# Advancements in Audio-Visual Active Speaker Detection: A Novel Approach for the ActivityNet Challenge

## Abstract

This document outlines our contribution to the ActivityNet Challenge, focusing on active speaker detection. We employ a 3D convolutional neural network (CNN) for feature extraction, combined with an ensemble of temporal convolution and LSTM classifiers to determine whether a person who is visible is also speaking. The results demonstrate substantial improvements compared to the established baseline on the AVA-ActiveSpeaker dataset.

## 1 Introduction

The field of multimodal speech perception has garnered significant attention in recent times, with major advancements in audio-visual methodologies facilitated by deep learning. The capacity to identify which individuals are speaking at any moment is crucial for a variety of applications. The introduction of the AVA-ActiveSpeaker dataset has been a significant development, allowing for the training of deep-learning-based active speaker detection (ASD) models with complete supervision. This document provides a concise analysis of this dataset and elaborates on the methodology behind our submission to the challenge.

### 1.1 Datasets

The model is developed using the AVA-ActiveSpeaker dataset, which is divided into training, validation, and test sets, as detailed in Table 1. The ground truth labels are available for the training and validation sets.

Table 1: Statistical Overview of the AVA-ActiveSpeaker Dataset

| Set | Videos | Frames |
|---|---|---|
| Train | 120 | 2,676K |
| Val | 33 | 768K |
| Test | 109 | 2,054K |

This dataset presents several challenges. The durations of speaking segments are notably brief, with an average of 1.11 seconds for segments that are both spoken and audible. Consequently, the system needs to deliver precise detection with a limited number of frames. Traditional methods, which depend on smoothing the output over a time window of several seconds, are not effective under these conditions.

Additionally, the dataset includes many older videos where the audio and video recordings appear to have been captured separately or are significantly out of sync. As a result, the temporal alignment between audio and visual speech representations is not a reliable indicator of a person's speaking status.

.

# 2 Methodology

The active speaker detection system is composed of two primary components: front-end feature extractors and a back-end classifier, each discussed in detail in the subsequent sections.

## 2.1 Front-end architecture

For the extraction of audio and video representations, pre-trained networks are employed. These encoder networks have undergone training for the audio-visual correspondence task through a self-supervised approach on unlabeled videos.

The video encoder utilizes a convolutional neural network (CNN), processing 5 RGB image frames to produce a 512-dimensional representation. The architecture draws inspiration from the VGG-M network, known for its compactness and efficiency, but incorporates a 3D convolution in the initial layer instead of the conventional 2D convolution.

The audio encoder receives an input comprising 20 frames in the temporal dimension and 13 cepstral coefficients in the other, generating a 512-dimensional representation that aligns with the video representation's embedding space.

## 2.2 Back-end architecture

Both the audio and video encoders process an input of 5 video frames (equivalent to 0.2 seconds), advancing 1 video frame (0.04 seconds) at a time. Consequently, for an input of T frames, the output dimensions are 512 x (T - 4). In this study, two straightforward back-end classifiers are evaluated. Although our experiments utilize T = 9, no significant performance variations were noted for T values within the range of 7 to 15.

LSTM classifier. The audio and video representations are channeled into two distinct bi-directional LSTM networks, each comprising 2 layers with a hidden size of 128. The outputs from these networks are merged and subsequently processed through a linear classification layer. This layer determines whether the individual is speaking, and it is trained using the softmax cross-entropy loss.

TC classifier. In place of LSTM layers, the encoder outputs are directed to two temporal convolution layers, each equipped with 128 filters. The outputs are similarly concatenated and forwarded to the classifier, mirroring the approach used with the LSTM classifier.

Ensemble. Ensemble methods in machine learning have been demonstrated to frequently surpass the performance of any individual classifier. In this approach, the predictions generated by both the LSTM and TC classifiers are averaged with equal weighting to produce the final prediction.

Smoothing. To mitigate noise within the predictions, the outputs of the classifiers undergo temporal smoothing using either a median or Wiener filter, both applied over 0.5-second intervals.

# 3 Experiments

Our model, implemented using the PyTorch library, was trained on a single Tesla M40 card with 24GB of memory. Training utilized the ADAM optimizer with default settings and a fixed learning rate of $10^{-2}$. To counteract any bias in the training data, the number of samples for positive and negative classes was balanced within each mini-batch during the training process.

The evaluation metric for this task is the mean Average Precision (mAP), with the evaluation code supplied by the challenge organizers.

Results on the validation set for the various back-end classifiers are presented in Table 2. The best model achieved an mAP of 0.878 on the sequestered test set for the challenge. In contrast, the GRU-based baseline model yielded an mAP of 0.821.

The qualitative outcomes of the proposed method significantly surpass those of existing correspondence-based methods on this dataset because it does not depend on accurate audio-to-video synchronization.

Table 2: Performance Evaluation on the AVA-ActiveSpeaker Validation Set

| Back-end | Smoothing | mAP |
|---|---|---|
| LSTM | X | 0.851 |
| TC | X | 0.855 |
| Ensemble | X | 0.861 |
| Ensemble | Median | 0.874 |
| Ensemble | Wiener | 0.878 |