Towards Robust Active Speaker Detection

Siva Sai Nagender Vasireddy[1] , Chenxu Zhang[2] , Xiaohu Guo[1], Yapeng Tian[1]
[1] The University of Texas at Dallas, [2] ByteDance, USA

1. Introduction

Active Speaker Detection (ASD) is the task of identifying the visible speakers in each frame of a video. It is an essential multimodal problem, and both facial dynamics and speech characteristics in sound provide strong cues.

Existing approaches for active speaker detection focus on designing effective neural networks that can leverage temporal and multimodal context information in videos [13, 14, 11, 2, 4], in which both audio and visual modalities are fully exploited. However, the input modalities could be noisy and unreliable, particularly the audio. In real-world scenarios, non-speech sounds are common in the environment surrounding the active speaker. As shown in Fig. 1, the audio track contains both speech and strong cafeteria noise. Besides speech, current approaches will also encode undesired audio information into the representation, which can negatively impact active speaker detection performance.

In this paper, we formulate the robust active speaker detection (rASD) problem to address the issue. The goal of rASD is to detect active speakers in videos with the presence of non-speech sounds in the surrounding environment. A naive solution to this problem is to separate the speech sound from the noisy audio mixture and then feed the separated sound to the existing ASD approaches. To train a separator, we can use the mix-and-separate strategy [15, 7] by randomly sampling non-speech audio and mixing it with clean speech sound. Recent audio-visual audio separation and speech enhancement models [6, 15, 1, 8] can be used as separators to address the problem. However, speech separation and enhancement is a challenging task, and even state-of-the-art methods can leave residual noise in the separated speech. The speech quality is also reduced compared to the original clean speech. Furthermore, the speech sounds in the training data used for ASD [13] can be noisy as they are collected from the web. Using these speech sounds with inherent noise as groundtruth for training the separator may lead to inferior performance.

To overcome these challenges, we propose a novel framework for robust active speaker detection aimed at addressing the issue of audio noises in the surrounding environment of the active speaker. Instead of separating speech from noisy audio, we utilize audio-visual speech separation as guidance to learn noise-free audio features. These features are then utilized in an active speaker detection model. As a result, we can learn speech separation and active speaker detection simultaneously in a multi-task learning manner. Both tasks will use the same audio features, and the features will be enforced to be clean and helpful for the active speaker detection task. This approach mitigates the residual noise and audio quality reduction issues and enables the two tasks to be jointly optimized in an end-to-end framework. To handle inherent noise in speech sounds and further enhance the robustness of audio features, we propose a dynamic weighted loss approach to train the speech separator. In this approach, we reduce the importance of audio samples with inherent noise during training using weights that are dynamically generated.

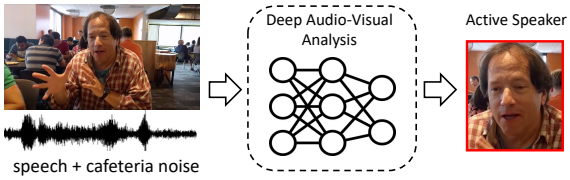

Figure 1: Given a video with both audio and visual tracks, we develop a robust deep audio-visual analysis model that can detect active speakers even in a noisy environment.

In addition, we collected a real-world noise audio dataset consisting of 1,350 non-speech sounds from 27 different categories. Experiments demonstrate that non-speech audio noises can significantly impact ASD models. Our approach is capable of learning robust speech sound features, which can improve ASD performance in noisy environments. Moreover, the proposed framework is general and can be applied to several different ASD approaches to improve their robustness. Please check out our demo.

2. Method

Problem Formulation. The goal of robust Active Speaker Detection (rASD) is to identify visible speaking faces in a given video frame, while accounting for the presence of audio noise. Specifically, the audio clip corresponding to the video frame is a sound mixture: $A = A_{\text{speech}} + A_n$, which contains speech from active speakers and irrelevant audio noise from other sound sources in the surrounding environ-

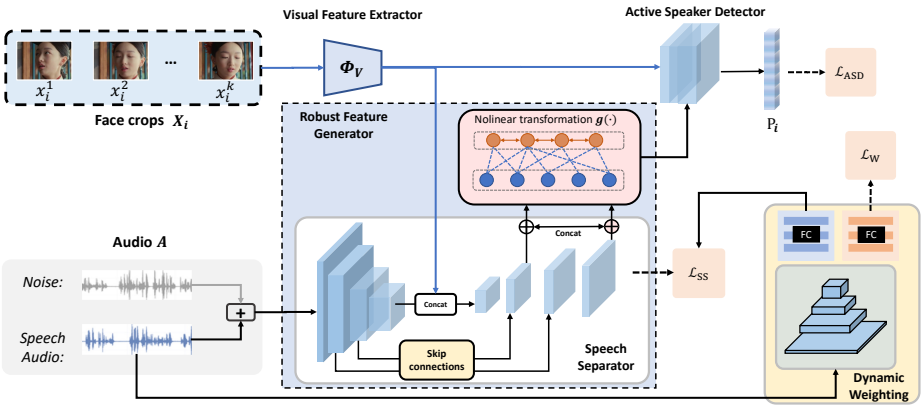

Figure 2: The proposed robust active speaker detection framework. In this framework, we utilize an audio-visual speech separator to guide the learning of noise-free speech features for active speaker detection. The framework includes a nonlinear transformation $g(\cdot)$ to bridge the features between the separator and the detector. In addition, a dynamic weighting mechanism is employed to generate dynamic weights for the separation loss, which helps handle inherent speech noises. The framework is general and can be applied to improve the robustness of any existing audio-visual active speaker detectors.

ment. The task of rASD is highly challenging, as it requires effective utilization of the spatial, temporal, and multimodal contexts present in both audio and visual modalities while accounting for potential noise in the audio signal.

2.1. Overview

Active Speaker Detection Pipelines. Most existing ASD methods [13, 14, 11, 2, 3, 4] can be generalized into four main modules: the visual feature extractor Φ_V, the audio feature extractor Φ_A, the audio-visual fusion module Φ_{AV}, and the speaker detection module Φ_D. These modules work together to detect the active speaker(s) in a video. The inputs to these methods are a set of face crops $X_i = \{x_i^1, x_i^2, ..., x_i^k\}$ for each person i, with a sequence of k face crops, and the corresponding audio A. The visual feature extractor Φ_V generates sets of feature representations $F_v^i = \Phi_V(X_i)$ of the face crops, while the audio feature extractor Φ_A generates sets of feature representations $F_a = \Phi_A(A)$ of their corresponding audio waveforms. The audio-visual fusion module Φ_{AV} then generates a set of audio-visual features $F_{av}^i = \Phi_{AV}(F_a, F_v^i)$ for each person i, which contains features of each of the face crops in X_i. Finally, the speaker detection module Φ_D generates the set of predictions $P_i = \Phi_D(F_{av}^i)$, which contains a prediction score p_i^k for each of the face crops x_i^k. A loss function $\mathcal{L}_{ASD}$ is computed using the standard Cross-entropy loss.

Our Approach. Rather than developing an advanced active speaker detection approach, our focus is on creating a sturdy audio-visual framework that can enhance the robustness of any existing active speaker detection method. Our approach centers on training a more robust feature generator to replace the existing Φ_A. To improve the robustness of our model, we employ an audio-visual speech separation model to guide the robust audio feature learning. To address the presence of inherent audio noise in speech sounds and further reduce the effect of noise on encoded audio features, we introduce a dynamic weighted loss for separation. Figure 2 illustrates an overview of our rASD framework.

2.2. Robust Audio Feature Generation

We propose a robust audio feature generation module that produces robust audio features F_a from noisy speech audio A. Our approach involves first computing the magnitude of the spectrogram χ_A of the audio and then generating the robust audio features $F_a = \Phi_{RFG}(\chi_A)$. The robust audio features F_a can be integrated into active speaker detection pipelines to improve their robustness against real-world audio noise. We incorporate a speech separator Φ_{SS} into our framework to ensure that the generated audio features contain noise-free speech information. The separator acts as guidance for our model.

Speech Separator. To learn a robust feature generator that can produce noise-free speech sound features for rASD, we use a speech separator Φ_{SS} as learning guidance, which can separate speech audio S_{speech} from noisy audio A. Similar to [15, 7], a U-Net [12] architecture is used as a speech separator module Φ_{SS}. We adopt the commonly used Mix-and-Separate strategy to train Φ_{SS}. In our implementation, we generate noisy speech audio $A = A_{speech} + \alpha A_n$ by mixing speech audio A_{speech} from the AVA-ActiveSpeaker [13] dataset with randomly sampled real-world noise samples A_n, where α controls the noise level being added to the speech sound. The separator takes the magnitude of the noisy spectrogram $\chi_A \in \mathbb{R}^{F \times T}$ as input and will output a separation mask $M_{\text{pred}} \in \mathbb{R}^{F \times T}$ with the help of tracked faces in video frames. The spectrogram of the separated speech sound can be reconstructed by $\chi_{A_{\text{speech}}} = M_{\text{pred}} \odot \chi_A$, where $\odot$ represents element-wise multiplication operation.

We compute the groundtruth ratio mask M_{gt} as an element-wise ratio of $\chi_{A_{speech}}$ and χ_A: $M_{gt}(p,q) = \frac{\chi_{A_{\text{speech}}}(p,q)}{\chi_A(p,q)}$. To train the separator, we use an L1 loss function to compute the loss L_{SS} between predicted and groundtruth masks: $\mathcal{L}_{SS} = \|M_{gt} - M_{pred}\|_1$.

Robust Feature Generator. Our approach to generating robust audio features is based on the premise that the set of intermediate feature maps F_{in} from the decoder layers of the speech separator Φ_{SS} contain rich information that is relevant to the speech sound A_{speech}. These feature maps F_{in} carry the information required to separate the speech audio A_{speech} from the input noisy audio A. This characteristic of F_{in} makes it an ideal resource for generating speech audio features that are robust to the noise A_n. The robust audio features F_a are generated from F_{in} using transformation layers in robust feature generation module, Φ_{RFG}.

The module Φ_{RFG} is built on top of the separator Φ_{SS} by adding a sequence of nonlinear transformations formulated by a function $g(\cdot)$ that generates robust audio features $F_a = g(F_{in})$ from a set of feature maps F_{in} generated by Φ_{SS}. Φ_{RFG} is trained in an end-to-end manner with the ASD loss. Intuitively, the additional nonlinear transformations in Φ_{RFG} bridge the domain gap between audio features that can separate speech audio from noise and the audio features that are required for the ASD task.

We investigated multiple combinations of feature maps from F_{in} that can be utilized as input to the nonlinear transformation function, $g(\cdot)$, and found that a combination of feature maps, one closer to the bottleneck of the encoder-decoder U-Net architecture of Φ_{SS} and the other closer to the output of Φ_{SS} produces the best results. We hypothesize that this combination of feature maps represents the higher-level and lower-level features of the speech audio, where the feature map nearer to the separation output contains the cleanest speech sound information, while the one near the bottleneck contains more high-frequency speech patterns.

2.3. Robustness to Inherent Noise

The speech separator Φ_{SS} is trained with audio pairs (A, A_{speech}), with A as input and A_{speech} as target. However, audio in web videos could be noisy, and it may contain inherent noise n_{in} even before adding external noise, *i.e.*, $A_{\text{speech}} = A_c + A_{n_{in}}$, where A_c is the clean speech component of A_{speech}. The presence of $A_{n_{in}}$ in A_{speech} reduces the quality of speech representation in the feature maps F_{in} of the speech separator Φ_{SS}. The degradation in the quality of F_{in} occurs because the feature maps also contain information about $A_{n_{in}}$ along with A_c and groundtruth mask as the learning label becomes noisy.

To alleviate the negative impact of $A_{n_{in}}$, we propose reducing the importance of audio samples with inherent noise by implementing a weighted separation loss approach. As different samples may have varying levels of noise, we present a dynamic weight generation approach designed to dynamically generate weights during training. This method helps to mitigate the adverse effects of inherent noise and enhances the robustness of our approach.

Weighted Loss For Separation. In web videos, a significant portion of speech sound contains either music or noise [5]. The noisy labels will decrease speech separation performance and, consequently, affect the audio features which will be used for rASD. To handle the inherent audio noises and strengthen the robustness of the speech sound representations, we use a weighted separation loss to optimize the speech separator. The separation loss is computed as the mean of losses of each sample in a batch, thus giving equal importance to each sample irrespective of the presence of inherent noise in speech audio. In our weighted loss approach, we assign a weight $w_k \in [0, 1]$ to each training sample k: $\mathcal{L}^k_{SS} = w_k \cdot \|M^k_{gt} - M^k_{pred}\|_1$. The loss will be used to train the speech separator to increase the robustness of the features to inherent noise.

Dynamic Weight Generator. One question that remained unanswered in implementing the weighted loss is how to obtain the weights. A naive strategy that we can employ is to set w_k as a fixed scalar that is less than 1 for the samples with non-speech noise, thereby lowering the importance of these samples with inherent noises. However, samples can differ significantly, even when they all contain noise.

Rather than using a fixed weight, we propose a dynamic weight generator, Φ_W, that can be trained to predict weights based on the given audio input. This method enables us to generate weights that are tailored to the specific characteristics of each sample, further improving the robustness of our approach. The architecture of Φ_W consists of a sequence of 2D convolutional layers, followed by two sets of fully connected layers, with ReLU being used as the activation function. It has two branches: one predicts the sound type class of A, while the other predicts the training sample weight. We train Φ_W using the loss $\mathcal{L}_W = \mathcal{L}_C + \frac{1}{b}\sum_{k=1}^{b} |w_k - 1|$, where b is the number of samples for which the loss is being computed, and $\mathcal{L}_C$ is the cross-entropy loss computed over the sound type classification. The second term serves to prevent an all-zero shortcut.

2.4. Loss function for our approach to rASD

The final loss function of our robust ASD model is $\mathcal{L} = \lambda_1\mathcal{L}_{ASD} + \lambda_2\mathcal{L}_{SS} + \lambda_3\mathcal{L}_W$, where λ_1, λ_2, and λ_3 are scalars to balance the loss terms, and they are empirically set as 0.1, 1, and 0.1, respectively.

3. Experiments

Datasets. To develop and evaluate rASD models, we utilize the AVA-ActiveSpeaker dataset [13] for ASD and create a

α	V	MAAS			ASC			SPELL			TalkNet			EASEE		
		Orig	NT	Ours	Orig	NT	Ours	Orig	NT	Ours	Orig	NT	Ours	Orig	NT	Ours
0	80.81	84.75	84.11	**85.36**	86.16	85.30	**86.31**	90.78	91.07	**91.37**	92.31	92.25	**92.85**	90.26	89.81	**90.55**
0.2		77.86	79.81	**82.56**	80.32	82.33	**84.38**	84.67	88.34	**89.71**	89.33	89.98	**91.37**	85.99	87.03	**89.20**
0.4		73.28	77.23	**80.59**	76.15	80.24	**82.79**	80.49	86.54	**88.30**	87.50	88.51	**90.30**	83.26	84.65	**87.71**
0.6		69.97	75.38	**79.05**	73.06	78.63	**81.54**	77.40	85.22	**87.19**	86.29	87.5	**89.39**	80.15	83.95	**86.46**
0.8		67.44	73.95	**77.44**	70.65	77.33	**80.47**	75.01	84.21	**86.26**	85.37	86.67	**88.63**	79.44	83.01	**85.40**
1		65.42	72.71	**76.56**	68.73	76.21	**79.51**	73.09	83.35	**85.39**	84.64	85.99	**87.97**	78.69	82.34	**84.98**

Table 1: Results of rASD. Here, V is a visual-only model from TalkNet that is not affected by audio noises. Our proposed framework can be applied to each of the five different ASD approaches, effectively enhancing their robustness.

Figure 3: Real-world visual examples, both of which are from actual video recordings. The first example has strong cafeteria noise, while the second contains background music sounds. Our approach, TalkNet+Ours, can be applied to handle these real-world examples (second row). In contrast, TalkNet with noisy training failed (first row).

Real-world Noise Audio (RNA) dataset from AudioSet [9] to simulate real-world noise. The AVA-ActiveSpeaker dataset contains 262 Hollywood movie videos split into 120 training, 33 validation, and 109 testing videos. For our RNA dataset, we extract 50 clips each from 27 classes of AudioSet that depict diverse real-world sounds, like baby cries, dog barks, and guitar strums, resulting in 1064 training, 133 validation, and 133 testing clips. During training, for each AVA-ActiveSpeaker audio sample, we randomly add a noise segment from RNA's training set. This noisy audio is produced as $A^t_{ij} = A^t_{\text{speech}_i} + \alpha_{ij} A^t_{n_{ij}}$, with α_{ij} being a uniformly sampled noise factor from $[0, 1]$. This factor dictates the noise intensity, aiding neural networks in understanding speech amidst varying noise levels without noise overfitting. For testing, speech from the AVA-ActiveSpeaker validation set pairs with random noise from RNA's validation set. We generate six noisy speech variants by adjusting the noise factor, α^v, from a set: $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Baselines. We select five recent state-of-the-art ASD methods, ASC [2], MAAS [10], TalkNet [14], SPELL [11], and EASEE [4] as our comparison methods. For each, we first train it without adding any noises and test the model on AVA ActiveSpeaker with noise from the RNA. The performance is displayed in the 'Orig' column of Tab. 1. Next, we incorporate RNA noise into the training data and retrain. These results are in the 'NT' column. Lastly, we integrate our robust audio feature generator into each method and train with the noisy data. Performance for this is in the 'Ours' column.

Evaluation. Following previous works on ASD [13, 14], we use mean average precision (mAP) as the metric and report our results on the AVA-ActiveSpeaker validation set.

3.1. Results and Analysis

Audio Noise Matters. As seen in Table 1, the performance of all five ASD approaches drops as the level of audio noise increases. When $\alpha = 1$, the performance of the original MAAS, ASC, SPELL, TalkNet, and EASEE models decreased by 19.3%, 16.6%, 17.4%, 17.7%, 7.7%, and 11.6%, respectively. Surprisingly, ASC and SPELL, as strong multimodal models, can achieve even worse performance than the visual-only unimodal baseline when α is large. These results demonstrate that speech noises can significantly weaken ASD performance, and joint audio-visual modeling may not always be helpful in a noisy environment.

Noisy Training. The results show that NT models can generally improve ASD results once the input speech sound becomes noisy. Thus, by adding randomly sampled audio noises into the training data, model robustness can be improved. Another observation is that NT models can decrease performance for most approaches when we do not add any external noises during testing. One possible reason for this is that NT models may overfit to some audio noises.

Our rASD framework is General and Effective. Table 1 shows that our robust ASD framework can significantly improve the detection performance of different ASD approaches. Therefore, learning robust audio features is crucial for detecting active speakers in noisy environments.

Real-world Data. To further validate the effectiveness of our framework, we apply our model to real-world scenes (see Fig. 3). We can find that our model successfully detects talking faces in videos with cafeteria noises and background music, while the baseline model fails. These results further illustrate the effectiveness and generalization capability of our robust ASD framework.

Acknowledgments. This work was supported in part by gifts from Cisco Systems and Adobe. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018.

[2] 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.

[3] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021.

[4] Juan León Alcázar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022.

[5] Sourish Chaudhuri, Joseph Roth, Dan Ellis, Andrew C. Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Christopher Reale, Loretta Guarino Reid, Kevin Wilson, and Zhonghua Xi. Ava-speech: A densely labeled dataset of speech activity in movies. In *Proceedings of Interspeech, 2018*, 2018.

[6] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[7] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019.

[8] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021.

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[10] Juan León-Alcázar, Fabian Caba Heilbron, Ali Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. *arXiv preprint arXiv:2101.03682*, 2021.

[11] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 371–387. Springer, 2022.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[13] 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.

[14] 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.

[15] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.