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

Audio-Visual Information Fusion Using Cross-modal Teacher-Student Learning for Voice Activity Detection in Realistic Environments

Hengshun Zhou[1], *Jun Du*[1,*], *Hang Chen*[1], *Zijun Jing*[2], *Shifu Xiong*[2], *Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]iFlytek Research, Hefei, Anhui, P.R.China
[3]Georgia Institute of Technology, Atlanta, GA. USA

zhhs@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn, ch199703@mail.ustc.edu.cn,
zjjing2@iflytek.com, sfxiong@iflytek.com, chl@ece.gatech.edu

Abstract

We propose an information fusion approach to audio-visual voice activity detection (AV-VAD) based on cross-modal teacher-student learning leveraging on factorized bilinear pooling (FBP) and Kullback-Leibler (KL) regularization. First, we design an audio-visual network by using FBP fusion to fully utilize the interaction between audio and video modalities. Next, to transfer the rich information in audio-based VAD (A-VAD) model trained with a massive audio-only dataset to AV-VAD model built with relatively limited multi-modal data, a cross-modal teacher-student learning framework is then proposed based on cross entropy with regulated KL-divergence. Finally, evaluated on an in-house dataset recorded in realistic conditions using standard VAD metrics, the proposed approach yields consistent and significant improvements over other state-of-the-art techniques. Moreover, by applying our AV-VAD technique to an audio-visual Chinese speech recognition task, the character error rate is reduced by 24.15% and 8.66% from A-VAD and the baseline AV-VAD systems, respectively.

Index Terms: Voice activity detection, audio-visual information fusion, factorized bilinear pooling, teacher-student learning, KL-regularization

1. Introduction

Voice activity detection (VAD) aims to separate speech and nonspeech segments from the target audio. It is an essential front-end component in many speech processing applications, such as speech recognition, speaker recognition, speech enhancement and sound event detection [1, 2, 3, 4]. The accuracy of speech/nonspeech detection is severely degraded when the speech signal is distorted by noise [5, 6]. In the realistic adverse environments, VAD is still a challenging problem which has been investigated for many years [7, 8].

The early approaches for VAD are generally based on statistical signal processing [9, 10], consisting of a feature extraction stage followed by a speech/nonspeech classifier. The common features used for VAD include energy, zero-crossing rate, cepstral coefficients, autocorrelation features and Mel-frequency cepstral coefficients (MFCC) [11, 12]. The classical Gaussian model was widely used for VAD [13]. Sohn et al. [9] devised the VAD based on a Gaussian statistical model by employing the decision rule based on the geometric mean of the likelihood ratio. An unsupervised learning framework was proposed to construct statistical models for VAD by a sequential Gaussian mixture model in [14] and the evaluations effectively showed its promising performance in comparison with some typical semi-supervised VAD. The machine learning-based methods are also considered for VAD back-ends, such as classification and regression tree (CART), Gaussian likelihood ratio test (LRT) [9, 15]. In [16], support vector machine (SVM) were investigated to improve receiver operating characteristic (ROC) curve.

Recently, deep learning (DL) based VAD in particular have attracted much attention [17, 18]. A joint training approach to VAD to address the issue of performance degradation due to unseen noise conditions was presented in [19]. Fan et al. [20] proposed to optimize the area under ROC curve (AUC) by deep neural networks (DNN), which can maximize the performance of VAD in terms of the ROC curve. Convolution neural networks (CNNs) and recurrent neural networks (RNNs) have several properties that make them popular choices for VAD [18, 21]. In [22], the authors proposed a novel adaptive VAD approach to control the speech/non-speech decision in naturalistic environments. For the most commonly used downstream tasks, automatic speech recognition and voice activity detection were integrated in an end-to-end manner in [23]. Considering the complexity of the realistic conditions, some researchers introduce the visual information to VAD system. Previous studies have demonstrated that adding visual-based voice activity detection systems relying on facial features can improve the performance in noisy environments [24, 25, 26, 27]. However, in real applications, audio-visual cross-modal databases are usually far smaller than available audio-only databases due to the complex recording conditions and high cost. So how to deeply transfer the rich information in audio-based VAD (A-VAD) model to audio-visual VAD (AV-VAD) model and fully utilize the information of both audio and video modalities is quite important to explore, which is also focus of this study.

Inspired by the work in audio-visual speech recognition [28], we present a novel information fusion approach to AV-VAD based on factorized bilinear pooling (FBP) and Kullback-Leibler (KL) regularization. Specifically, we first design an audio-visual network by using FBP fusion to fully utilize the information interaction between audio modality and video modality. Next, a cross-modal teacher-student learning framework by minimizing the cross entropy (CE) with KL-divergence regularization is utilized to deeply transfer the rich information in A-VAD model trained using massive audio data to AV-VAD model built with relatively limited multi-modal data. Evaluated on an in-house dataset for AV-VAD task, the proposed approach can yield consistent and significant improvements on the standard VAD metrics compared with other single-modal and multi-modal methods. Furthermore, by applying our AV-VAD system to an audio-visual Chinese speech recognition task, character error rate (CER) is reduced by 24.15% and 8.66% compared

http://dx.doi.org/10.21437/Interspeech.2021-592

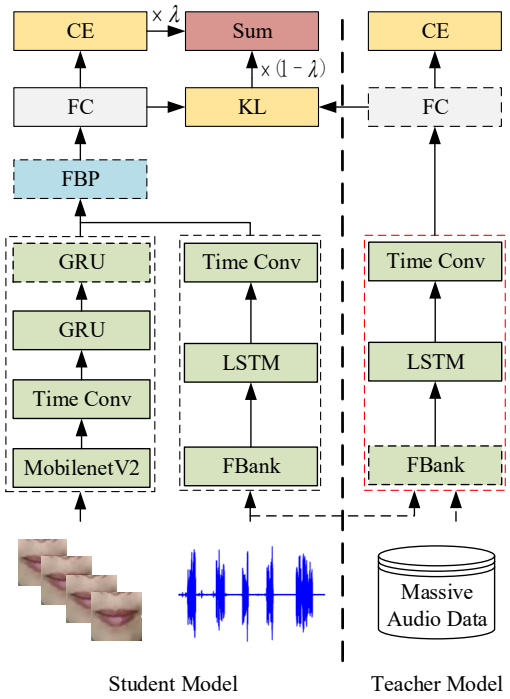

Figure 1: *AV-VAD architecture based on FBP audio-visual fusion and KL-regularized teacher-student learning.*

with A-VAD system and the baseline AV-VAD system respectively.

The remainder of this paper is organized as follows. Section 2 gives detailed description of our proposed approach. In Section 3, experimental results and analysis are discussed. Finally, we conclude in Section 4.

2. Proposed Approach for AV-VAD

The overall flowchart of proposed AV-VAD model architecture is illustrated in Figure 1, consisting of teacher model training using audio-only data and student model training using audio-visual data. In the training stage, first the large-scale external audio data is used to build the teacher model using CE criterion. Next, audio and video streams are fused as the student model to utilize multimodal information via FBP. Finally, teacher-student learning using weighted CE with KL-regularization is conducted to improve the student model. The details are elaborated in the following subsections.

2.1. Teacher Model Training Using Extensive Audio Data

For the teacher model training, inspired by [21], the extracted acoustic features are fed into a LSTM layer followed by a time convolution layer which is used to utilize acoustic context information, and then a fully connected (FC) layer is adopted to classify speech/nonspeech frames. We use 40-dimensional filter bank (FBank) features normalized by global mean and variance to train the audio teacher model. The LSTM layer has 416 cells. The kernel size of the time convolution layer is set to 5×1. The whole network is optimized by minimizing the cross entropy criterion as:

$$l_{\mathrm{CE}}^{\mathrm{T}} = -\sum_{t} \log(P_{\mathrm{T}}(y_t|\boldsymbol{x}_{a,t})), \quad y_t \in \{0,1\} \tag{1}$$

where $P_{\mathrm{T}}(y_t|\boldsymbol{x}_{a,t})$ is the speech/nonspeech class posterior of y_t given $\boldsymbol{x}_{a,t}$ generated by the teacher audio model, $\boldsymbol{x}_{a,t}$ denotes the acoustic feature vector at the t-th frame, and y_t is the corresponding VAD label at the t-th frame. The teacher model is trained using 60,000 hours audio data, which is commercially used in iFlytek's speech products.

2.2. Audio-Visual Student Model Using FBP Fusion

For the audio-visual student model, the structure of audio stream initially copies from the teacher model, except that the final fully connected layer is removed. For the video stream, considering practicality and lightweight, we choose a combination of MobileNetV2 [29], one time convolution layer and two GRU layers, as shown in Figure 1. In this study, 13 linear bottlenecks are adopted in MobileNetV2 as a feature extractor. The grayscale lip reshaped to 64×64 is used as the video network input, and the output is an encoded vector by using average pooling. For details, please refer to [29]. We employ cross entropy loss to train the video-based VAD (V-VAD) system by adding a fully connected layer at the end of the video stream in Figure 1, which will be mentioned in the experiments.

Inspired by [30], a direct concatenation for audio-visual fusion at the encoder is first considered. Due to the effectiveness of bilinear pooling in visual question answering (VQA) tasks, we apply factorized bilinear pooling for AV-VAD fusion. Bilinear pooling is introduced in [31] and initially used for feature fusion. Then the fused vectors are used for classification. Although the system performance is improved, it also brings a huge amount of computation. Some researches on reducing computational cost have achieved the considerable results [32, 33]. According to [33], given two feature vectors in different modalities, i.e. the audio encoder vector $\boldsymbol{a} \in \mathbb{R}^M$ (M=384) and the visual encoder vector $\boldsymbol{v} \in \mathbb{R}^N$ (N=64), the simplest cross-modal bilinear model is defined as follows:

$$z_i = \boldsymbol{a}^\top \boldsymbol{W}_i \boldsymbol{v} \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{M\times N}$ is a projection matrix, $z_i \in \mathbb{R}$ is the output of bilinear model. We use the Eq. (3) to obtain the output fusion vector $\boldsymbol{z} = [z_1, \cdots, z_O]$. The formula derivation from formula Eq. (2) to Eq. (3) was described in [33].

$$\boldsymbol{z} = \mathrm{SumPooling}(\widetilde{\boldsymbol{U}}^\top \boldsymbol{a} \circ \widetilde{\boldsymbol{V}}^\top \boldsymbol{v}, K) \tag{3}$$

where the function $\mathrm{SumPooling}(\boldsymbol{x}, K)$ applies sum pooling within a series of non-overlapped windows to $\boldsymbol{x}$. $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{M\times KO}$ and $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{N\times KO}$ are reshaped parameter 2-D matrices to be learned. In our experiment, O and K are set to 128 and 4 respectively. $\circ$ represents the element-wise multiplication of two vectors. Besides, dropout is adopted to prevent over-fitting. The L2-normalization is used to normalize the energy of $\boldsymbol{z}$ to avoid the dramatical variation of the output magnitude. The whole network is optimized in an end-to-end manner by minimizing cross entropy similar to Eq. (1).

2.3. Cross-Modal Teacher-Student Training

Considering that audio-visual cross-modal databases are usually far smaller than available audio-only databases due to the complex recording conditions and high cost, we explore how to use the abundant external audio data to improve the performance of the cross-modal model. We present a teacher-student learning framework based on KL regularization and FBP fusion. We find that for the VAD task in realistic noisy conditions, adding video modality can bring significant performance

Table 1: *Detailed data distribution of our AV-VAD training corpus in different environments.*

Environment	TV Program	Mobile Phone	Car
Size	220h	65h	235h

improvements (see section 3.2 for details). Unlike in [28] where using video modality can not bring significant gains for audio-visual speech recognition task, it is not adequate to use just KL-divergence as the loss function for the teacher-student learning in AV-VAD. Accordingly, here KL-divergence is adopted as a regularization term for traditional cross entropy loss, which is calculated by using the output of the teacher network and the audio stream in student network. In this way, the impact of video modality can be enhanced and the trade-off between CE and KL-divergence leads to better VAD performance. The KL-divergence and CE loss are calculated as:

$$l_{\mathrm{KL}}^{\mathrm{S}} = \sum_{t} P_{\mathrm{T}}(y_t|\boldsymbol{x}_{a,t}) \log \frac{P_{\mathrm{T}}(y_t|\boldsymbol{x}_{a,t})}{P_{\mathrm{S}}(y_t|\boldsymbol{x}_{a,t}, \boldsymbol{x}_{v,t})} \quad (4)$$

$$l_{\mathrm{CE}}^{\mathrm{S}} = -\sum_{t} \log(P_{\mathrm{S}}(y_t|\boldsymbol{x}_{a,t}, \boldsymbol{x}_{v,t})), \quad y_t \in \{0, 1\} \quad (5)$$

where $P_{\mathrm{T}}(y_t|\boldsymbol{x}_{a,t})$ and $P_{\mathrm{S}}(y_t|\boldsymbol{x}_{a,t}, \boldsymbol{x}_{v,t})$ are the speech or non-speech class posteriors generated by teacher and student networks, $\boldsymbol{x}_{a,t}$ and $\boldsymbol{x}_{v,t}$ denote the acoustic feature vector and visual feature vector at the t-th frame. Then, the final loss function is:

$$L = \lambda \times l_{\mathrm{CE}}^{\mathrm{S}} + (1 - \lambda) \times l_{\mathrm{KL}}^{\mathrm{S}} \quad (6)$$

where λ is a weighting factor set to 0.7 in our experiments, which ranges from 0 to 1. L is used to optimize the whole student audio-visual network and the parameters of the teacher network are fixed.

3. Experiments and Result Analysis

3.1. Databases and Implementation Details

We conduct VAD experiments on an audio-only dataset of about 60,000 hours and an audio-visual dataset of about 520 hours collected in realistic conditions. All these in-house data are collected by iFlytek. Table 1 reports detailed data distribution of our AV-VAD training corpus in different environments. We use the following standard metrics for evaluation of VAD performance: accuracy, precision, recall and AUC. Both the validation set and test set include about 6-hour audio-visual data recorded in the car environment.

We further evaluate VAD using character error rate (CER) by applying it to an audio-visual Chinese speech recognition (AVCSR) task with a large vocabulary of 14,835 words. The AVCSR model architecture [34], having a hybrid CNN-HMM acoustic model is with 15,003 tied states. To train the AVCSR model, we first initialize the audio-stream parameters by pre-training using about 50,000h audio-only data and the video-stream parameters by pre-training using about 300h video-only data. Then about 150 hours of audio-visual data recorded in car environments are employed to fine-tune the whole AVCSR model. The test set consists of about 3-hour of audio-visual data in the same car environments. A 3-gram language model is adopted for decoding.

We use pytorch to train all VAD networks and minimize the loss function using the Adam optimization method. The batch size is 128 and the dropout probability is equal to 0.2 to prevent over-fitting. The learning rate is 0.005 for A-VAD and 0.0001 for V-VAD and AV-VAD. For the selection of O and K, we take $O = 128$ as the benchmark and the range of K is $2 \sim 10$. For λ, we select it from 0 to 1 with a step of 0.1, and the value with the best result on the validation set is determined.

3.2. Results on Standard VAD Metrics

First we evaluate the performance of the single-modal systems. We train the A-VAD and the V-VAD system respectively, corresponding to the audio stream and video stream of the student model in Figure 1, by minimizing the cross entropy loss. The A-VAD model is trained by external audio data in addition to the audio portion of the given audio-visual training corpus. The V-VAD model is built based on video portion of audio-visual training corpus. Our results are presented in Table 2. We can observe that the better results were achieved by A-VAD system compared with V-VAD in all four metrics. For example, the performance gap for precision measure is 11.71% between A-VAD and V-VAD.

Table 2: *Test set performance on standard metrics for VAD systems. [Acc: accuracy, Pre: precision, Rec: recall]*

System	Acc(%)	Pre(%)	Rec(%)	AUC(%)
V-VAD	88.59	77.80	90.51	95.49
A-VAD	94.74	89.51	94.86	98.85
Concat	94.99	89.48	95.75	98.99
FBP	95.54	90.53	96.28	99.05
TS-Concat	95.06	89.48	96.00	99.03
TS-FBP	95.73	91.03	96.28	99.16
TS [28]	88.03	74.46	95.84	92.48

Next, we compare different AV-VAD systems. Inspired by [30], a direct concatenation for audio-visual fusion at the encoder is implemented and denoted as 'Concat' in Table 2, which yielded remarkable improvements for accuracy, recall and AUC measures. Moreover, the audio-visual fusion using FBP achieved consistent performance gains over audio-visual concatenation for all four VAD metrics.

Finally, our proposed cross-modal teacher-student (TS) learning is applied for both AV-VAD systems based on feature concatenation ('TS-Concat') and FBP fusion ('TS-FBP'). The teacher-student (TS) learning method in [28] which has been successfully used for audio-visual speech recognition is also utilized for comparison. Our TS learning methods yielded additional improvements on top of both 'Concat' and 'FBP' systems. However, the performance of TS method in [28] was much worse as it only employs KL-divergence loss for optimization without the constraint of CE loss. Overall, 'TS-FBP' achieved the best performance for all VAD measures. The similar performance trend can be also observed from Figure 2, which shows the ROC curves for different systems, obtained by varying the classification threshold. Note that the 'TS-FBP' curve is superior to other systems, which is consistent with our previous results in Table 2.

Furthermore, we make a comparison of the learning curves between 'TS-FBP' and TS system in [28] on the validation set.

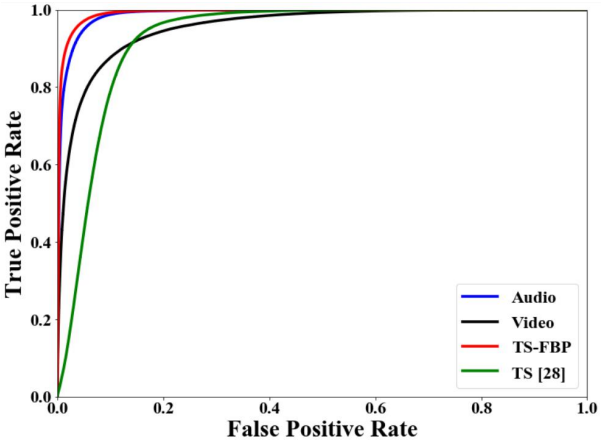

Figure 2: *Comparison of ROC curves of different VAD systems.*

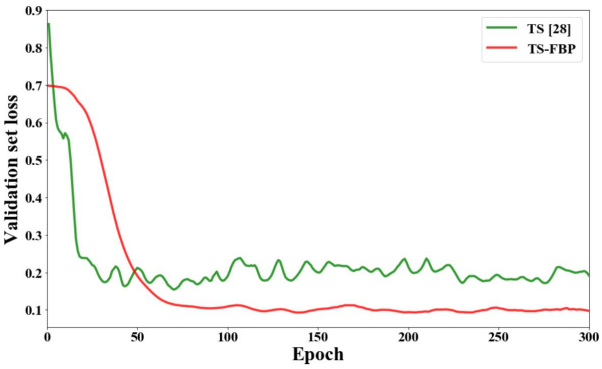

Figure 3: *Comparison of learning curves between TS-FBP and TS [28] system.*

As illustrated in Figure 3, the generalization ability of 'TS-FBP' can be enhanced, and the fluctuation of the curve is smaller. Accordingly, our approach can achieve a smaller loss which leads to better system performance.

Table 3: *Test set performance of different VADs for AVCSR. [I: insertion error, D: deletion error, S: substitution error]*

System	I(%)	D(%)	S(%)	CER(%)
V-VAD	1.76	13.96	8.75	24.47
A-VAD	6.90	6.26	9.37	22.53
Concat	1.52	8.83	8.36	18.71
FBP	1.54	8.07	8.39	18.00
TS-Concat	1.77	8.35	8.28	18.40
TS-FBP	1.54	7.30	8.25	17.09
TS [28]	6.59	7.01	8.89	22.49

3.3. Results on CER for AVCSR System

We further evaluate the VAD system performance on AVCSR and the results are shown in Table 3. Different from audio-visual speech recognition task in [28], the performance of VAD systems have been significantly improved after fusing the video modality. Specifically, according to Table 3, compared with A-VAD, the system performance of the direct concatenation at the encoder is relatively improved by 16.96%. This is our motivation to introduce the weighted CE loss with KL-regularization. After using FBP for audio-visual fusion, the CER yielded an absolute accuracy gain of 0.71% over the 'Concat' based approach. After using the proposed teacher-student learning through KL regularization, the performance of these two audio-visual fusion systems has been further improved. And 'TS-FBP' achieved the best CER result of 17.09%.

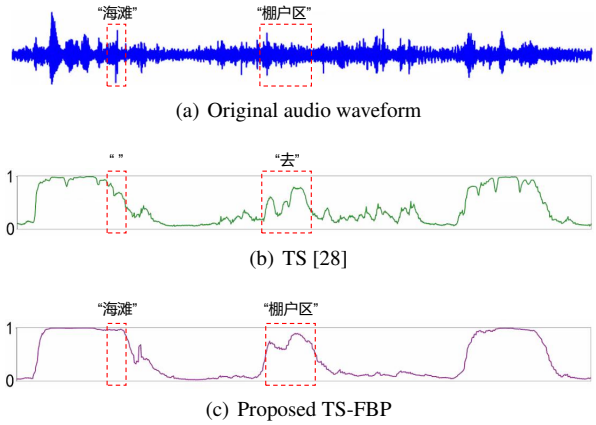

Figure 4: *Analysis on one utterance example.*

Figure 4 gives an utterance example from the real test set. Figure 4(a) plots the audio waveform with background music noise. The frame-level VAD probabilities of the utterance provided by the TS method in [28] and the proposed TS-FBP are shown in Figure 4(b) and Figure 4(c) respectively. Clearly, the TS-FBP output is much smoother when compared to TS in [28]. Based on the AVCSR results, it seems that the TS method in [28] led to more deletion and substitution errors due to speech segments missing after AV-VAD, as shown in the two red dotted boxes in Figure 4(a)-(c). Our proposed TS-FBP for AV-VAD could well preserve those speech segments and generate correct recognition results. The reason is that our approach can fully utilize the video information by introducing a weighted loss of CE and KL-divergence. In this example, the input for the video stream only contains the target speaker's lip, and does not include the interferer's lip (e.g., singers). This way it is more robust to noise interferences.

4. Conclusion

We present a novel information fusion approach to audio-visual voice activity detection based on teacher-student learning through KL regularization and factorized bilinear pooling fusion. FBP fusion demonstrates its superiority over the simple audio-visual feature concatenation. Then cross-modal teacher-student learning is verified to be effective to transfer the rich information in audio-only teacher model to the audio-visual student model. Evaluated on a real-world recorded dataset, the proposed approach yields consistent improvements over state-of-the-art techniques in terms of standard VAD metrics and CER for audio-visual speech recognition.

5. Acknowledgements

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.

6. References

[1] J. Ramirez, J. C. Segura, J. M. Gorriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[2] M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 295–313, 2014.

[3] R. Lin, C. Costello, C. Jankowski, and V. Mruthyunjaya, "Optimizing Voice Activity Detection for Noisy Conditions," in *20th Annual Conference of the International Speech Communication Association*, 2019, pp. 2030–2034.

[4] M. W. Yefei Chen, Heinrich Dinkel and K. Yu, "Voice activity detection in the wild via weakly supervised sound event detection," in *21st Annual Conference of the International Speech Communication Association*, 2020, pp. 3665–3669.

[5] W. A. Jassim and N. Harte, "Voice activity detection using neurograms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5524–5528.

[6] H. Z. Tianjiao Xu and X. Zhang, "Joint training rescnn-based voice activity detection with speech enhancement," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC*, 2019, pp. 1157–1162.

[7] I. C. Amir Ivry and B. Berdugo, "Evaluation of deep-learning-based voice activity detectors and room impulse response models in reverberant environments," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 406–410.

[8] R. B. H. B. Marvin Lavechin, Marie-Philippe Gill and L. P. García-Perera, "End-to-end domain-adversarial voice activity detection," in *21st Annual Conference of the International Speech Communication Association*, 2020, pp. 3685–3689.

[9] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[10] L. E. A. Scott Wisdom, Greg Okopal and J. W. Pitton, "Voice activity detection using subband noncircularity," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4505–4509.

[11] H. Ghaemmaghami, D. Dean, S. Kalantari, S. Sridharan, and C. Fookes, "Complete-linkage clustering for voice activity detection in audio and visual speech," in *16th Annual Conference of the International Speech Communication Association*, 2015, pp. 2292–2296.

[12] R. Muralishankar, D. Ghosh, and S. Gurugopinath, "A novel modified mel-dct filter bank structure with application to voice activity detection," *IEEE Signal Processing Letters*, vol. 27, pp. 1240–1244, 2020.

[13] A. Sholokhov, M. Sahidullah, and T. Kinnunen, "Semi-supervised speech activity detection with an application to automatic speaker verification," *Comput. Speech Lang.*, vol. 47, no. C, p. 132156, 2018.

[14] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[15] P. N. L. Kaavya Sriskandaraja, Vidhyasaharan Sethu and E. Ambikairajah, "A model based voice activity detector for noisy environments," in *16th Annual Conference of the International Speech Communication Association*, 2015, pp. 2297–2301.

[16] J. Wu and X. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 466–469, 2011.

[17] G. Gelly and J. Gauvain, "Optimization of rnn-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, 2018.

[18] F. Martinelli, G. Dellaferrera, P. Mainar, and M. Cernak, "Spiking neural networks trained with backpropagation for low power neuromorphic implementation of voice activity detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8544–8548.

[19] Q. Wang, J. Du, X. Bao, Z. Wang, L. Dai, and C. Lee, "A universal VAD based on jointly trained deep neural networks," in *16th Annual Conference of the International Speech Communication Association*, 2015, pp. 2282–2286.

[20] Z. Fan, Z. Bai, X. Zhang, S. Rahardja, and J. Chen, "Auc optimization for deep learning based voice activity detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6760–6764.

[21] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7378–7382.

[22] R. K. D. Bidisha Sharma and H. Li, "Multi-Level Adaptive Speech Activity Detector for Speech in Naturalistic Environments," in *20th Annual Conference of the International Speech Communication Association*, 2019, pp. 2015–2019.

[23] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with ctc-based voice activity detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6999–7003.

[24] R. T. D. Dov and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 23, no. 4, pp. 732–745, 2015.

[25] F. Tao and C. Busso, "Bimodal recurrent neural network for audiovisual voice activity detection," in *18th Annual Conference of the International Speech Communication Association*, 2017, pp. 1938–1942.

[26] M. Buchbinder, Y. Buchris, and I. Cohen, "Adaptive weighting parameter in audio-visual voice activity detection," in *Science of Electrical Engineering*, 2017.

[27] C. B. Fei Tao, "Audiovisual speech activity detection with advanced long short-term memory," in *19th Annual Conference of the International Speech Communication Association*, 2018, pp. 1244–1248.

[28] W. Li, S. Wang, M. Lei, S. M. Siniscalchi, and C. Lee, "Improving audio-visual speech recognition performance with cross-modal student-teacher training," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6560–6564.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[30] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "Ava active speaker: An audio-visual dataset for active speaker detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4492–4496.

[31] T. Y. Lin, A. Roychowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[32] G. Yang, O. Beijbom, Z. Ning, and T. Darrell, "Compact bilinear pooling," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[33] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *IEEE International Conference on Computer Vision (CVPR)*, 2017, pp. 1839–1848.

[34] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.