INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Look Who's Talking: Active Speaker Detection in the Wild

You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung[*]*, Yoohwan Kwon*[*]*,*
Bong-Jin Lee, Youngki Kwon, Joon Son Chung

Naver Corporation, South Korea
youjin.kim117@navercorp.com

Abstract

In this work, we present a novel audio-visual dataset for active speaker detection in the wild. A speaker is considered **active** when his or her face is visible and the voice is audible simultaneously. Although active speaker detection is a crucial pre-processing step for many audio-visual tasks, there is no existing active speaker detection dataset to evaluate the performance using natural human speech. We therefore curate the *Active Speakers in the Wild* (ASW) dataset which contains videos and co-occurring speech segments with dense speech activity labels. Videos and timestamps of audible segments are parsed and adopted from VoxConverse, an existing speaker diarisation dataset that consists of videos in the wild. Face tracks are extracted from the videos and active segments are annotated based on the timestamps of VoxConverse in a semi-automatic way. Two reference systems, one is self-supervised and the other is supervised system, are evaluated on the dataset to provide the baseline performances of ASW. Cross-domain evaluation and case study are conducted, in order to show the negative effect of the dubbed videos that are excluded in ASW.

Index Terms: active speaker detection, audio-visual dataset, multi-modal speech processing.

1. Introduction

How can an AI-enabled agent interact with the user in a public area? There are many individuals in the crowd who can be confused as the user, and the agent must recognise who is talking to them in order to address the command.

Perceiving the rich information in a conversation is a core task in human computer interaction (HCI) [1–3], speech recognition [4–8], speaker recognition [9–12] and video understanding [13–15]. It conveys identity, intents or emotions as well as the truth or knowledge. Active speaker detection (ASD) detects when a person speaks in a video, providing the crucial information on who is speaking to understand conversations in context (Figure 1).

In addition to being an interesting problem in its own right, ASD is also an important pre-processing step for multi-modal speech processing – it is a key component of pipelines to create VoxCeleb [16–18] and LRS datasets [19–21], and is essential for applications such as audio-visual speech enhancement [22–27] and audio-visual speaker diarisation [28–31].

ASD has some challenges in both audio and visual elements. Audio can contain background noises such as laugh, clap, mumbling, camera shutter sound, or overlapping speech. Visual clues can be loss due to the following reasons. A speaker can turn his/her head aside or down while talking and cover the face by his/her hand. Also, some words are pronounced by only tongue movements. Therefore, instead of utilizing audio and visual elements individually, both information need to be used at the same time.

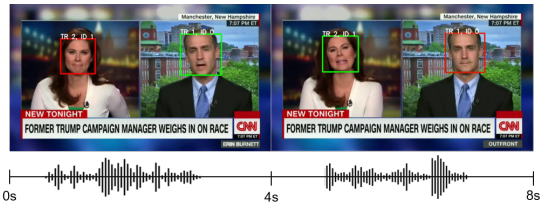

Figure 1: *Who is speaking? ASD detects when a person is speaking. The person who is active is in green boxes, and non-active is in red. Between* 0 *and* 4 *seconds, the left woman is speaking (**active**), and the right man is not speaking (**non-active**) . On the other hand, between* 4 *and* 8 *seconds, the left woman is not speaking (**non-active**), and the right man is speaking (**active**).*

AVA-ActiveSpeaker [32] is a recently released audio-visual dataset for ASD. It contains large amount of conversational videos, but is not suitable for the aforementioned tasks. A key limitation of this dataset is that a large proportion of the videos are dubbed movies, often into another language. Therefore, the audio and the video are not in correspondence or synchronisation. Without correspondence between audio and video, the network learns to detect whether there is speech in the audio and whether the lips are moving, but not whether the two correspond (*i.e.* originate from the same person). Although the sections of voice-over or dubbing are labelled as 'positive' in this dataset, in the context of audio-visual speech enhancement, speaker recognition and many related applications, they should be considered false positives. For example, a model trained on this dataset would consider a silent video of Donald Trump's interview with the voice-over of a CNN narrator as positive. These issues in the existing dataset raise the necessity for a new dataset of natural human speech.

In this paper, we present *Active Speakers in the Wild* (ASW), an audio-visual dataset for ASD that overcomes the limitations in the previous work. The dataset is created using a semi-automatic pipeline and will be released to the public [1].

The videos in this dataset are based on VoxConverse [33], an audio-only dataset for speaker diarisation, which solves the problem of "*who spoke when*". The same list of video is used to create ASW, in order to make use of the speech activity labels from the VoxConverse dataset.

In order to perform ASD using audio and video at the same time, a framework that can effectively model information common to both signals is required. The speech activity labels are combined with SyncNet [34] predictions to generate initial annotations, which are then corrected by human annotators. The

[*] work done while at Yonsei University

[1] The second round manual verification is being processed.

http://dx.doi.org/10.21437/Interspeech.2021-2041

Table 1: *Statistics of ASW. The number of videos each set has are 106, 53, and 53 respectively. The duration of the active tracks and the non-active tracks are denoted in* ***hours****, and the ratio is in the parentheses.*

Set	# of videos	Active	Non-active
Dev	106	6.5 (43.6%)	8.4 (56.4%)
Val	53	4.0 (39.0%)	6.1 (60.9%)
Test	53	3.6 (42.0%)	4.9 (58.0%)

Table 2: *Statistics of face tracks from ASW. The number of tracks of each set is summarised. The min, max and mean of the duration of the track are denoted in* ***seconds****.*

Set	# of face tracks	Min	Max	Mean
Dev	4,676	2.1	311.0	11.9
Val	3,483	2.1	327.9	10.7
Test	3,392	2.1	156.3	9.2

annotations are double-checked in order to minimise human errors.

We provide a number of baselines for this task, including a SyncNet-based ASD and a pre-trained model from the winning entry for the ASD track of the ActivityNet challenge in 2019. We report a number of metrics including average precision (AP), area under the receiver operating characteristic (AUROC), and equal error rate (EER).

2. Dataset Description

The ASW dataset consists of 212 videos randomly selected from the VoxConverse dataset. 212 videos are divided into three sets, 106, 53 and 53 for development, validation and test set. Face bounding-boxes are detected for every frame of 212 videos and grouped together. As a result, 33.7 hours are generated and annotated. The development set has 14.9 hours in which 6.5 hours (43.6%) are active tracks, and 8.4 hours (56.4%) are non-active tracks. The validation set consists of 10.1 hours where 4.0 hours (39.0%) are active, and 6.7 hours (60.9%) are non-active. The test set is comprised with 8.5 hours where 3.6 hours (42.0%) are active, and 5.0 hours (58.0%) non-active. The summary of the statistics can be found in Table 1.

The total number of the face tracks is 11551, and the number of the tracks in development, validation, and test set are 4676, 3483 and 3392, respectively. The min, max, mean of the tracks from development set are 2.1, 311.0, 11.9 seconds, validation set are 2.1, 327.9, 10.7 seconds, and test set are 2.1, 156.3, 9.2 seconds. The summary of the statistics can be found in Table 2.

Videos in the dataset are recorded in the wild and include news, political debates, press conferences, panel discussions, interviews, talk-shows and so on. Therefore, it contains various types of background noise, such as laughter, applause and camera shutter sound, as well as visual challenges such as occlusion and the deterioration of video quality as shown in Figure 2.

Annotations are binary - **active** (positive) and **non-active** (negative). An audio-visual segment in a face track is active when it is audible and visible. The segment is considered to be audible when the utterance can be transcribed in words and visible when a face detection appears in the face track. The ASW dataset will be released publicly, including the annotations.

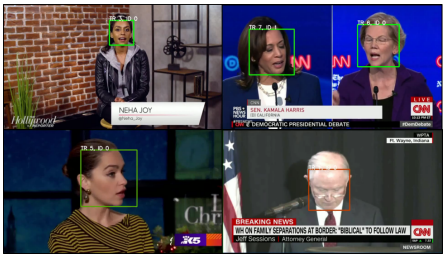

Figure 2: *Examples from ASW. The videos are recorded in the wild, including political debates, talk-shows, press conferences and so on. It leads various types of audio challenges, such as laugh, applaud, cross-talking and visual challenges, such as profile faces and changes in light. (Green and red boxes indicate active or non-active.)*

3. Dataset Creation

3.1. Automatic pipeline

The automatic part of the pipeline generates initial annotations to be checked by human annotators.

3.1.1. Face track extraction

The visual pipeline to create the dataset is based on VoxConverse. The key stages are described in the following paragraphs.

(1) Shot detection. A video is split when the scene or camera angle changes. A shot boundary detection library, PySceneDetect [35] is used, and it splits a video based on intensity or brightness changes.

(2) Face detection. The face in a clip are spot by Single Shot Scale-invariant Face Detector (S3FD) [36]. It can detect small or turned faces and achieves high performance on several benchmark datasets.

(3) Face tracking. The face detected by S3FD are concatenated, based on the position in consecutive frames. Intersection over union (IoU) defined as diving the area of overlap to the area of union between bounding-boxes is calculated between face bounding-boxes of the previous and the current frames. Pair of faces that has high IoU score are joined to form a track.

3.1.2. Active speaker annotation

We can find speech activity labels from the RTTM file of VoxConverse, since the RTTM files contain the start and the duration of all speaking segments. Speech activity labels from the VoxConverse labels must be positive for any speaking segment, but there can be no corresponding speaking segment for some speech activity regions if the face of the speaker is not visible.

SyncNet has been proposed in [34] and has shown good performance in ASD. SyncNet has originally been trained for audio-to-video synchronisation, the task of aligning the two modalities by locating the most relevant audio segment given a short video stream. For that purpose, this system learns audio and video signals as vectors in the joint embedding space, sharing the common linguistic information across two modalities. It predicts whether or not the visible speaker is speaking based on the correlation between the audio and the video embeddings. The correlation is high when a speaker is visible in the audible segment and low in other cases. The segments of active speech are detected when the correspondence is larger than a pre-defined

threshold, and the speech activity labels from the VoxConverse dataset is positive.

3.2. Manual verification

To minimise errors in the dataset, 6 human annotators have gone over the dataset to correct the errors from the automatic annotation and cross-check others' corrections. In all processes, the multimedia variation of VGG image annotator (VIA) [37] is used.

3.2.1. Guidelines and tool

The result of the automatic detection is verified by human annotators. The guideline for the audio stream is adopted from VoxConverse [33], while for the visual stream is newly defined. The speech is considered to be audible if it can be written down in words. The video is defined as visible when a speaker appears, even if the speaker's mouth is barely visible. A speaker sometimes covers his/her lips with a hand, or turns his/her face side ways. It is still defined as visible, as long as the speaker's head appears in the video. A segment is labeled as **active**, when it is audible and visible at the same time.

Original videos are provided, on which face bounding-boxes and prediction of SyncNet-ASD are overlaid. Annotators are asked to annotate the true boundary of an active segment to 0.1-second accuracy and divide a segment into two parts when there is a pause longer than 0.25 second.

3.2.2. Cross-checking

The human corrections are cross-checked by the other annotators. Disagreements between the model and the annotators usually occur when the segment is particularly challenging. The second annotator is asked to re-check the videos, with particular attention to the sections modified by the first annotator.

4. Active Speaker Detection

In this section, we describe two baseline methods for ASD, one using a cross-modal representation model trained by self-supervision and the other adopting a two-way classifier trained on an external dataset.

4.1. Self-supervised method

In [34, 38], the authors propose SyncNet to learn audio-visual representation using cross-modal self-supervision. The network configurations are shown in Figure 3. The audio and visual streams are trained in a multi-way matching method [38, 39] for the audio-to-video synchronisation task, where we use 40 candidates; 1 positive pair and 39 negative pairs. The networks are trained to minimise the distance between positive embeddings and maximise the distance between negative embeddings simultaneously, using a multi-way matching (MWM) loss. Its training criterion is as follows,

$$L_{MWM} = -\frac{1}{2N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_{n,m}\log(p_{n,m})$$
$$p_{n,m} = \frac{\exp(d_{n,m}^{-1})}{\sum_{k=1}^{M}\exp(d_{n,k}^{-1})}, \quad (1)$$

where $d_{n,m}$ is the pairwise distance between audio and visual embeddings, and $y_{n,m}$ is a similarity metric where 1 and 0 indicate positive and negative pairs, respectively. N is the number of samples, and M is the number of candidates. In face tracks, active segments can be detected where the speech and the face of a person are synchronised, showing high similarity. We adopt either the inverse of L2 distance or cosine similarity to calculate the similarity score.

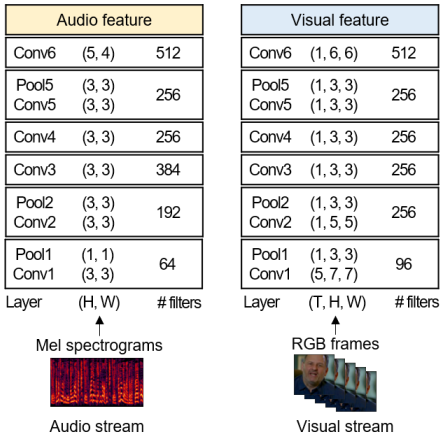

Figure 3: *Architecture of the self-supervised method. It consists of audio and visual streams, trained using 0.2 second inputs. 40-dimensional mel-spectrogram is used as the input of an audio stream and RGB frames are used as the input of an visual stream.*

4.2. Supervised method

In [40], it uses SyncNet as a front-end feature extractor, and places back-end classifier to provide explicit labels. Back-end classifier can be found in Figure 4. It consists of 2 bi-directional gated recurrent units (BGRUs) [41] and 2 fully-connected (FC) layers. GRUs are adopted instead of long short-term memory (LSTM) [42] used in [40] for efficiency. The joint representation is derived by concatenating GRU outputs along the channel axis. It is then projected into a two node output layer for binary classification; label 1 is for active, while label 0 is for non-active frames. Note that the cross-entropy loss is only used to train the back-end classifier, not to fine-tune the SyncNet layers.

5. Experiments

In this section, the performances of the reference systems on the ASW dataset are evaluated using both self-supervised and supervised methods. We use three metrics to evaluate the ASD performance - AP, AUROC, and EER.

Self-supervised method. SyncNet extracts audio and visual features from 0.2 second segment, moving 0.04 second at a time. The similarities between audio-visual features are measured using either the inverse of L2 distance or cosine similarity. An audio-visual pair is determined to be active, when the similarity score of it is higher than a pre-defined threshold.

Supervised method. Here, we adopt the front-end feature extractor from the self-supervised method and add the back-end classifier on the top of it. The back-end classifier consists of BGRUs and FC layers, as shown in Figure 4. BGRUs have two layers with 128 nodes each. FC layers have 128 and 2 nodes, respectively. The front-end feature extractor is fixed, and only the back-end classifier is fine-tuned. The numbers of active and non-active labels are balanced in a batch, training the back-end

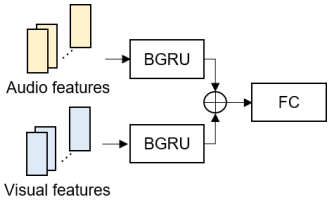

Figure 4: *Back-end classifier of the supervised method. The audio and the visual features are generated from the front-end feature extractor, SyncNet, and ingested into BGRU layers. The outputs of BGRU are concatenated and fed into two FC layers, in order to determine activeness.*

Table 3: *Reference performances on the ASW dataset using the self-supervised and supervised methods. The similarity score of the self-supervised method is measured by the inverse of L2 distance or cosine similarity. The supervised method is a variant of the Activity challenge winning system in 2019 using BGRUs.*

method		**AP**	**AUROC**	**EER**
Self-supervised	L2	0.890	0.944	0.116
	cosine	0.905	0.951	0.098
Supervised		0.966	0.972	0.062

classifier.

Results. Table 3 shows the performance of two methods. The use of cosine similarity shows better performance than the inverse of L2 distance in case of the self-supervised method. The supervised method shows 6.3% higher AP than the self-supervised method.

5.1. Discussion

One of the main properties that differentiates the ASW dataset from the AVA-ActiveSpeaker dataset is that ASW does not include the dubbed videos, in which the audio is recorded on different languages. Note that the dubbed videos in the AVA-ActiveSpeaker dataset are annotated as active frames without any indicator. When ASD is used as a pre-processing step for audio-visual speech processing, the positive labels allocated to dubbed videos can be regarded as mislabeled, and we hypothesise that it might cause the trained network to produce more false alarms. We propose two experiments to show the effect of the dubbed videos.

First, we conduct a cross-domain evaluation. Specifically, an identical model is trained using each dataset, and a threshold is tuned to meet EER using the corresponding validation set. FAR and FRR of ASW are calculated using the test set to avoid overfitting, whereas FAR and FRR of AVA-ActiveSpeaker are calculated using the validation set, as the annotation of test set is not available. Then we observe how the FAR and FRR change when the evaluation sets are the same. Table 4 describes the results. Comparing the first and the fourth row, we can analyze the effect of different training datasets on ASW dataset. It is acceptable that the total error increases because of domain differences of datasets. However, it is noteworthy that the specific type of error has increased significantly. In particular, in the model trained with the AVA-ActiveSpeaker dataset, the proportion of FAR has increased significantly, and we interpret that this result is caused by the dubbed videos in the AVA-ActiveSpeaker dataset. This is because, in the training process using the AVA-ActiveSpeaker dataset, dubbed videos with no correspondence between audio-visual pairs are forced to be classified as active frames, making the model easily mis-detect non-active frames. We also find the same tendency by comparing the second row and the third row.

Table 4: *Cross-domain evaluation. The thresholds of the first and the third row are adopted to the second and the fourth row, respectively, in order to see how the FAR and FRR changes. (AVA: AVA-ActiveSpeaker)*

Train data	**Eval data**	**Thres.**	**FAR**	**FRR**
ASW	ASW	0.65	0.05	0.08
ASW	AVA	0.65	0.04	0.74
AVA	AVA	0.43	0.21	0.21
AVA	ASW	0.43	0.30	0.07

Table 5: *The ratio of active frames detected in dubbed videos. All frames in dubbed videos should not be detected as active since there is no correspondence between audio and visual streams. (AVA: AVA-ActiveSpeaker)*

Video id \ **Train data**	**ASW**	**AVA**
1odS6ynbWNo	0.23	0.31
KQViP7O8t98	0.43	0.67
krCpn6RrNX8	0.27	0.39
x3D6dr-feaU	0.27	0.48
ZiOJq6PMkls	0.12	0.28
Average	0.26	0.43

The second experiment uses five dubbed videos downloaded from YouTube. Note that these videos are included in neither ASW nor AVA-ActiveSpeaker. In the viewpoint of speech pre-processing, all active frames detected from the dubbed video are false alarms because there is no correspondence between the video and the audio streams. Therefore, we expect that the precision of the method trained by ASW or AVA-ActiveSpeaker can be estimated using the ratio of detecting active frames from the dubbed video (*lower is better*). Table 5 shows the ratio of active frames detected by the method trained by each dataset and their average value. The method trained by ASW shows 17% lower false alarms in average. It shows that false positive labels included in AVA-ActiveSpeaker reduce the precision of the system.

6. Conclusion

In this paper, we present a novel ASD dataset, ASW, and the corresponding reference systems. ASW contains 33.7 hours of videos, in which positive and negative labels are balanced. The baseline performances are measured by AP, AUROC and EER, adopting the self-supervised and supervised systems. ASW does not include dubbed videos unlike the existing dataset, AVA-ActiveSpeaker. Cross-domain evaluation and case study show that the dubbed videos cause the threshold shift and lower the precision of the systems. We expect that ASW will be well-applicable to real-world scenarios since it contains natural human speech and minimum false positive labels.

7. References

[1] R. Yan, Y. Song, and H. Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016.

[2] H. Gamboa and A. L. Fred, "An identity authentication system based on human computer interaction behaviour." in *Proceedings of the Pattern Recognition in Information Systems*, 2003.

[3] H. Gamboa and A. Fred, "A behavioral biometric system based on human-computer interaction," in *Biometric Technology for Human Identification*, vol. 5404. International Society for Optics and Photonics, 2004, pp. 381–392.

[4] S.-W. Chung, H. G. Kang, and J. S. Chung, "Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision," in *INTERSPEECH*, 2020.

[5] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019.

[6] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[7] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," in *INTERSPEECH*, 2017.

[8] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[9] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[10] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[12] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, and I. Sturdy, "Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers," *arXiv preprint arXiv:1706.00079*, 2017.

[13] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the International Conference on Computer Vision*, 2019.

[14] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the International Conference on Computer Vision*, 2019.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[18] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[19] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proceedings of the Asian Conference on Computer Vision*, 2016.

[20] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[21] J. S. Chung and A. Zisserman, "Lip reading in profile," in *Proceedings of the British Machine Vision Conference*, 2017.

[22] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[23] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *INTERSPEECH*, 2018.

[24] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[25] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[26] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Facefilter: Audio-visual speech separation using still images," in *INTERSPEECH*, 2020.

[27] J. Lee, S.-W. Chung, S. Kim, H.-G. Kang, and K. Sohn, "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[28] Y. Ding, Y. Xu, S.-X. Zhang, Y. Cong, and L. Wang, "Self-supervised learning for audio-visual speaker diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[29] J. S. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," in *INTERSPEECH*, 2019.

[30] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[31] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[32] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[33] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," in *INTERSPEECH*, 2020.

[34] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[35] B. Castellano, "Pyscenedetect," 2018.

[36] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proceedings of the International Conference on Computer Vision*, 2017.

[37] A. Dutta, A. Gupta, and A. Zissermann, "Vgg image annotator (via)," *URL: http://www. robots. ox. ac. uk/˜ vgg/software/via*, 2016.

[38] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[39] ——, "Perfect match: Self-supervsied embeddings for cross-modal retrieval," *IEEE Journal of Selected Topics of Signal Processing*, vol. 14, no. 3, pp. 568–576, 2020.

[40] J. S. Chung, "Naver at activitynet challenge 2019–task b active speaker detection (ava)," *arXiv preprint arXiv:1906.10555*, 2019.

[41] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.