Active Speaker Detection as a Multi-Objective Optimization with Uncertainty-based Multimodal Fusion

Baptiste Pouthier\textsuperscript{1,2}, Laurent Pilati\textsuperscript{1}, Leela K. Gudupudi\textsuperscript{1}, Charles Bouveyron\textsuperscript{2}, Frederic Precioso\textsuperscript{2}

\textsuperscript{1}NXP Semiconductors, France
\textsuperscript{2}Université Côte d’Azur, Inria, CNRS, LJAD, I3S, Maasai, France
\{baptiste.pouthier, laurent.pilati, leela.k.gudupudi\}@nxp.com
\{charles.bouveyron, frederic.precioso\}@univ-cotedazur.fr

Abstract

It is now well established from a variety of studies that there is a significant benefit from combining video and audio data in detecting active speakers. However, either of the modalities can potentially mislead audiovisual fusion by inducing unreliable or deceptive information. This paper outlines active speaker detection as a multi-objective learning problem to leverage best of each modalities using a novel self-attention, uncertainty-based multimodal fusion scheme. Results obtained show that the proposed multi-objective learning architecture outperforms traditional approaches in improving both mAP and AUC scores. We further demonstrate that our fusion strategy surpasses, in active speaker detection, other modality fusion methods reported in various disciplines. We finally show that the proposed method significantly improves the state-of-the-art on the AVA-ActiveSpeaker dataset.

Index Terms: active speaker detection, audiovisual, multimodal fusion, multi-objective

1. Introduction

Active Speaker Detection (ASD) contemplates on identifying active speakers in a video by analyzing both visual and audio features. Hence, ASD is inherently multimodal in nature, where video and audio data are essential attributes. In recent years there has been considerable interest in the ASD methods based upon audio-visual cues [1, 2, 3, 4, 5]. Despite this interest, the lack of large-scale in-the-wild datasets impeded scientific progress in the field. Unfortunately, most of the prior works are challenged by skewed results owing to the poor quality of the considered datasets.

The recent AVA-ActiveSpeaker dataset [6] can potentially overcome these limitations and reshape the ASD field of study. Together with the large in-the-wild dataset, the authors presented a baseline model based on a two-stream network that merges video and audio modalities in an end-to-end fashion. Within the annual AVA-ActiveSpeaker challenge [7], Chung [8] and Zhang et al. [9] improved this baseline using hybrid 3D-2D CNNs pre-trained on large-scale multimodal datasets [10]. Unfortunately, this approach encounters two major challenges in practice, as illustrated in Fig.1: (1) multi-speaker scenario where multiple persons in a video frame are speaking, and (2) low-resolution and/or indiscernible faces in video frames. In [11], Alcázar et al. addressed the multi-speakers scenario by learning long-term relationships between speakers, and Huang et al. [12] handled the uncertainty in video modality by adding optical flow to raw pixel representation to strengthen face characterisation. Nevertheless, these studies focused on ad-hoc objectives with limited scope and little attention has been paid to the aggregated approaches which exploit correlation between both the challenges. In this paper, we propose to learn this correlation using a cross-modal fusion that involves simultaneous learning of the uncertainty in both the modalities using a multi-objective learning scheme as described in Section 2.1.

Traditionally, in multi-task learning, uncertainty is handled by learning adaptive weighting between each task’s loss functions [13, 14]. In the field of Automatic Video Description, researchers have investigated attention-based [15] fusion mechanisms that capture the importance of each of the modalities. Hore et al. [16] proposed a model that selectively uses features from different modalities. This approach was later improved using hierarchical attention fusion in [17]. Despite the growing interest of these fusion mechanisms in other disciplines, to the best of our knowledge, no studies have been conducted on ASD problem. The present paper aims to investigate different fusion schemes to solve the ASD problem. We also propose a novel audiovisual fusion mechanism as a first attempt to enrich the self-attention model with uncertainty information to disentangle the practical challenges.

The proposed method consists of learning a self-attention [15] and uncertainty-based fusion mechanism that weights video and audio embeddings in an end-to-end fashion. We use the typical two-stream DNN architecture from the literature [6, 11], with a stream per each modality, to encode both embeddings. By characterizing uncertainty of each modality, we aim at disentangling the problem presented in Fig.1. Indeed, the fusion scheme we propose determines the viability of each modality at any given time to learn a comprehensive understanding of every situation towards ASD disambiguation. Our solution significantly outperforms state-of-the-art in the AVA-ActiveSpeaker dataset by 4.8% and 1.7% on validation and test sets, respectively.
2. Active Speaker Detection

Fig.1 demonstrates video and audio modalities may conflict, with non-talking faces affiliated with speech-labeled audio. Conflicting video and audio information lead to a compromised audio-visual learning. Most recent ASD studies [6, 8, 9, 11, 12] defined a trade-off optimizing both video and audio streams towards video label, sometimes using auxiliary classifiers to increase the discriminative power of the individual streams towards the unique objective [6, 11]. However, this approach undermines the learning of the audio stream. In this section, we detail the multi-objective optimization of the video and audio streams towards their respective ground-truth labels to learn unbiased and accurate representations of video and audio modalities. Then, we introduce a novel audiovisual fusion mechanism that uses self-attention and uncertainty indicators to better disentangle the ambiguous scenarios.

2.1. Multi-Objective Learning

Let \((X_t = (x_t^n)_{n \in \mathbb{N}})_{n \in \mathbb{N}}\) be a series of consecutive frames, where \(x_t^n\) denotes the \(n^{th}\) face detected in the \(t^{th}\) frame, and \((A_t)_{n \in \mathbb{N}}\) is the corresponding audio segment. Each audiovisual sequence is noted \(M_t = (x_t^n, A_t)_{n \in \mathbb{N}}\). During training, each multimodal sample \(M_t\) is associated with a video-based ground-truth label \(y_t^V\) where \(y_t^V = 1\) if the face \(x_t^n\) is speaking and \(y_t^A = 0\) otherwise. To cater multi-objective learning, we also define audio-based ground-truth by aggregating video-based labels of each frame using Eq.1:

\[
y_t^A = \begin{cases} 
0 & \text{if } \sum_{n} y_t^n = 0, \\
1 & \text{otherwise.} 
\end{cases}
\]  

We train a standard two-stream architecture, presented in Fig.2, that leverages auxiliary classifiers (Aux) to encourage video and audio networks to minimize their own loss function, thus optimizing both modalities intermediate feature representations. The outputs of the Video, Audio and Multimodal networks are denoted as \(y_t^V, y_t^A\), and \(y_t^M\) respectively; we formulate these quantities as \(y_t^M = f^W (x_t^n)\), \(y_t^A = f^W (A_t)\), and \(y_t^M = f^W (M_t)\) where \(W_V\), \(W_A\), and \(W_M\) are the weights of the respective networks. To train these weights, we define the loss function \(L\) as a cross-entropy loss function \(L(y, \hat{y}) = -\sum y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)\). The final multi-objective loss-function \(L_f\) is formulated as \(L_f = L_M + L_V + L_A\), where \(L_M = L(y^V, \hat{y}^V)\), \(L_V = L(y^V, \hat{y}^V)\), and \(L_A = L(y^A, \hat{y}^A)\) respectively.

2.2. Cross-Modality Fusion

This section describes a multimodal fusion that focuses on disentangling ambiguous scenarios in ASD problem with self-attention and uncertainty estimation. The key strategy here is to consider dynamically varying weights for each modality at any given time by computing an estimate of their relative uncertainty. Fig.4 illustrates the concept of our fusion mechanism, inspired by the method presented in [17], that fuses the information from both modalities in a hierarchical fashion. First, we feed each video \((u_t^V)_{t=1}^{T}\) and audio \((u_t^A)_{t=1}^{T}\) embedded sequences to isolated 128-dim BiGRU layers with hidden state \(h_t^V, h_t^A\). Then, we compute the multimodal embedding \(u_t^M\) using weighted concatenation as given by Eq.2:

\[
u_t^M = \lambda_t^V h_t^V \oplus \lambda_t^A h_t^A
\]  

where the weights \(\lambda_t^V\) and \(\lambda_t^A\) characterize, at each time \(t\), the uncertainty of the video and audio modalities. Finally, we use post-fusion 100-dim BiGRU layers with hidden state \(h_t^M\), a 2-dimensional Fully Connected (FC) layer, and a softmax.

In [16, 17], the \(\lambda\) weights are learnt using perceptrons that fully-connect, at any given time, the hidden representation of each modality. This paper presents a novel approach to compute the weights of the modalities using: (1) an estimation of the uncertainty of both video and audio embedded representations, and (2) self-attention to measure the importance of video and audio modalities in their local temporal context.

High-Level Attributes: To characterize the uncertainty, we first rely on a high-level assessment of the quality of the video and audio data. We associate each multimodal training sample \(M_t^n\) with two attributes \(\eta_t^n\) and \(\mu_t^n\) that characterize the number of detected faces in frame \(X_t\) and the resolution of the targeted face \(x_t^n\), respectively. Indeed, higher the number of potential speakers lower the chance for the audio to be discriminative. Similarly, low-resolution face coincides with prediction uncertainty as it is more difficult to interpret facial cues and lips movements on a smaller face thumbnail. We denote \(\eta_t = \text{card}(X_t)\) and \(\mu_t^2\) the number of pixels in the face thumbnail \(x_t^n\) before any resizing. To scale the two quantities \(\eta\) and \(\mu\), we first use data binning then target encoding [18] to replace each value with a blend of posterior and prior...
Auxiliary Uncertainty: We leverage auxiliary classifiers predictions to estimate the uncertainty in video and audio embeddings. Intuitively, the data leading to a highly polarized auxiliary decision can be considered as reliable. The output of a softmax layer cannot be used reliably as true probabilities [19, 20, 21]. Therefore, we use a simple yet effective auxiliary decision can be considered as reliable. The output of a softmax layer cannot be used reliably as true probabilities [19, 20, 21]. Therefore, we use a simple yet effective auxiliary classifier.

\[ \delta_t = \frac{\tilde{y}_{V,A} - y_{V,A}}{\max(\tilde{y}_{V,A}, 1.7\%)} \]

Self-Attention: We aim at using dot-product global self-attention [15] mechanism to evaluate how important video and audio auxiliary predictions, respectively. Given \( \tilde{y}_{V,A} \) the temperature-scaled probabilities of the video or audio auxiliary output, we define \( \delta_t \) the uncertainty values linked to the video and audio auxiliary predictions, respectively.

\[ S_{V,A} = \text{softmax}(H_{V,A}H_{V,A}^T + B) \]

Binarization: DAGRAGD optimizer [22] is used with a learning rate of 0.015 to train the network in end-to-end fashion for 20 epochs with mini-batches of 16 sequences and without any pre-training. Roth et al. [6] demonstrated that stacking few consecutive frames as input of the first 2D convolutional layer is beneficial to learn short temporal motion. Therefore, the input to the visual network (Fig.3b) is a stack of 3 consecutive grayscale face thumbnails. The faces are extracted using the provided bounding boxes and resized to 224x224. We feed the audio embedding network (Fig.3c) with 13 MFCC features extracted from the preceding 0.5s of audio with a 25ms window and a 10ms step. Our BiGRUs are trained with 1.12s long segments (28 frames) to capture most of the different speech patterns within the A V A-Active-Speaker dataset, the average continuous speech duration being 1.11s [6].

Metrics: For ease of comparison with previous studies, we evaluate the proposed method using the official ActivityNet Challenge evaluation script [6] that computes the mean Average Precision (mAP) score. When available, Area Under Receiver Operating Characteristic Curve (AUC) score is also provided in Table 1.

3. Evaluation and Analysis

3.1. Experimental Setup

AVA-ActiveSpeaker Dataset [6]: It is the most comprehensive, largest, and challenging publicly available dataset for audiovisual ASR problem. It consists of 262 movies divided into training (120), validation (33) and test (109) sets. In total, 5,498K faces are labeled with normalized bounding boxes. For training and validation sets, ground-truths on whether someone is speaking are also provided. The ground-truths for the test set are withheld for the annual ActivityNet Challenge [7].

Training Strategy: The ADAGRAD optimizer [22] is used with a learning rate of 0.015 to train the network in end-to-end fashion for 20 epochs with mini-batches of 16 sequences and without any pre-training. Roth et al. [6] demonstrated that stacking few consecutive frames as input of the first 2D convolutional layer is beneficial to learn short temporal motion. Therefore, the input to the visual network (Fig.3b) is a stack of 3 consecutive grayscale face thumbnails. The faces are extracted using the provided bounding boxes and resized to 224x224. We feed the audio embedding network (Fig.3c) with 13 MFCC features extracted from the preceding 0.5s of audio with a 25ms window and a 10ms step. Our BiGRUs are trained with 1.12s long segments (28 frames) to capture most of the different speech patterns within the A V A-Active-Speaker dataset, the average continuous speech duration being 1.11s [6].

Table 1: Performance comparison of the proposed method with the state of the art. Results are reported on both validation and ActivityNet Challenge hidden test sets.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Validation subset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our method</td>
<td>91.9</td>
<td>96.3</td>
</tr>
<tr>
<td>Active Speakers Context [11]</td>
<td>87.1</td>
<td>-</td>
</tr>
<tr>
<td>Huang et al. [12]</td>
<td>-</td>
<td>93.2</td>
</tr>
<tr>
<td>Google Baseline [6]</td>
<td>86.3</td>
<td>92.0</td>
</tr>
<tr>
<td>Naver Corp. (Temporal Convolutions) [8]</td>
<td>85.5</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al. [9]</td>
<td>84.0</td>
<td>-</td>
</tr>
<tr>
<td><strong>ActivityNet Challenge Leaderboard</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our method</td>
<td>89.5</td>
<td>-</td>
</tr>
<tr>
<td>Naver Corp. [8]</td>
<td>87.8</td>
<td>-</td>
</tr>
<tr>
<td>Active Speakers Context [11]</td>
<td>86.7</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al. [9]</td>
<td>83.5</td>
<td>-</td>
</tr>
<tr>
<td>Google Baseline [6]</td>
<td>82.1</td>
<td>-</td>
</tr>
</tbody>
</table>

3.2. Comparison with State Of The Art

Experimental results in Table 1 show that the proposed method outperforms all existing approaches on both test and validation sets, in terms of mAP and AUC scores. It is worth highlighting that the proposed method outperforms one of the best approaches in [11] by a significant margin of 4.8% and 2.8% in mAP score on the validation and test sets respectively. Furthermore, our approach surpasses Naver’s method [8] by 1.7% and ranks first in the A V A-Active-Speaker Leaderboard.
Table 2 compares the number of parameters of the top-ranked models on the ActivityNet Challenge Leaderboard. Our results are very favorable since our approach has significantly fewer parameters compared to the state of the art and does not necessitate any pre-training and/or ensemble-models.

Table 2: Comparison of the gross number of parameters.

<table>
<thead>
<tr>
<th>Method</th>
<th>#params</th>
<th>pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>2M</td>
<td></td>
</tr>
<tr>
<td>Naver Corp. [8]</td>
<td>13M</td>
<td>✔</td>
</tr>
<tr>
<td>Zhang et al. [9]</td>
<td>22M</td>
<td>✔</td>
</tr>
</tbody>
</table>

3.3. Performance Breakdown

Multi-objective learning, as formulated in Section 2.1, aims at (1) learning accurate representations of each modality, and (2) allowing unbiased estimation of the uncertainty of video and audio intermediate features. We therefore evaluate the discriminative power of video and audio embedded representations. Figures 5a and 5c detail the performance of the embedding networks detailed in Fig.3 when used alone. The presented results are with 128-dim BiGRU added on top of each embedding network. We compare the performance of the audio embedding network optimized towards either video or audio ground-truth labels as discussed in Section 2.1. As expected, the performance of the audio embedding network trained with video ground-truths strongly degrades on the number of speakers. It suffers in the ambiguous scenario presented in Fig.1 where multiple persons share the same audio track. On the contrary, the performance of the audio network trained with audio labels is almost constant while increasing the number of speakers. We also observed a major mAP score increase of 26.9% compared with using video labels. Thus, the multi-objective approach using an independent audio-based labels allows the reliability on the audio network in difficult/ambiguous scenarios.

Figures 5b and 5d compare the performance of our fusion method with the Naïve and Hierarchical [17] Fusion schemes in different scenarios. Our multi-objective model is first evaluated using a naïve concatenation fusion (NF) to combine video and audio modalities. Here it is crucial to note that the method improves Active Speakers Context [11] mAP by 3.1% on validation subset (Table 1). This result highlights the effectiveness of our end-to-end multi-objective learning. Hierarchical Fusion (HF) refers the fusion scheme presented in [17]. Note that our adaptation implies a slight modification of the initial method to match our “many-to-many” architecture. As shown in Fig.5b, the proposed fusion scheme clearly has an advantage over NF and HF. Fig.5d presents additional comparative analysis results by varying the number of faces detected (left) and the size of the face thumbnails (right). The proposed method clearly outperforms both NF and HF schemes, especially in challenging scenarios.

4. Conclusion

In this paper, we proposed a self-attention and uncertainty-based fusion mechanism that learns a comprehensive understanding of every situation towards ASD disambiguation. Our approach catered multi-objective optimization to encourage the learning of unbiased multimodal features. Experimental results on the challenging AVA-ActiveSpeaker dataset demonstrate that the proposed method achieves superior performance than existing methods. Besides, the proposed method outperformed the state-of-the-art on both validation and test datasets and ranked first in the ActivityNet Challenge, despite having fewer parameters and without any pre-training.

5. Acknowledgements

This work has been supported by the French government, through the 3IA Côte d’Azur Investment in the Future project managed by the National Research Agency (ANR) with the reference numbers ANR-19-P3IA-0002.
6. References


