ICTCAS-UCAS-TAL Submission to the AVA-ActiveSpeaker Task at ActivityNet Challenge 2021

Yuanhang Zhang*[1,2], Susan Liang*[1,2], Shuang Yang[1,2], Xiao Liu[3], Zhongqin Wu[3], Shiguang Shan[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China
[3]Tomorrow Advancing Life (TAL) Education Group, Beijing, China
{zhangyuanhang15, liangsusan18}@mails.ucas.ac.cn, {shuang.yang, sgshan}@ict.ac.cn, ender.liux@gmail.com, 30388514@qq.com

Abstract

This report presents a brief description of our method for the AVA Active Speaker Detection (ASD) task at ActivityNet Challenge 2021. Our solution, the Extended Unified Context Network (Extended UniCon) is based on a novel Unified Context Network (UniCon) designed for robust ASD, which combines multiple types of contextual information to optimize all candidates jointly. We propose a few changes to the original UniCon in terms of audio features, temporal modeling architecture, and loss function design. Together, our best model ensemble sets a new state-of-the-art at 93.4% *mAP on the AVA-ActiveSpeaker test set* ***without any form of pretraining****, and currently ranks first on the ActivityNet challenge leaderboard.*

1. Introduction

Active Speaker Detection (ASD) is the task of identifying which visible person is speaking in a video, which requires careful analysis of face motion and voices. It has a variety of modern practical applications, and has gained increased popularity in the audio-visual community. Although many effective methods have been proposed and verified [1, 2, 12] (especially on the large-scale AVA-ActiveSpeaker dataset [9]), they do not sufficiently consider the relationships among the visible candidates, which heavily limits their performance in challenging scenarios with low-resolution faces, multiple candidates, etc.

Our submission is based on a novel *Unified Context Network (UniCon)* [11], which leverages multiple sources of *contextual* information to analyze all speaker candidates simultaneously. As part of our final solution, we extend the original UniCon formulation in the following ways:

- We replace 13-dimensional MFCCs with 80-dimensional log-Mel spectrograms, and apply SpecAugment during training to increase model robustness against noise.
- We replace the original Bi-GRU-based *temporal context* module with a Conformer-based counterpart, and sample longer training examples to enhance long-term temporal modeling.
- We adjust the loss function to alleviate the previously observed over-fitting problem with the audio branch.
- We ensemble predictions from both temporal convolution and Conformer back-ends to generate more reliable results.

With the above proposed changes, our final submission achieves state-of-the-art performance on the AVA-ActiveSpeaker test set with 93.4% mAP. We now describe our approach in detail.

2. Proposed Approach

To keep this report concise, we only provide a short review of the UniCon framework, and describe the changes we made leading to the new Extended UniCon.

2.1. UniCon Review

Fig. 1 provides an overview of UniCon. First, for each candidate, the scale and position of all candidates' faces are introduced as global *spatial* context to complement facial information and help learn the relationships among the speakers. Each candidate is then contrasted with others

*Equal contribution.

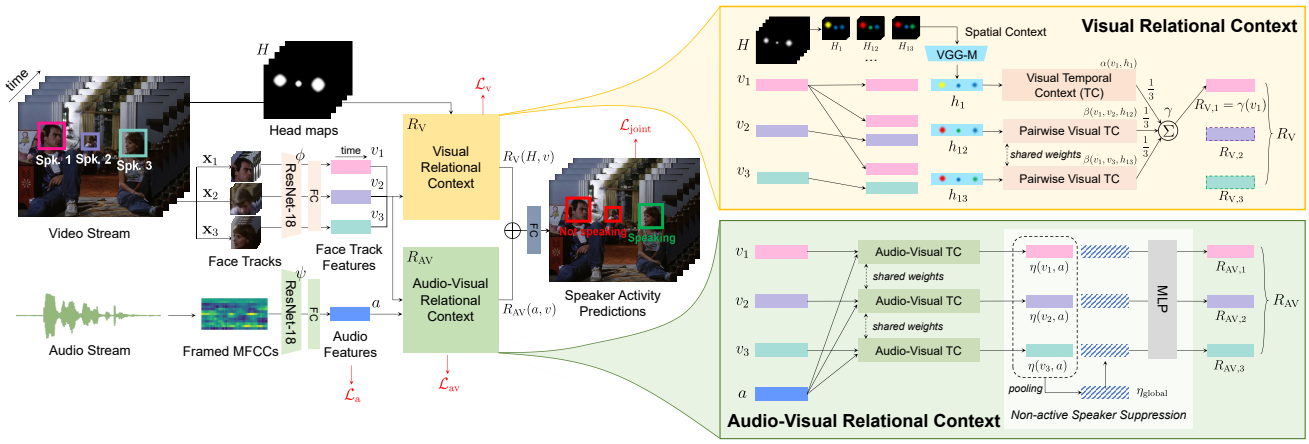

Figure 1. The original UniCon architecture [11].

from both a visual and audio-visual perspective in the *relational* context modeling component. To further improve the robustness of the model's predictions, *temporal* context is integrated. Finally, based on the aggregated contextual features, speaker activity predictions for all candidates are generated simultaneously using a shared prediction layer.

Encoders: Given an input video clip, audio tracks and face tracks are first extracted for each candidate speaker. The resulting audio and visual frames are transformed into 512-d average-pooled features with 2D ResNet-18s [5, 1], whose dimensions are reduced to 128 with fully-connected layers. For the visual stream, every five consecutive face crops are stacked together to encode short-term dynamics.

Spatial Context: To represent the relative visual saliency of the candidates in the scene and reflect gaze-related information, face positions and sizes of all candidates are encoded using 64×64 coordinate-normalized maps of 2D Gaussians, which is motivated by [7]. Color-coded versions of the above head maps are then generated for each candidate i and candidate pair (i, j) with $i \neq j$, and embedded into 64-dimensional vectors with a VGG-inspired convolutional neural network [7]. The resulting embeddings are termed each candidate's **spatial context**.

Relational Context: The **relational context** component completes two natural sub-tasks for ASD: learning a contextual visual representation for *visual voice activity detection*, and a contextual audio-visual representation for *audio-visual affinity modeling*. For *visual* relational context (R_{V}), a permutation-equivariant layer aggregates each speaker's locally perceived activity and his/her pairwise interactions with other candidates. For *audio-visual* relational context (R_{AV}), element-wise max-pooling is applied over all candidates' initial A-V affinity features to obtain a global representation. A shared multi-layer perceptron (MLP) contrasts each candidate's initial features with the global information and refines the initial features, suppressing the non-active speakers. Finally, the resulting representations R_{V} and R_{AV} are concatenated and passed through a fully-connected layer, yielding the final prediction.

Temporal Context: The **temporal context** components (α, β, and η in Fig. 1), instantiated as Bidirectional Gate Recurrent Units (Bi-GRUs), are embedded in R_{V} and R_{AV}. They improve the temporal consistency of the relational context modeling process, and smooth out local, instantaneous noises. In addition, they alleviate synchronization errors between the audio and the video which are common with in-the-wild videos and old films.

Loss function: The model is trained end-to-end with a multi-task loss formulation, with two auxiliary losses associated with auxiliary prediction layers for audio and visual-based ground truths (see Fig. 1), and a joint loss for the final speaker activity prediction layer:

$$\mathcal{L} = \mathcal{L}_{\mathrm{a}} + \mathcal{L}_{\mathrm{v}} + \mathcal{L}_{\mathrm{joint}}. \tag{1}$$

Here, each loss term applies the standard binary cross-entropy (BCE) loss, averaged over all time steps.

2.2. Extended UniCon

We now describe in detail the changes we made to the original UniCon model.

Audio features: At each time step, instead of 13-dimensional Mel Frequency Cepstrum Coefficients (MFCCs), we calculate a 80-dimensional log-Mel spectrogram from a 400ms window preceding the video frame to obtain the audio representation. The spectrograms, which contain richer acoustic information compared to MFCCs, are then processed with a ResNet-18 encoder and passed

through a fully-connected layer, same as UniCon. During training, we apply SpecAugment [8] (frequency and time masking only) to increase our model's robustness to noise and different acoustic environments.

Temporal context: Instead of using a single Bi-GRU layer, we adopt the recently proposed Convolution-augmented Transformer (Conformer) [3] architecture, which combines local convolutions with global self-attention for higher temporal modeling capacity. We use 3 Conformer layers for α, β, and η. Among them, we set the number of attention heads in the multi-head self-attention (MHSA) module to 4 for α, and 8 for β and η. The kernel size of the convolutions is set to 27.

We also increase the lengths of the randomly sampled examples during training. The original version of UniCon samples 28 frames at 25fps (1.12s, around the mean speech segment duration), while here we increase the number twofold to 56 frames, to better capture speaker alternation patterns and long-term dependencies. Empirically, we find that using longer training samples benefits Conformer-based models.

Loss function: The original UniCon formulation applies an auxiliary binary classification loss on the low-level 128-d audio features ($\mathcal{L}_{\text{a}}$ in Fig. 1), which is dropped when training on multiple candidates (*i.e.* when relation context is deployed), as it leads to serious over-fitting within the audio modality. We find that removing this auxiliary audio prediction layer and adding a new auxiliary active speaker prediction layer after the contextual audio-visual features (R_{AV}) resolves the over-fitting problem ($\mathcal{L}_{\text{av}}$ in Fig. 1), and leads to a unified loss formulation for both one candidate and multiple candidates:

$$\mathcal{L} = \mathcal{L}_{\text{av}} + \mathcal{L}_{\text{v}} + \mathcal{L}_{\text{joint}}. \quad (2)$$

2.3. Ensembling

To further boost performance, we apply model ensembling by combining the predictions from a temporal convolution (TCN) back-end and a Conformer back-end. Our intuition is that the self-attention based Conformer model is better at modeling long-term dependencies, while the TCN-based model has a local receptive field and is more sensitive to short utterances. Here, we average Wiener-smoothed (over 11-frame, or 0.44s windows) TCN logits and the raw Conformer logits, and pass the mean logits through the sigmoid function to generate the final predictions.

2.4. Implementation Details

We carry out our experiments on the large-scale AVA-ActiveSpeaker dataset [9] which consists of 262 YouTube movies from film industries around the world. Our data preprocessing scheme is identical to that described in [11]. We implement our model with PyTorch and the `pytorch-lightning` package. All models are trained from scratch, using the AdamW optimizer [6] and automatic mixed precision (AMP) on a single NVIDIA Titan RTX GPU with 24GB memory. The network parameters are initialized using He initialization [4]. We use a `fairseq`-style Transformer learning rate schedule [10], warming up linearly to a maximum learning rate of 0.0003 (when training on one candidate) or 0.0001 (when training on multiple candidates) over $2,000$ updates, and decaying proportionally to the inverse square root of the step number thereafter.

During training, we augment the data via random horizontal flipping and uniform corner cropping along the input face tracks, followed by random adjustments to brightness, contrast, and saturation. All cropped face tracks are resized to 144×144, and randomly cropped to 128×128 for training [1]. We use a central patch for testing.

3. Results

The official metric for the task is Mean Average Precision (mAP). We obtain the numbers using the official evaluation tool, after interpolating our predictions to the timestamps in the original annotations.

Table 1. The performance of previous state-of-the-art, the original UniCon model, and our extension on the AVA-ActiveSpeaker validation and test sets.

Method	**Val mAP (%)**	**Test mAP (%)**
UCAS [12]	84.0	83.5
ASC [1]	87.1	86.7
Naver Corporation [2]	87.8	87.8
UniCon [11]	92.0	90.7
Ours (Conformer)	93.6	93.3
Ours (Conformer + TCN)	**93.8**	**93.4**

As shown in Table 1, our extensions give our final solution a competitive edge over the original UniCon model, improving it by 1.6% on validation set and 2.7% on test set in terms of mAP. The new temporal context module and stronger acoustic features are crucial to the improvement. Remarkably, our best single model submission improves over the previous state-of-the-art [2] by as much as 5.5% *without any pre-training or ensembling*. Finally, our best model ensemble further yields small improvements (0.2% on the validation set, and 0.1% on the test set).

An interesting observation is that the Conformer-based models achieve very similar performance on the validation and the test set, while the original Bi-GRU-based UniCon performs slightly worse on the test set. One possibility is that the Conformer back-end has a higher capacity, allowing it to generalize better on the more diverse test set.

References

[1] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *CVPR*, pages 12462–12471. IEEE, 2020. 1, 2, 3

[2] Joon Son Chung. Naver at ActivityNet Challenge 2019 - Task B active speaker detection (AVA). *CoRR*, abs/1906.10555, 2019. 1, 3

[3] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for speech recognition. In *INTERSPEECH*, pages 5036–5040. ISCA, 2020. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *ICCV*, pages 1026–1034. IEEE Computer Society, 2015. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 2

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019. 3

[7] Manuel J. Marín-Jiménez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. LAEO-Net: Revisiting people looking at each other in videos. In *CVPR*, pages 3477–3485. Computer Vision Foundation / IEEE, 2019. 2

[8] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, pages 2613–2617. ISCA, 2019. 3

[9] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew C. Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. AVA-ActiveSpeaker: An audio-visual dataset for active speaker detection. In *ICASSP*, pages 4492–4496. IEEE, 2020. 1, 3

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3

[11] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen. UniCon: Unified context network for robust active speaker detetion. *to appear in arXiv*, 2021. 1, 2, 3

[12] Yuan-Hang Zhang, Jingyun Xiao, Shuang Yang, and Shiguang Shan. Multi-task learning for audio-visual active speaker detection. *The ActivityNet Large-Scale Activity Recognition Challenge 2019*, pages 1–4, 2019. 1, 3