TALKNCE: IMPROVING ACTIVE SPEAKER DETECTION WITH TALK-AWARE CONTRASTIVE LEARNING

Chaeyoung Jung[1*], *Suyeon Lee*[1*], *Kihyun Nam*[1], *Kyeongha Rho*[1],
You Jin Kim[2], *Youngjoon Jang*[1], *Joon Son Chung*[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Naver Cloud Corporation, South Korea

ABSTRACT

The goal of this work is Active Speaker Detection (ASD), a task to determine whether a person is speaking or not in a series of video frames. Previous works have dealt with the task by exploring network architectures while learning effective representations has been less explored. In this work, we propose TalkNCE, a novel talk-aware contrastive loss. The loss is only applied to part of the full segments where a person on the screen is actually speaking. This encourages the model to learn effective representations through the natural correspondence of speech and facial movements. Our loss can be jointly optimized with the existing objectives for training ASD models without the need for additional supervision or training data. The experiments demonstrate that our loss can be easily integrated into the existing ASD frameworks, improving their performance. Our method achieves state-of-the-art performances on AVA-ActiveSpeaker and ASW datasets.

Index Terms— Active Speaker Detection, Multi-Modal Speech Processing, InfoNCE loss

1. INTRODUCTION

In recent years, there has been a shift in the way we communicate, transitioning from in-person exchanges to audio-visual interactions online. With this paradigm shift, identifying the active speaker has become crucial to enable effective communication and understanding of conversations in context. In multi-modal conversations, active speaker detection (ASD) serves as a fundamental pre-processing module for speech-related tasks, including audio-visual speech recognition [5], speech separation [6, 7], and speaker diarization [8, 9].

ASD in real-world scenarios is a challenging task that requires effective integration of audio-visual information and leveraging their long-term relationships. In response to these specific requirements, ASD model architectures are designed to create corresponding audio-visual features and comprehensively analyze these features over long periods to capture crucial temporal context. The general ASD frameworks begin with modality-specific encoders extracting embedding and subsequent audio-visual fusion techniques enable the seamless interaction of heterogeneous modalities. A prevalent fusion method involves the utilization of a cross-attention mechanism [2, 4, 10], facilitating the connection between visual streams and synchronized audio streams. After the two features are combined, self-attention [2, 4] or RNN-based architectures [3, 11] are employed to ensure consistent tracking of the active speaker throughout the utterance. These aforementioned approaches take advantage of the complementary information of both modalities. However, there has been less exploration into learning high-quality representations for the multi-modal task of ASD.

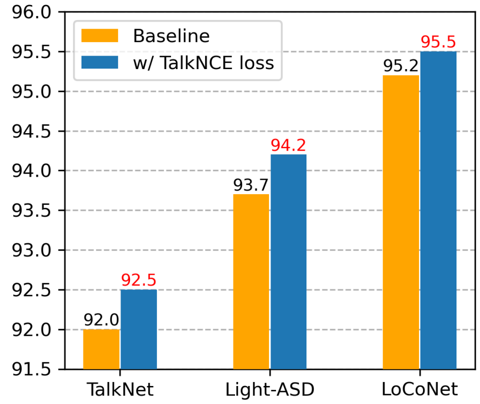

Fig. 1: Comparison of performances on the validation set of AVA-ActiveSpeaker [1]. Our TalkNCE loss improves the performance of existing ASD models [2, 3, 4] consistently.

In this paper, we focus on learning strong phonetic representations for the ASD task. To this end, we propose a novel supervised talk-aware contrastive learning strategy by devising a new loss function, named TalkNCE loss. The loss acts on audio and visual features in a frame-wise manner rather than a chunk-wise manner [12, 13] to learn transient phonetic representations required for audio-visual matching. Furthermore, the suggested loss is computed exclusively on the ac-

[*] These authors contributed equally.

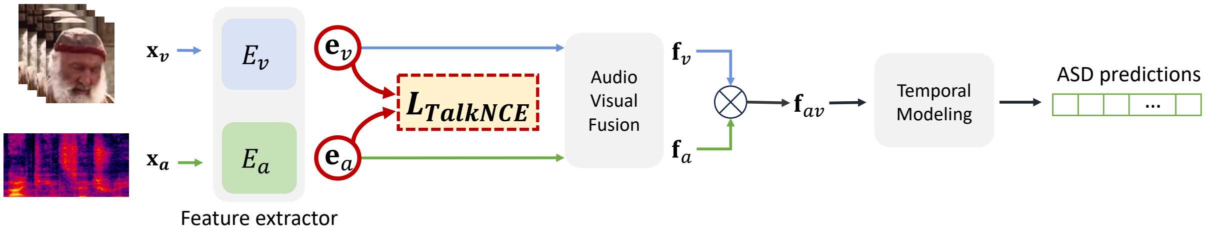

Fig. 2: General framework for the active speaker detection task. Our loss is applied to the audio and visual embedding before the fusion. $\otimes$ denotes the concatenation of two features along the temporal dimension.

tive speaking section extracted by using the ASD labels. Supervision imposed by the TalkNCE loss enables the encoders to concentrate on the fine details of audio-visual correspondence, thus enhancing the effect of audio-visual fusion. In contrast to the approaches that use pre-trained audio-visual embeddings [11, 14], our method trains the models in an end-to-end manner by combining the proposed TalkNCE loss with the existing ASD classification losses. Our contrastive learning strategy is model-agnostic, on many of which the addition of the TalkNCE loss brings performance improvements as shown in Fig. 1. In particular, combined with [4], our method outperforms the previous state-of-the-art ASD method on the AVA-ActiveSpeaker and ASW datasets.

Our contributions can be summarized as follows: (1) We propose TalkNCE loss, a novel contrastive loss for ASD that enforces the model to exploit information from the alignment of audio and visual streams. (2) The proposed loss can be plugged into multiple existing ASD frameworks and the additional supervision imposed by the loss improves the performances without any additional data. (3) Our method achieves state-of-the-art performance on both the AVA-ActiveSpeaker and ASW datasets.

2. METHOD

2.1. Preliminaries

This section describes the general ASD framework widely used in literature [1, 2, 3, 4, 15]. As shown in Fig. 2, ASD frameworks usually consist of feature extractors, audio-visual fusion, and temporal modeling. Given the input video frames, every face is detected, cropped, stacked, and transformed to grayscale images $\mathbf{x}_v$. The corresponding audio waveform is transformed into mel-spectrogram $\mathbf{x}_a$ by the short-time Fourier transform. Then the audio encoder $E_a(\cdot)$ and visual encoder $E_v(\cdot)$ encode inputs into audio and visual embeddings respectively,

$$\mathbf{e}_a = E_a(\mathbf{x}_a), \quad \mathbf{e}_v = E_v(\mathbf{x}_v) \tag{1}$$

To integrate the information between two modalities, the two features are fed into an audio-visual fusion module. Recent methods such as TalkNet [2] and LoCoNet [4] utilize cross-attention layers for multi-modal fusion. By cross-attention mechanism, audio embedding is effectively aligned with the visual information of the corresponding speaker. Then the attention-weighted features $\mathbf{f}_a$ and $\mathbf{f}_v$ are concatenated to make the fused audio-visual features $\mathbf{f}_{av}$.

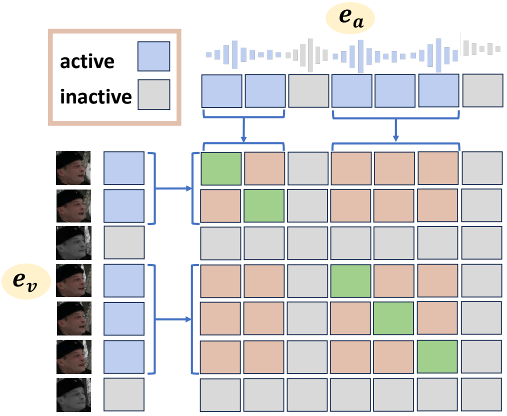

Fig. 3: Our TalkNCE loss is applied only to active speaking sections (blue) in a frame-wise manner. Audio and visual embeddings of synchronized video frames are positive pairs (green), while others are negative pairs (orange).

Finally, several works [2, 4, 3, 16, 17] employ a temporal modeling module by utilizing longer context to accurately predict the active speaker. Including self-attention modules in LoCoNet [4], a Gated Recurrent Unit (GRU), a Long Short-Term Memory (LSTM), and a Bidirectional LSTM (BiLSTM) are frequently exploited for temporal modeling.

2.2. Contrastive Learning with TalkNCE Loss

The training objective typically used in the ASD task [1, 2, 3, 4, 15] can be represented as:

$$\mathcal{L}_{model} = \mathcal{L}_{av} + \lambda_a \cdot \mathcal{L}_a + \lambda_v \cdot \mathcal{L}_v, \tag{2}$$

where $\mathcal{L}_{av}$ denotes the cross-entropy loss between the ground-truth labels and the final frame-level predictions after temporal modeling. $\mathcal{L}_a$ and $\mathcal{L}_v$ are auxiliary cross-entropy losses that are computed with uni-modal features $\mathbf{f}_a$ and $\mathbf{f}_v$ to make the model utilize both modalities in a balanced manner.

In addition to the cross-entropy losses, we introduce TalkNCE loss, a talk-aware contrastive loss that provides supervision for encoders to learn audio-visual correspondence. The proposed loss brings audio-visual embedding pairs closer

Method	Multiple candidates?	E2E?	mAP(%)
ASC* [15]	✓	✗	87.1
MAAS* [18]	✓	✗	88.8
TalkNet* [2]	✗	✓	92.3
ASDNet* [19]	✓	✗	93.5
EASEE-50* [16]	✓	✓	94.1
SPELL* [17]	✓	✗	94.2
Light-ASD* [3]	✗	✓	94.1
TS-TalkNet [10]	✗	✗	93.9
LoCoNet [4]	✓	✓	95.2
Ours	✓	✓	**95.5**

Table 1: Comparison of ASD performance on the AVA-ActiveSpeaker validation set. *Performance of previous methods are from [3]. E2E refers to end-to-end training.

if they come from the same video frame or pushes each other away otherwise. This encourages the encoders to detect salient information regarding audio-visual synchronization and to concentrate on phonetically fine details including frame-level correlation between audio and visual modalities.

Our new loss optimizes frame-level matching between paired audio and visual streams as denoted in Fig. 3. Given the audio and visual embeddings $\mathbf{e}_a$ and $\mathbf{e}_v$, the active speaking regions are extracted using the original ASD labels. Let T_{act} be the length of active speaking region and $\{\mathbf{e}_{a,i}\}_{i\in\{1,\ldots,T_{act}\}}$, $\{\mathbf{e}_{v,i}\}_{i\in\{1,\ldots,T_{act}\}}$ be the frame-level audio and visual embeddings for the active speaking region. Then the TalkNCE loss $\mathcal{L}_{TalkNCE}$ is as follows:

$$\mathcal{L}_{TalkNCE} = -\frac{1}{T_{act}}\sum_{i=1}^{T_{act}} \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{T_{act}} \mathbb{1}_{[j\neq i]} \exp(s_{i,j}/\tau)}, \qquad (3)$$

where $s_{i,j} = \|\mathbf{e}_{v,i}\|^T \|\mathbf{e}_{a,j}\|$ and τ represents a temperature constant fixed to 1. With our proposed TalkNCE loss and the model's loss, the final training objective used in this paper can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{model} + \lambda \cdot \mathcal{L}_{TalkNCE}, \qquad (4)$$

where λ is the weight value of TalkNCE loss.

3. EXPERIMENTS

3.1. Datasets

AVA-ActiveSpeaker dataset [1] is a benchmark for evaluating active speaker detection performance extracted from Hollywood movies. It contains 262 videos that span 38.5 hours, of which 120, 33, and 109 videos are divided into training, validation, and test set respectively. As the videos are from movies, the dataset includes some dubbed videos.

Active Speakers in the Wild (ASW) dataset [11] is a dataset derived from the existing audio-only speaker diariazation dataset VoxConverse [9] by using a subset of the videos and exploiting the diarization annotations. Unlike AVA-ActiveSpeaker, ASW excludes dubbed videos to ensure synchronization between lip movements and speech. The total duration of videos in the dataset is 30.9 hours, and the ratio of an active track duration is 56.7%, 60.4%, and 57.0% for the training, validation, and test set respectively. Note that four videos in the dataset cannot be downloaded at present because they are no longer available online.

Method	mAP(%)
TalkNet†	92.0
TalkNet+$\mathcal{L}_{TalkNCE}$	92.5
LightASD†	93.9
LightASD+$\mathcal{L}_{TalkNCE}$	94.2
LoCoNet†	95.2
LoCoNet+$\mathcal{L}_{TalkNCE}$	**95.5**

Table 2: Performance comparison of other baseline models trained with TalkNCE Loss. † Results are reproduced using the codes released by the original works.

3.2. Experimental Setup

Implementation details. Visual input $\mathbf{x}_v \in \mathbb{R}^{n_s \times T \times H \times W}$ is pre-processed to size of $H = 112$ and $W = 112$ for n_s speakers and T video frames. The corresponding audio signal is transformed into mel-spectrogram $X_a \in \mathbb{R}^{4T * C_{mel}}$. The mel-spectrogram is extracted to have $4T$ time frames, by adjusting the hop length of short-time Fourier transform. The frame-level audio embedding $e_a \in \mathbb{R}^{T\times C}$ and visual embedding $e_v \in \mathbb{R}^{T\times C}$ are extracted with a feature size of $C = 128$.

We reproduce the baselines for performance comparison following the official reporting of each model [4, 3, 2]. We train each model end-to-end for 25 epochs using the Adam optimizer [20]. Feature dimensions of both audio and visual embeddings are set to 128 for all three models. The value for λ is set to 0.3 for LoCoNet to make the scale similar with $\mathcal{L}_{av}$. Similarly, λ values for other models [2, 3] are assigned considering the magnitudes of the existing loss function.

Evaluation metrics. Following the common protocol suggested by previous works [4, 3, 2, 19], mean Average Precision (mAP) is used as an evaluation metric for the AVA-ACtiveSpeaker validation set. Evaluation metrics reported for the ASW validation and test sets are mAP, Area Under the ROC Curve (AUC), and Equal Error Rate (EER).

3.3. Performance on AVA-ActiveSpeaker Dataset

The performance of our method is compared with that of existing ASD methods on the AVA-ActiveSpeaker validation set, and we show our results in Table 1. As indicated in Table 1, LoCoNet trained with our method attains 95.5% mAP, achieving a state-of-the-art result. Note that the model can be

	Method	mAP(%)↑	AUC(%)↑	EER(%)↓
Val	TalkNet [2]	96.4	98.2	6.0
	TS-TalkNet [10]	97.7	98.7	5.1
	Ours	**98.8**	**99.3**	**3.2**
Test	TalkNet [2]	97.7	98.6	5.1
	TS-TalkNet [10]	98.5	99.0	4.3
	ASW-BGRUs [11]	96.6	97.2	6.2
	LoCoNet [4]	93.4	95.1	9.8
	Ours	**99.3**	**99.5**	**3.0**

Table 3: Comparison of ASD performance on the ASW dataset. Performance of the previous methods are from [10]. ↑ denotes higher is better, and ↓ denotes lower is better.

trained in an end-to-end manner using the combination of the proposed contrastive loss with the cross-entropy loss, without the need for a multiple-stage training strategy. We also apply our contrastive learning strategy for TalkNet and Light-ASD to verify the general capability of our method. It can be seen in Table 2 that our loss provides a consistent increase in performances compared to our reproduced models [2, 3].

3.4. Performance on ASW Dataset

We also compare our method with other ASD methods on the ASW dataset. In Table 3, our method achieves state-of-the-art results by a margin of 1.1% on the ASW validation set and 0.8% for the test set. Applying our talk-aware contrastive loss $\mathcal{L}_{TalkNCE}$ to LoCoNet shows the highest performance. Unlike the AVA-ActiveSpeaker dataset, the ASW dataset does not contain any dubbed video. As a result, our loss is more effective for the ASW dataset. This is because our loss is devised to optimize performance by learning audio-visual phonetic representations that synchronize each other.

3.5. Ablation Studies

Choice of positive and negative samples. By using the ASD labels for a certain speaker, a sequence of audio and visual embeddings can be divided into 'active' and 'inactive' regions, as shown in Fig. 3. Our proposed method only attends to the 'active' regions for calculating the contrastive loss $\mathcal{L}_{TalkNCE}$, and Table 4 shows the results obtained by different sampling strategies. The experimental result demonstrates that the model can learn richer meaningful representations from the active part of the inputs, rather than from the inactive part. In particular, utilizing inactive frames of audio embedding degrades performance, as the audio in inactive regions includes sounds that are irrelevant to the speaker's lip motions. Similarly, the visual input from inactive regions contains non-speaking faces, which can have an adverse effect on discriminative learning.

Visual embedding	Audio embedding	mAP(%)
Act	Act	**95.5**
Act	Act + Inact	94.3
Act + Inact	Act	95.1
Act + Inact	Act + Inact	94.9

Table 4: Comparisons between combinations of visual and audio embeddings used for calculating TalkNCE loss. 'Act' refers to the active region, and 'Inact' refers to inactive regions for a certain speaker.

Location of $\mathcal{L}_{TalkNCE}$	mAP(%)
None (Baseline)	95.2
After CA	93.4
Before CA	**95.5**

λ	mAP(%)
0.15	95.2
0.3	**95.5**
0.6	95.2

Table 5: Ablation on the position and the λ value of the TalkNCE loss, using LoCoNet [4] as a baseline. CA denotes the cross-attention module for audio-visual fusion.

Location of $\mathcal{L}_{TalkNCE}$. Our proposed loss is particularly effective for refining the audio and visual embeddings before the audio-visual fusion stage as indicated in the right side of Table 5. TalkNCE loss function has the capability to enhance uni-modal features by taking information from other modalities through frame-level contrastive learning. However, applying our loss after the audio-visual fusion stage gets an even lower score than the baseline. This is because the two modalities have already been combined in the attention-weighted features $\mathbf{f}_a$ and $\mathbf{f}_v$, so applying the loss function after the fusion stage results in a negative impact on the multi-modal guidance provided by the features.

Weight value of $\mathcal{L}_{TalkNCE}$. We also show the performance of LoCoNet trained using our loss with varying weight value λ for $\mathcal{L}_{TalkNCE}$. As shown in the left side of Table 5, the λ value of 0.3 gives the best result. Our loss and the ASD classification loss $\mathcal{L}_{av}$ can be trained in balance at this value.

4. CONCLUSION

In this paper, we propose TalkNCE, a talk-aware contrastive loss for active speaker detection. The objective function utilizes active speaker labels to select audio and visual embeddings from which natural co-occurrences are learnt. Our loss is applicable to a range of existing ASD systems, for which we demonstrate a consistent improvement in performance. In particular, we achieve a new state-of-the-art score of 95.5% on the AVA-ActiveSpeaker validation set by combining the proposed method with the LoCoNet model.

5. ACKNOWLEDGEMENTS

We would like to thank Hee-Soo Heo, Bong-Jin Lee, and Icksang Han for their helpful comments, and Yidi Jiang for the help with the evaluation of the models on the ASW dataset.

6. REFERENCES

[1] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew 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 *Proc. ICASSP*, 2020. 1, 2, 3

[2] 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 *Proc. ACM MM*, 2021. 1, 2, 3, 4

[3] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen, "A light weight model for active speaker detection," in *Proc. CVPR*, 2023. 1, 2, 3, 4

[4] Xizi Wang, Feng Cheng, Gedas Bertasius, and David Crandall, "LoCoNet: Long-short context network for active speaker detection," *arXiv preprint arXiv:2301.08237*, 2023. 1, 2, 3, 4

[5] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018. 1

[6] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. ECCV*, 2018. 1

[7] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018. 1

[8] Joon Son Chung, Bong-Jin Lee, and Icksang Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," in *Proc. Interspeech*, 2019. 1

[9] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman, "Spot the conversation: speaker diarisation in the wild," in *Proc. Interspeech*, 2020. 1, 3

[10] Yidi Jiang, Ruijie Tao, Zexu Pan, and Haizhou Li, "Target active speaker detection with audio-visual cues," in *Proc. Interspeech*, 2023. 1, 3, 4

[11] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung, "Look Who's Talking: Active speaker detection in the wild," in *Proc. Interspeech*, 2021. 1, 2, 3, 4

[12] Joon Son Chung and Andrew Zisserman, "Out of time: automated lip sync in the wild," in *ACCV 2016 Workshops*, 2017. 1

[13] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang, "Perfect match: Self-supervised embeddings for cross-modal retrieval," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 568–576, 2020. 1

[14] 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*, vol. 4, 2019. 2

[15] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem, "Active speakers in context," in *Proc. CVPR*, 2020. 2, 3

[16] Juan León Alcázar, Moritz Cordes, Chen Zhao, and Bernard Ghanem, "End-to-end active speaker detection," in *Proc. ECCV*, 2022. 2, 3

[17] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar, "Learning long-term spatial-temporal graphs for active speaker detection," in *Proc. ECCV*, 2022. 2, 3

[18] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem, "Maas: Multi-modal assignation for active speaker detection," in *Proc. ICCV*, 2021. 3

[19] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *Proc. ICCV*, 2021. 3

[20] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015. 3