution of speech segments in video. If a speech event is detected, we holistically analyse every speaker along with the audio track to discover the most likely active speaker.

lem in a principled manner. Our key insight is that, instead of optimizing active speaker predictions over individual audiovisual embeddings, we can jointly model a set of visual representations from every speaker in the scene along with a single audio representation extracted from the shared audio track. While simple, this modification allows us to map the active speaker detection task into an assignment problem, whose goal is to match multiple visual representations with a singleton audio embedding. Figure 1 illustrates some of the challenges in active speaker detection and provides a general insight for our approach.

Our approach, dubbed “Multi-modal Assignment for Active Speaker detection” (MAAS) relies on multi-modal graph neural networks [27, 50] to approach the local (frame-wise) assignment problem, but it is flexible enough to also propagate information from a long-term analysis window by simply updating the underlying graph connectivity. In this framework, we define the active speaker as the local visual representation with the highest affinity to the audio embedding. Our empirical findings highlight that reformulating the problem into a multi-modal assignment problem brings sizable improvements over current state-of-the-art

methods. On the AVA Active speaker benchmark, MAAS outperforms all other methods by at least 1.7%. Additionally, when compared with methods that analyze a short temporal span, MAAS brings a performance boost of at least 1.1%.

Contributions. This paper proposes a novel strategy for active speaker detection, which explicitly learns multi-modal relationships between audio and facial gestures by sharing information across modalities. Our work brings the following contributions: **(1)** We devise a novel formulation for the active speaker detection problem. It explicitly matches the visual features from multiple speakers to a shared audio embedding of the scene (Section 3.2). **(2)** We empirically show that this assignation problem can be solved by means of a Graph Convolutional Network (GCN), which endows flexibility on the graph structure and is able to achieve state of the art results (Section 4.1). **(3)** We present a novel dataset for active speaker detection, called “Talkies”, as a new benchmark composed of 10,000 short clips gathered from challenging and diverse scenes (Section 5).

To ensure reproducible results and promote future research, all the resources of this project, including source code, model weights, official benchmark results, and labeled data will be publicly available.

2. Related Work

In the realm of multi-modal learning, different information sources are fused with the goal of establishing more effective representations [36]. In the video domain, a common multi-modal paradigm involves combining representations from visual and audio features [4, 8, 22, 33, 34, 37, 49]. Such representation allows the exploration of new approaches to well established problems, such as person re-identification [33, 25, 55], audio-visual synchronization [1, 9, 10], speaker diarization [44, 48, 59], bio-metrics [34, 40], and audio-visual source separation [4, 22, 37, 41, 49]. Active speaker detection is a special instance of audiovisual source separation, where the sources are people in a video, and the goal is to detect and assign a segment of speech to one of those sources [41].

Active Speaker Detection. The work of Cutler *et al.* [12] pioneered research in active speaker detection in the early 2000s. It detected correlated audiovisual signals by means of a time-delayed neural network [47]. Follow up works [14, 42] approached the task relying only on visual information, focusing strictly on the evolution of facial gestures. Such visual-only modeling was possible as they addressed a simplified version of the problem with a single candidate speaker. Recent works [5, 10] have approached the more general multi-speaker scenario and relied on fusing multi-modal information from individual speakers. A parallel corpus of work has focused on audiovisual feature alignment, which resulted in methods that rely on audio as the primary

source of supervision [4], or as an alternative to jointly train a deep audiovisual embedding [8, 10, 35, 45].

The work of Roth *et al.* [41] introduced the AVA-ActiveSpeaker dataset and benchmark, the first large-scale video dataset for the active speaker detection task. In the AVA-ActiveSpeaker challenge of 2019, Chung *et al.* [6] presented an improved architecture of their previous work [10], which trains a large 3D model with the need for large-scale audiovisual pre-training [35]. Zhang *et al.* [60] also leveraged a hybrid 3D-2D architecture with large-scale pre-training [10, 11]. This method achieved its best performance when the feature embedding was optimized using a contrastive loss [18]. Follow up works focused on modeling an attention process over face tracks, where attention was estimated either from the audio alignment [1] or from an ensemble of speaker features [2]. We approach the active speaker problem in a more principled manner, as we go beyond the aggregation of contextual information from multiple-speakers and propose an approach that explicitly seeks to model the correspondence of a shared audio embedding with all potential speakers in the video.

Datasets for Active Speaker Detection. Apart from the development of the AVA-ActiveSpeaker benchmark, there are few public datasets specific to this problem. The most well known alternative is the Columbia dataset [5], which contains 87 minutes of labeled speech from a panel discussion. It is much smaller and less diverse than AVA. Modern audiovisual datasets [35, 10, 7] have been adapted for the large scale pre-training of some active speaker methods [6]. Nevertheless, these datasets were designed for related tasks such as speaker identification and speaker diarization. In this paper, we present the Talkies dataset as a new benchmark for active speaker detection gathered from social media clips. It contains 800,000 manually labeled face detections and includes challenging scenarios that contain multiple speakers, occlusion, and out of screen speech.

Graph Convolutional Networks (GCNs). GCNs [27] have recently gained popularity, due to the greater interest in non-Euclidean data. In computer vision, GCNs have been successfully applied to scene graph generation [23, 32, 39, 53, 57], 3D understanding [17, 30, 50, 52], and action recognition in video [21, 54, 56]. In MAAS, we design a DeepGCN-like architecture [28, 29, 31], which addresses a special scenario, namely the multi-modal nature of audiovisual data. We rely on the well-known EdgeConv operator [50] to model interactions between different modalities for graph nodes identified across multiple frames. This enables us to model both the multi-modal relations and the temporal dependencies in a single graph structure.

3. Multi-modal Active Speaker Assignation

Our approach is based on a straightforward idea. Instead of assessing the likelihood of individual audiovisual

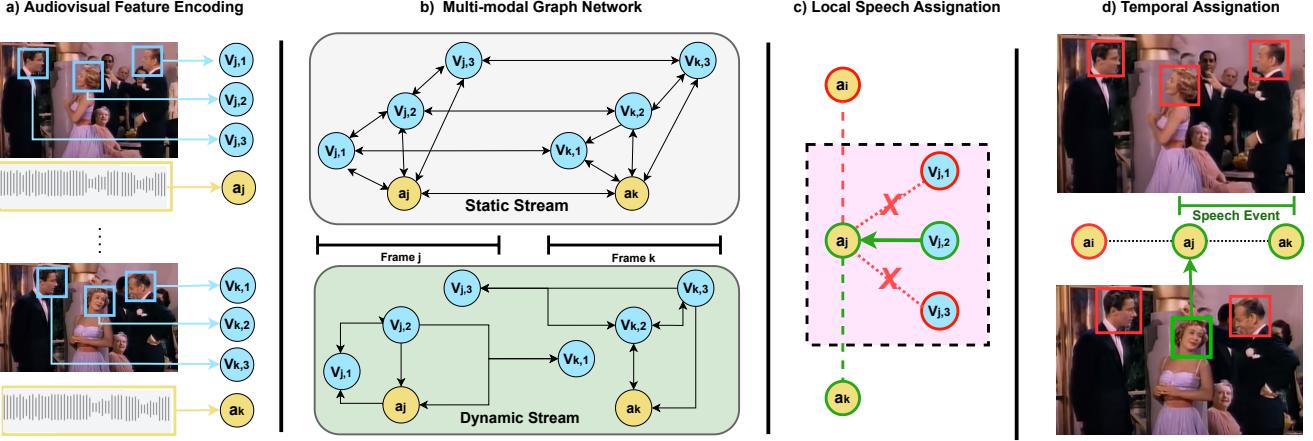


Figure 2. Overview of MAAS Pipeline. **a)**: Our approach begins by sampling independent audio and video features. Video features (cyan) are extracted from a stack of face crops that belong to a single person. Audio features (yellow) are extracted from the audio-spectrogram and are shared at the frame-level. **b)**: We create two feature graphs: one with static connections that model local temporal relations between the audio track and the visible persons; in parallel, we allow a secondary stream in the network to discover relationships given the estimated feature embeddings. **c)**: We estimate a frame-level affinity between the visual nodes and the local audio node such that the active speaker will have the highest affinity with the audio node. **d)**: Finally, we extend the network by modelling a longer temporal window. We jointly optimize the local affinities, while enforcing temporal consistency. We select the active speaker (green bounding box) as the most likely speaker to have generated the sequence of speech events.

patterns to belong to an active speaker, we directly model the correspondence between the local audio and the facial gestures of all the individuals present in the scene. This approach is motivated by the nature of the active speaker problem, which first identifies if any speech patterns are present, and then attributes those patterns to a single speaker.

Overall, our approach simultaneously solves three sub-tasks. First, we detect speech events in a short-term temporal window. Second, we iterate over all the visible speakers in a single frame, and decide which one is most likely to be an active speaker given the local information. Third, we extend this frame-level analysis along the temporal dimension, leveraging the inherent temporal consistency of video data to improve frame-level predictions. Figure 2 illustrates an overview of our MAAS approach.

3.1. Frame-Level Video Features

Following recent works [41, 60], we extract the initial frame-level features from a two-stream convolutional encoder. The visual stream takes as input a tensor of dimensions $H \times W \times (3c)$, where H and W are the image width and height, and c is the number of time consecutive face crops sampled from a single tracklet. Similar to [41], we transform the original audio waveform into a Mel-spectrogram and use it as input for the audio stream.

Our approach relies on independent audio and video features. To obtain these independent features (and to make fair comparison to state-of-the-art techniques), we train a joint model as described by [41], but drop the final two layers at inference time. These layers are responsible for the

feature fusion and final prediction.

At time t , a forward pass of our feature encoder yields $N + 1$ feature vectors for a frame with N possible speakers (detected persons). One shared audio embedding (\mathbf{a}_t) and N independent visual descriptors $\mathbf{v}_t = \{\mathbf{v}_{t,0}, \mathbf{v}_{t,1}, \dots, \mathbf{v}_{t,n-1}\}$ one for each of the N visible persons (see Figure 2-a). We define (\mathbf{s}_t) as the local set of features at time t , such that $\mathbf{s}_t = \{\mathbf{a}_t \cup \mathbf{v}_t\}$. The feature set \mathbf{s}_t is used for the optimization of the basic graph structure in MAAS, the Local Assignment Network, described next.

3.2. Local Assignment Network (LAN)

We model the local assignment problem by generating a directed graph over the feature set \mathbf{s}_t . Our local graph consists of an audio node and one video node for each potential speaker. We create a bidirectional connectivity between the audio node and each visual node thus leveraging a GCN that operates on a directed graph generated from \mathbf{s}_t . Figure 3 (left) illustrates this graph structure. We call this graph structure the Local Assignment Graph and the GCN that operates over it the Local Assignment Network (LAN).

The goals of LAN are two-fold: (i) to detect local speech events, (ii) if there is a speech event, to assign the most likely speaker from the set of candidates. We achieve these two goals by fully supervising every node in LAN. Visual nodes are supervised by the ground-truth, l_{tv} , of the corresponding speaker. On the other hand, each audio node receives a binary ground-truth label indicating whether there is at least one active speaker, *i.e.* $\max(\{l_{t,0}, l_{t,1}, \dots, l_{t,n-1}\})$; otherwise, there is silence.

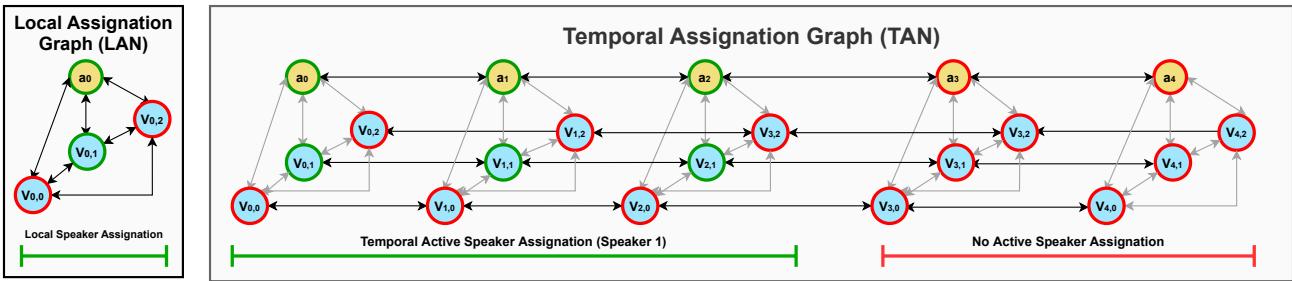


Figure 3. Assignment Graphs. The base static graphs for MAAS are composed of multi-modal nodes, visual nodes (Cyan), and audio nodes (Yellow). The local Assignment Graph (left) defines frame level connectivity of the individual features. The Temporal Assignment Graph (right) is composed of multiple local graphs (5 in the figure) and defines a temporal extension of the frame-level relations (we depict local relations in light gray to avoid visual clutter). While a local graph solves an instantaneous assignation problem. The temporal graph optimizes a subset of nodes thereby incorporating temporal information in the individual local graphs.

LAN is tasked to discover active speakers at the frame-level (*i.e.* t is fixed).

3.3. Temporal Assignment Network (TAN)

While the LAN is effective at finding local correspondences between audio patterns and visible faces, it models information sampled from short video clips (s_t). This sampling strategy can lead to inaccurate predictions from noisy or ambiguous local estimations (*e.g.* audio noise, blurred faces, ambiguous facial gestures, etc.). Therefore, we extend our proposed approach to include temporal information from adjacent frames.

We extend the local graph in LAN by sampling s_t over a temporal window (w) centered around time t . $w = [i, i+1, \dots, t, \dots, j]$ and define a temporal feature set $b_w = [s_i, s_{i+1}, \dots, s_t, \dots, s_j]$. Following the LAN structure outlined in 3.2, we can build $(j - i)$ independent local graph structures out of b_w (one for every time step). We augment this set of independent graphs by adding temporal links between time adjacent representations of frame-level features. We follow two rules to build these connections: we create temporal connections between time adjacent audio nodes, and we create temporal connections between time adjacent video nodes, only if they belong to the same person. No additional cross-modal connections are built. We call the resulting graph, the Temporal Assignment Graph, which allows for information flow between time adjacent audio and video features, thereby allowing for temporal consistency in the audio and video modalities. Figure 3 (right) illustrates this graph structure.

We build a GCN over the extended graph topology and call it the Temporal Assignment Network (TAN). TAN allows us to directly identify speech segments as continuous positive predictions over audio nodes. Likewise, it detects active speech segments over continuous predictions for same-speaker video features.

3.4. Dynamic Stream & Global Prediction

Finally, we account for potential connection patterns that go beyond our initial insights. We augment our architecture and define a second stream that will operate on the very same data as the static stream (including multiple temporal timestamps). However, we do not define a fixed connectivity pattern for this stream. Instead, we aim at creating a dynamic graph structure based on the node distribution in feature space. In this stream, we allow the GCN to estimate an arbitrary graph structure by calculating the K nearest neighbors in feature space for each node, and by establishing edges based on these neighbouring nodes. In practice, we replicate the static stream, drop the definition of the static graph, and use the dynamic version of the edge-convolution [50], allowing for independent dynamic graph estimation at every layer.

The final prediction is achieved through slow fusion [24, 54]. At every GCN layer, we merge the feature set from the dynamic layer with the feature set from the static layers. The final prediction is achieved using a shared fully connected layer and softmax activation over every node. This architecture is depicted in Figure 4.

3.5. Training and Implementation Details

Following [41], we implement a two-stream feature encoder based on the ResNet-18 architecture [19] pre-trained on ImageNet [13]. We perform the same modifications at the first layer to adapt for the extended input tensor (stack of face crops and spectrogram). We train the network end-to-end using the Pytorch library [38] for 100 epochs with the ADAM optimizer [26] using Cross-Entropy Loss. We use 3×10^{-4} as initial learning rate that decreases with annealing of $\gamma = 0.1$ at epochs 40 and 80. We empirically set $c = 11$ and augment the input videos via random flips and corner crops. Unlike other methods, MAAS does not require any large-scale audiovisual pre-training. We also incorporate the sampling strategy proposed by [2] in train-

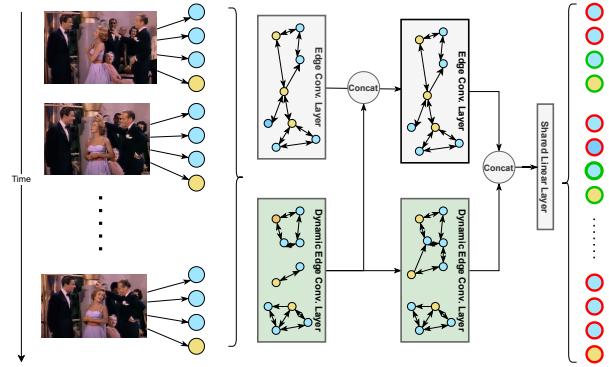


Figure 4. GCN Architecture in MAAS. Our graph neural network implements a two stream architecture. The first stream (top) uses the edge convolution operator and operates over static local and temporal graphs. The second stream (bottom) relies on dynamic edge convolutions and complements the feature embedding discovered by the static stream by means of slow fusion. After every GCN layer, we fuse the features from the dynamic and static streams and use them as input to the next layer.

ing to alleviate overfitting. During training, we follow the supervision strategy outlined by [41], where two extra auxiliary loss functions ($\mathcal{L}_a, \mathcal{L}_v$) are adopted to supervise the final layer of the audio and video streams. This favors the estimation of useful features from both streams.

Training MAAS After optimizing the feature encoder, we implement MAAS (LAN and TAN networks) using the PyTorch Geometric library [15]. We choose edge-convolution [51] to propagate the neighbor information between nodes. Our network model contains 4 GCN layers on both streams, each with filters of 64 dimensions. We apply dimensionality reduction to map features from their original 512 dimension to 64 using a fully connected layer. We find that this dimensionality reduction favors the final performance and largely reduces the computational cost.

Since we process data from different modalities, we use two different dimensionality reduction layers, one for video features and another for audio features. We train the MAAS-LAN and MAAS-TAN networks using the same procedure and set of hyper-parameters, the only difference being their underlying graph structure. We use the ADAM optimizer with an initial learning rate of 3×10^{-4} and train for 4 epochs. Both GCNs are trained from random weights and use a pre-activation [20] linear layer (Batch Normalization \rightarrow ReLU \rightarrow Linear Layer) to map the concatenated node features inside the edge convolution.

4. Experimental Validation

In this section, we provide an empirical analysis of our proposed MAAS method. We focus on the large-scale AVA-ActiveSpeaker dataset [41] to assess the performance of MAAS and present additional evaluation results on Talkies.

Method	mAP
<i>Validation Set</i>	
MAAS-TAN (Ours)	88.8
Alcazar <i>et al.</i> [2]	87.1
Chung <i>et al.</i> (Temporal Convolutions) [6]	85.5
Alcazar <i>et al.</i> (Temporal Context) [2]	85.1
Chung <i>et al.</i> (LSTM) [6]	85.1
MAAS-LAN (Ours)	85.1
Zhang <i>et al.</i> [60]	84.0
Sharma <i>et al.</i> [43]	82.0
Roth <i>et al.</i> [41]	79.2

<i>Test Set</i>	
MAAS-TAN (Ours) [6]	88.3
Naver Corporation [6]	87.8
Active Speaker Context [2]	86.7
University of Chinese Academy of Sciences [60]	83.5
Google Baseline [41]	82.1

Table 1. State-of-the-art Comparison on AVA-ActiveSpeaker. We compare MAAS against state-of-the-art methods on the AVA-ActiveSpeaker validation set. Results are measured with the official evaluation tool as published by [41]. We report an improvement of 1.7% mAP over the current state-of-the-art.

This section is divided in three parts. First, we compare MAAS with state-of-the-art techniques. Then, we ablate our proposal to analyse all of its individual design choices. Finally, we test MAAS on known challenging scenarios to explore common failure modes.

AVA-ActiveSpeaker Dataset. The AVA-ActiveSpeaker dataset [41] is the first large-scale testbed for active speaker detection. It is composed of 262 Hollywood movies: 120 of those on the training set, 33 on validation, and the remaining 109 on testing. The AVA-ActiveSpeaker dataset contains normalized bounding boxes for 5.3 million faces, all manually curated from automatic detections. Facial detections are manually linked across time to produce face tracks (tracklets) depicting a single identity. Each face detection is labeled as speaking, speaking but not audible, or non-speaking. All AVA-ActiveSpeaker results reported in this paper were measured using the official evaluation tool provided by the dataset creators, which uses mean average precision (mAP) as the main metric for evaluation.

4.1. State-of-the-art Comparison

We begin our analysis by comparing MAAS to state-of-the-art methods. The results reported for MAAS-TAN are obtained from a two-stream model composed of 13 temporally linked local graphs, which span about 1.59 seconds. We set $K = 3$ for the number of nearest neighbors in the dynamic stream and limit the number of video nodes to 4 per frame. The results reported for MAAS-LAN are obtained from a two-stream model, which includes a single timestamp and 4 video nodes. For sequences with 5 or more visible speakers, we make sure that one video node contains

Network Depth	mAP	Filters in Layer	mAP	Dynamic Graph	Static Graph	mAP	nNeighbors	mAP
1 Layer	88.0	32	88.5	✓	✗	66.5	2	88.5
2 Layer	88.2	64	88.8	✗	✓	87.9	3	88.8
3 Layer	88.4	128	88.6	✓	✓	88.8	4	88.4
4 Layer	88.8	256	88.1				5	88.4
5 Layer	87.5							

a) mAP by Network Depth

b) mAP by Network Width

Dynamic Graph	Static Graph	mAP	nNeighbors	mAP
✓	✗	66.5	2	88.5
✗	✓	87.9	3	88.8
✓	✓	88.8	4	88.4
			5	88.4

c) mAP by Individual Stream

d) mAP by Neighbors

Table 2. **Architecture choices in MAAS.** We ablate the design choices in our proposed GCN-based MAAS-TAN network. We analyse the network depth in a), and empirically find that a deeper network favors the final result, but saturates at 4 layers. We also analyse the number of filters per layer in b) and find the optimal to be at 64. From c), we observe that the static stream is far more effective by itself than the dynamic stream; however, the latter stream still incorporates information that is complementary leading to overall improvement. In d), we empirically find the most suitable number of neighbors in the dynamic stream and set it to 3.

the features from the active speaker, and randomly sample the remaining three. If no active speaker is present, we just randomly sample 4 speakers without replacement. At inference time, we split the speakers in non overlapping groups of 4, and perform multiple forward passes. Results in the validation set are summarized in Table 1.

Our best model, MAAS-TAN, ranks first on the AVA-ActiveSpeaker validation set. We highlight two aspects of these results. First, at 88.8% mAP, MAAS-TAN outperforms the best results reported on this dataset by at least 1.7%. It must be noted that some state-of-the-art methods [6, 60] rely on large 3D models and large-scale audiovisual pre-training, while MAAS uses only the standard ImageNet initialization for both streams. Second, while the MAAS-LAN network does not achieve state-of-the-art performance, it outperforms every other method that does not rely on long-term temporal processing [41, 60]. It also remains competitive with those methods that rely only on long-term context [6, 2], being outperformed only by the temporal version of [2] by a margin of 0.6% and falling 2.1% behind the full method of [2] (temporal context and multi-speaker).

4.2. Ablation Analysis

After evaluating the performance of our MAAS method against state-of-the-art techniques, we ablate our best model (MAAS-TAN) to validate the individual contributions of each design choice, namely: network depth, network width, independent stream contributions, and the number of neighbors for the dynamic stream.

Network Architecture We begin by ablating the proposed architecture. We explore the effects of changing the network depth, layer size, and the number of neighbors (K) for the dynamic stream. We also control for the individual contribution of each stream.

We summarize our ablation results for the MAAS-TAN network in Table 2. In 2-a), we identify the depth of the network as a relevant hyper-parameter for its performance. Shallow networks underperform, but increasingly get better as depth increases, reaching an optimal value at 4 layers. Deeper networks have a better capacity for estimating use-

ful features and have the chance to propagate relevant features over a large number of connected nodes, not only the immediate neighbors. In 2-b), we show that wider networks have a beneficial effect but saturate quickly with 64 or more filters. Beyond that size, the networks do not yield improvements at the expense of additional network complexity. In 2-c), we demonstrate the complementary nature of the two stream approach in MAAS. While the static stream has the best individual performance, the dynamic stream is capable of finding relationships that are beyond the insights we use to create the static graph structure, thus increasing the final performance by 0.9%. Finally, 2-d) shows how the selected number of clusters on the dynamic stream affects the final performance of MAAS. Interestingly, the optimal number of neighbors ($K = 3$) matches the number of valid assignments in the active speaker problem (audio with speech, active speaker), (audio with speech, silent speaker) and (audio with silence, silent speaker).

Graph Structure After assessing the design choices in the architecture, we proceed to evaluate the proposed graph structure. Here, we test for the incremental addition of LAN graphs into a TAN graph that analyses N timestamps. Additionally, we test for the maximum number of video nodes that get linked to an audio node at training time. Table 3 summarizes these results. Overall, we notice that MAAS benefits from modelling longer temporal sequences or modelling more visible speakers. We interpret this as a consequence of our modelling strategy that focuses on the assignation of locally consistent visual and audio patterns, while remaining compatible with the mainstream approach of modelling long-term temporal sequences.

4.3. Dataset Properties

We continue our analysis following the evaluation protocol of [41] and report MAAS-TAN results in known hard scenarios, namely multiple possible speakers and small faces.

In Table 4, we provide a breakdown of MAAS results according to the number of possible speakers. Overall, we see a significant performance increase when comparing MAAS to the AVA baseline [41] and improvements in all scenarios

Number of LANs	Per LAN Video Nodes				
	1	2	3	4	5
1	80.2	84.3	84.9	85.1	85
5	85.4	87.1	87.3	87.4	87.3
9	86.6	87.8	87.9	88.3	88.5
13	87.1	88.1	88.4	88.8	88.5
15	87.1	87.9	88.2	88.5	88.4

Table 3. **Graph Structure in MAAS.** We ablate the size of the MAAS-TAN network which is the core data structure of our approach. We empirically find it beneficial to model multiple speakers at the same time, and find the optimal number of speakers to be 4. Likewise, longer temporal sampling favors the performance but diminishes with sequences longer than 13 frames.

Number of Faces	MAAS	AVA Baseline [41]	ASC [2]
1	93.3	87.9	91.8
2	85.8	71.6	83.8
3	68.2	54.4	67.6

Table 4. **Performance evaluation by number of faces.** We evaluate MAAS according to the number of faces visible in the video frame. While performance decreases with more visible people, our method outperforms the AVA baseline and current state-of-the-art.

Face Size	MAAS	AVA Baseline [41]	ASC [2]
S	55.2	44.9	56.2
M	79.4	68.3	79.0
L	93.0	86.4	92.2

Table 5. **Performance evaluation by face size.** We evaluate MAAS in another challenging scenario: small and medium sized faces, which cover less than 128×128 pixels and 64×64 pixels, respectively. We observe that MAAS outperforms the current state-of-the-art, in most scenarios.

when compared to the multi-speaker stack of [2]. Clearly, the multi-speaker scenario is still quite challenging, but the improvements highlight that our speech assignation-based method is especially effective when two or more possible speakers are present.

In Table 5, we provide a breakdown of MAAS results according to the size of the face crop. We follow the evaluation procedure of [41] and create 3 sets of faces: (S) denotes faces smaller than 64×64 pixels, (M) denotes faces between 64×64 and 128×128 pixels, and (L) denotes any face larger than 128×128 pixels. Although MAAS does not explicitly addresses specific face sizes, we observe a large performance gap when compared to the AVA baseline, and we improve in most scenarios when compared to the method of Alcazar *et al.* [2]. We think this increase in performance is a consequence of better predictions in related faces, *i.e.* smaller faces are typically seen in cluttered scenes with multiple other visible individuals, so our method improves the prediction on these smaller faces by integrating more reliable information from other speakers.

5. The Talkies Dataset

Given the scarcity of in-the-wild active speaker datasets, we introduce “Talkies”, a manually labeled dataset for the active speaker detection task. Talkies contains 23,507 face tracks extracted from 421,997 labeled frames that yield a total of 799,446 individual face detections.

In comparison, the Columbia dataset [5] has about 150,000 face crops, while AVA-ActiveSpeaker [41] contains about 5.3 millions (760,000 in validation). Although AVA-ActiveSpeaker has a larger number of individual samples, we argue that Talkies is an interesting, complementary benchmark for three reasons. First, Talkies is more focused on the challenging multi-speaker scenario with 2.3 speakers per frame on average, while AVA-ActiveSpeaker averages only 1.6 speakers per frame. Second, Talkies does not focus on a single source of videos, as in AVA-ActiveSpeaker (Hollywood movies). As a consequence, Talkies contains a more diverse set of actors and scenes, with actors rarely overlapping between clips. This strikes a hard contrast with Hollywood movies, where a small cast takes most of the screen time. Finally, out of screen speech (another challenge for active speaker detection) is not common in Hollywood movies, but it appears more often in Talkies. More details about Talkies are provided in the **supplementary material**.

Training	MAAS	AVA Baseline [41]	ASC [2]
AVA	79.1	71.5	77.4
AVA augmented	79.7	N/A	N/A

Table 6. **Performance on Talkies.** We evaluate MAAS performance on the Talkies dataset. Without any fine-tuning on Talkies, MAAS (pre-trained on AVA-ActiveSpeaker) outperforms the baseline by 7.6% and the state-of-the-art by 1.7%. A simple augmentation targeting out of screen speech during AVA-ActiveSpeaker training leads to a direct improvement in the challenging scenes of Talkies.

Now, we evaluate the transferability of our MAAS method, trained on AVA-ActiveSpeaker, to the Talkies dataset. No fine-tuning is performed in this case. In Table 6, we compare the results of our best model (MAAS-TAN) against the AVA baseline of [41] and the ensemble model of [2] on our new dataset. MAAS outperforms these models by 7.6% and 1.7%, respectively. These results suggest that the core strategy proposed in MAAS is not domain-specific and can be applied to diverse scenarios without fine-tuning. Moreover, we explore an interesting attribute of Talkies, namely out of screen speech. To do so, we augment the training of MAAS on AVA-ActiveSpeaker, such that we randomly replace a silent audio track (its corresponding frames do not show any active speaker) with a track that contains speech. This simulates out of screen speech scenarios. This artificial substitution is done only during training and with a probability of 20%. This augmentation does not increase the amount of supervision at training time and

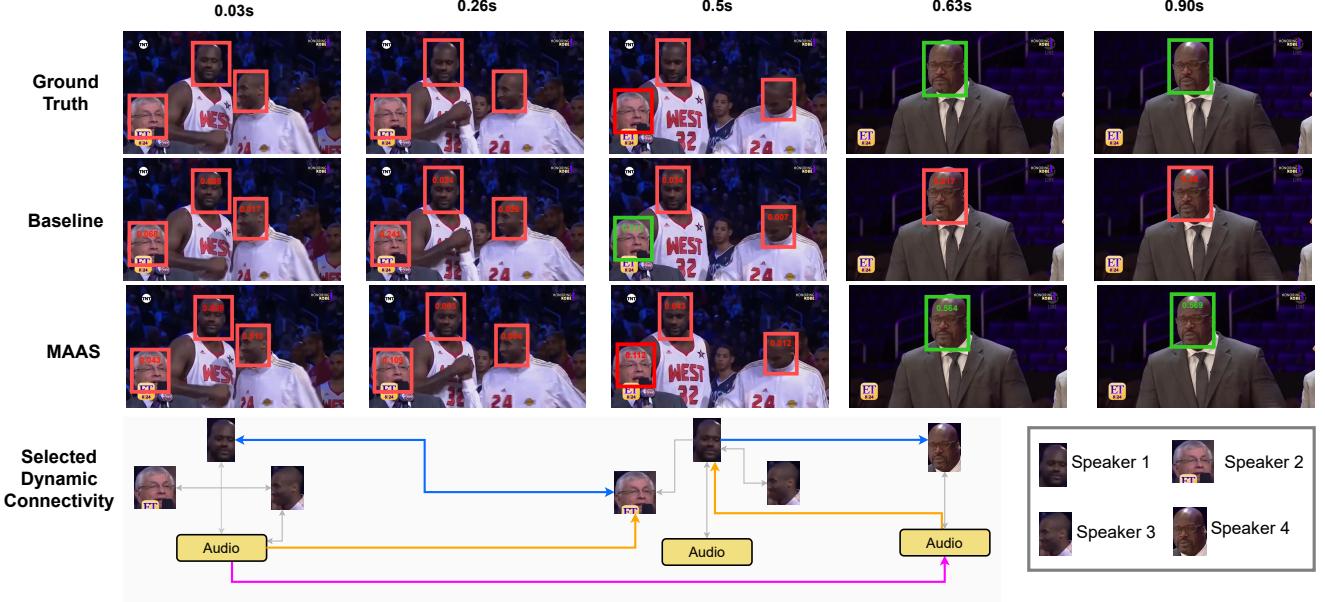


Figure 5. Qualitative Results. MAAS-TAN includes a dynamic stream that estimates graph structures from nearest neighbours in feature space. The connection patterns in this stream are very diverse and often create edges not present in our static graph. We find that such connectivity allows for information flow between distant audio clips (magenta), inter-speaker relations over multiple time-stamps, and cross-modal arcs that involve nodes in different frames (orange). For easier visualization, we only show a subset of all the dynamic connections.

has no empirical impact on the performance of MAAS on AVA-ActiveSpeaker, since out of screen speech is not common in Hollywood movies. However, this augmentation strategy does bring an improvement of 0.6% on Talkies, indicating that MAAS may be flexible enough to handle scenarios more general than those in AVA-ActiveSpeaker.

5.1. Qualitative Analysis

We conclude our assessment of MAAS by briefly looking at the connectivity patterns estimated by the dynamic stream. In Figure 5, we show a complex clip from the Talkies dataset, in this clip speaker 4 is the only active speaker. In fact, he narrates over the first frames of this clip. This scenario (out of screen speech) makes it very difficult for the baseline to generate accurate predictions on the first frames resulting in some false positive predictions (see speaker 2). MAAS on the other hand performs significantly better, reducing the false positives and effectively detecting the active speaker.

Empirically, we find that this clip mAP increases from 64% (baseline) to 97.9% (MAAS). We think this improvement is explained by two factors. First, MAAS builds more consistent temporal relationships for individual speakers, as its graph structure enforces consistent assignments across the temporal dimension. Second, the dynamic stream allows for unconventional, yet useful connectivity patterns. We show some of these patterns in the bottom row of Figure 5. In blue, we highlight inter-speaker connections across

timestamps. These connections are not part of the static graph structure, and can potentially encode semantic relationships between face crops. In magenta, we highlight audio-to-audio connections that go beyond our initial insight of linking adjacent audio clips. We think these connections allow for long-term temporal consistency between audio clips. Such consistency is key to resolve complex scenarios, as is the case with the narration in the selected clip. Finally, we highlight in orange cross-modal connections of nodes at different time-stamps. These connections also differ from those modeled in our static graph, and they reflect semantic similarities in the audiovisual embedding of MAAS.

6. Conclusion

We introduced MAAS, a novel multi-modal assignation technique based on graph convolutional networks, for the active speaker detection task. Our method focuses on directly optimizing a graph that simultaneously detects speech events and estimates the best source (active speaker). Additionally, we present Talkies, a novel benchmark with challenging scenarios for active speaker detection, which serves as a challenging transfer dataset for future research.

Acknowledgments. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding.

References

- [1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020. 1, 2
- [2] Juan Leon Alcazar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbelaez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2020. 2, 4, 5, 6, 7
- [3] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012. 1
- [4] Punarjay Chakravarty, Sayeh Mirzaei, Tinne Tuytelaars, and Hugo Van hamme. Who’s speaking? audio-supervised classification of active speakers in video. In *International Conference on Multimodal Interaction (ICMI)*, 2015. 2
- [5] Punarjay Chakravarty, Jeroen Zegers, Tinne Tuytelaars, et al. Active speaker detection with audio-visual co-training. In *International Conference on Multimodal Interaction (ICMI)*, 2016. 2, 7
- [6] Joon Son Chung. Naver at activitynet challenge 2019–task b active speaker detection (ava). *arXiv preprint arXiv:1906.10555*, 2019. 1, 2, 5, 6
- [7] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the conversation: speaker diarisation in the wild. *arXiv preprint arXiv:2007.01216*, 2020. 2
- [8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2
- [9] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, 2017. 2
- [10] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 2
- [11] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 2
- [12] Ross Cutler and Larry Davis. Look who’s talking: Speaker detection using video and audio correlation. In *International Conference on Multimedia and Expo*, 2000. 2
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4
- [14] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009. 2
- [15] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 5
- [16] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4930–4934. IEEE, 2017. 1
- [17] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. *arXiv preprint arXiv:1906.02739*, 2019. 2
- [18] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5
- [21] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016. 2
- [22] Arindam Jati and Panayiotis Georgiou. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1577–1589, 2019. 2
- [23] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018. 2
- [24] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 4
- [25] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *ACCV*, 2018. 2
- [26] D Kinga and J Ba Adam. A method for stochastic optimization. In *ICLR*, 2015. 4
- [27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1, 2
- [28] Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. Deepgcn: Can gcns go as deep as cnns? In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2
- [29] Guohao Li, Matthias Müller, Guocheng Qian, Itzel C. Delgadillo, Abdulellah Abualshour, Ali Thabet, and Bernard Ghanem. Deepgcn: Making gcns go as deep as cnns, 2019. 2
- [30] Guohao Li, Guocheng Qian, Itzel C. Delgadillo, Matthias Müller, Ali Thabet, and Bernard Ghanem. Sgas: Sequential greedy architecture search, 2019. 2
- [31] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. Deepgcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*, 2020. 2
- [32] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient

- subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018. 2
- [33] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*, 2018. 2
- [34] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018. 1, 2
- [35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 2
- [36] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011. 2
- [37] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2
- [38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS-Workshop*, 2017. 4
- [39] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5199–5208, 2017. 2
- [40] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *IEEE Spoken Language Technology Workshop (SLT)*, 2018. 1, 2
- [41] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava-activespeaker: An audio-visual dataset for active speaker detection. *arXiv preprint arXiv:1901.01342*, 2019. 1, 2, 3, 4, 5, 6, 7
- [42] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James Glass, and Trevor Darrell. Visual speech recognition with loosely synchronized feature streams. In *ICCV*, 2005. 2
- [43] Rahul Sharma, Krishna Somandepalli, and Shrikanth Narayanan. Crossmodal learning for audio-visual speech event localization. *arXiv preprint arXiv:2003.04358*, 2020. 1, 5
- [44] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass. Unsupervised methods for speaker diarization: An integrated and iterative approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2015–2028, 2013. 1, 2
- [45] Fei Tao and Carlos Busso. Bimodal recurrent neural network for audiovisual voice activity detection. In *INTERSPEECH*, pages 1938–1942, 2017. 2
- [46] Sue E Tranter and Douglas A Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565, 2006. 1
- [47] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989. 2
- [48] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno. Speaker diarization with lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5239–5243. IEEE, 2018. 1, 2
- [49] Quan Wang, Hannah Muckenhirk, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. *arXiv preprint arXiv:1810.04826*, 2018. 2
- [50] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay Sarma, Michael Bronstein, and Justin Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 2018. 1, 2, 4
- [51] 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. 5
- [52] Zhuyang Xie, Junzhou Chen, and Bo Peng. Point clouds learning with attention-based graph convolution networks. *arXiv preprint arXiv:1905.13445*, 2019. 2
- [53] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 2
- [54] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 2, 4
- [55] Sarthak Yadav and Atul Rai. Learning discriminative features for speaker identification and verification. In *Interspeech*, 2018. 2
- [56] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2
- [57] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018. 2
- [58] Chengzhu Yu and John HL Hansen. Active learning based constrained clustering for speaker diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2188–2198, 2017. 1
- [59] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. Fully supervised speaker diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019. 2
- [60] Yuan-Hang Zhang, Jingyun Xiao, Shuang Yang, and Shiguang Shan. Multi-task learning for audio-visual active speaker detection. 1, 2, 3, 5, 6