See the silence: improving visual-only voice activity detection by optical flow and RGB fusion*

Danu Caus[0000−0003−0597−8844], Guillaume Carbajal[0000−0003−4077−1742], Timo Gerkmann[0000−0002−8678−4699], and Simone Frintrop[0000−0002−9475−3593]

Department of Informatics, University of Hamburg, Germany
{danu.caus, guillaume.carbajal, timo.gerkmann, simone.frintrop}@uni-hamburg.de

Abstract. In this work, we propose a novel approach for visual voice activity detection (VAD), which is an important component of audio-visual tasks such as speech enhancement. We focus on optimizing the visual component and propose a two-stream approach based on optical flow and RGB data. Both streams are analyzed by long short-term memory (LSTM) modules to extract dynamic features. We show that this setup clearly improves the one without optical flow. Additionally, we show that focusing on the lower face area is superior to processing the whole face, or only the mouth region as usually done. This aspect involves practical advantages, since it facilitates data labeling. Our approach especially improves the true negative rate, which means we detect frames without speech more reliably — we see the silence.

Keywords: Visual voice activity detection · Optical flow · Ensemble learning

1 Introduction

Voice activity detection (VAD) is the task of identifying the presence or absence of human speech segments in a stream of input data. Depending on the input modality used, we can distinguish between visual VAD, which uses only images to detect voice activity, and audio VAD, which separates the voice audio signal from the background noise signal. VAD is an important component in a variety of applications, such as speech enhancement [1, 2] or source separation [3, 4]. While most approaches in this field have focused on processing the audio data [5], there is a currently growing effort to combine audio VAD and visual VAD sub-systems. Both modalities complement each other and the joint processing obtains superior performance. Visual voice activity detection in particular is more robust to noisy

* This work was supported by the Alliance of Hamburg Universities for Computer Science (ahoi.digital) as part of the Adaptive Crossmodal Sensor Data Acquisition (ACSDA) research project and by the German Research Foundation (DFG) in the Transregional project Crossmodal Learning (TRR 169).

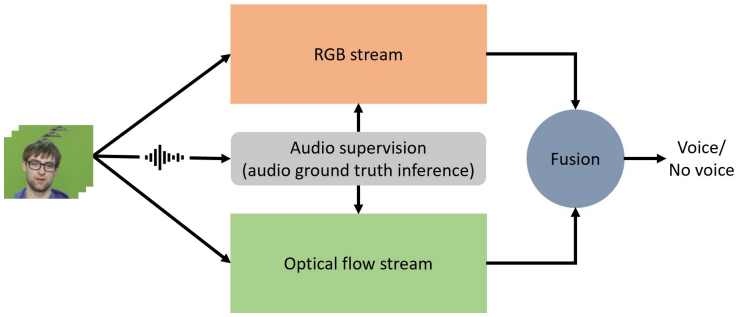

Fig. 1. Overview of our approach, consisting of an RGB and an optical flow stream. Both streams are fused to determine if voice activity is present in a frame or not, exploiting the complementary information of the streams. The audio channel from the video is used to generate ground truth labels that supervise the training of RGB and optical flow streams.

speech [6] and identifying whisper speech [7] than its audio-only counterpart. The merged audio-visual systems will generally be better if each of the individual components is optimized. However, most research focuses on the audio part, whereas the visual modality of VAD systems is often overlooked and standard solutions are applied, like extracting visual features with a ResNet before fusing them with the audio signals [8].

In this work, we focus on optimizing the visual VAD component. In particular, our aim is to improve the ability of the network to detect frames in which there is no speech, which will help reduce noise more effectively. We propose a novel visual VAD system, which combines RGB features and optical flow. Both modalities focus on different aspects of the data and are thus complementary. While the RGB stream extracts color features, the optical flow stream focuses on motion features, which is especially useful in detecting when someone is speaking. LSTM modules allow the system to exploit dynamic features and probabilistic-based fusion combines the two streams in an ensemble manner. Fig. 1 shows an overview of our system.

We focus our attention on facial features for voice activity detection, as opposed to body language and body motion, which are useful for large, in-the-wild scenes. Hence, we use a constrained dataset, were people's faces are clearly visible: the TCD-TIMIT dataset [9]. The contribution of this paper is twofold:

1. We show that adding optical flow to a deep learning approach via probabilistic fusion results in improving the true negative rate of the system (TNR) by as much as 7%. This is useful in audio-visual systems that rely on TNR in order to reduce transient noise and enhance speech.
2. We show that focusing on the lower face area of the speaker, including not only the mouth, but also nose and chin, offers better results than considering the mouth-only region, or entire face as done in previous work. Compared to extracting mouth-only regions of interest, our pre-processing is also more

practical for real-world applications, avoiding precise mouth croppings and fine alignments, which require manual adjusting.

2 Related Work

The task of voice activity detection has a variety of solutions, which can broadly be classified as: audio-only, visual-only, and audio-visual combined approaches. We give an overview of relevant work in this section.

Audio-only VAD The task of voice activity detection has a rich history in the audio community. One important solution is based on sound energy thresholding to infer the VAD label such as [10]. Other approaches such as [11] have used statistical models which include the background noise statistics. These statistical models are robust to noise and also allow for computing the a posteriori probability of speech presence [12], i.e. to have a soft decision instead of a hard decision. More recently, researchers have leveraged deep learning for audio-only VAD and overcame traditional audio methods [13]. Additionally considering the time dimension via recurrent neural networks improved the results even more, as presented in [14].

Visual-only VAD with traditional techniques Early computer vision methods have studied VAD using manually extracted features and traditional machine learning techniques. The work of Joosten et al. [15] has shown that using mouth regions of interest is better than using full faces of speakers, although the authors did not apply their approach to sequence of frames, but rather on individual frames. We will show later that this also holds for videos, but even better results can be obtained using larger crops.

Tao et al. [7] have shown that it is important to have dynamic features rather than static ones. The manual features extracted from the RGB inputs such as mouth width and height, mouth area and perimeter were concatenated in feature vectors and the authors showed that computing first order differences between the vectors, so called dynamic representations, is beneficial for the task of visual VAD. They also used optical flow in this context as a natural dynamic feature.

Chakravarty and Tuytelaars [16] have used a support vector machine approach on full-body images as opposed to looking only at the face of speakers. The rational of the authors is that body language such as hand gesticulation and upper body motion are important especially in a multi-speaker environment such as a conference.

Visual-only VAD using deep learning Surprisingly, visual only voice activity detection using neural networks has not been studied abundantly. The recent work of Guy et al. [17] is the most prominent work in this sense. They have investigated two types of datasets: a) unconstrained, in-the-wild datasets and b) constrained, in-the-lab datasets; as well as two types of architectures: 1) facial landmarks extractor coupled with LSTM and 2) optical flow VGG16 ConvNet

without any LSTM. They come to the conclusion that facial landmarks fed to an LSTM is the best all-round method for both types of datasets. Interestingly, the optical flow approach they investigated registered a particularly unexpected true negative rate for the constrained dataset that they used [18], roughly 15 %, while the natural expectation would have been something in the range of 80%, which the other non-optical flow model they investigated did indeed reach.

Other works with deep learning approach have focused exclusively on in-the-wild scenes. For instance [19] use Hollywood movies which do not have a fixed structure. These types of datasets allow extraction of VAD labels from the subtitle timestamps of the movie, which makes the audio channel unnecessary for ground truth computation. They used 3D CNNs with bidirectional LSTMs on 1 second RGB video snippets and empirically showed that in the process of learning VAD classifications, the neural network also learns to pay attention to the face of the characters.

The recent work of Shahid et al. [20] has a similar use case as the work of [16], namely detecting active speakers during conferences, but uses a deep learning approach instead. Moreover, the work deliberately does not detect body parts of speakers, but rather focuses only on motion segmentation instead and learns to associate body motion with speech activity.

Audio-Visual VAD Merging audio-only and visual-only VAD solutions has yielded synergistic results. In [21] the authors show improvement over single modality baselines by using a rule-based fusion approach which dictates what modality should be treated as the primary component depending on whether face or lip movements are detected or not. Tao et al. [22] show that if both modalities are used together, it is possible to improve the precision of boundary detection between speech and non-speech regions via the Bayesian information criterion/BIC algorithm. In a recent deep-learning framework [8], Ariav and Cohen combined raw audio data with cropped mouth regions as visual input via compact bilinear pooling [23] and obtained better results than with a single modality.

3 Our Approach

Our approach for visual voice activity detection consists of two streams: an RGB stream and an optical flow stream (see Fig. 2). Both streams are trained independently, and combined during the test phase in a probabilistic manner.

It is partially inspired by the ideas of [7], in which the authors use optical flow and manually engineered visual features on which they subsequently compute delta features. We build on this idea of delta features and integrate it into a deep learning framework. We keep the optical flow as a separate stream and use another stream that learns suitable RGB features depending on the dataset. By adding two LSTMs for the RGB and the optical flow stream, we allow the system to explore dynamic features, which is emulating the aforementioned delta features, but in a deep learning context. We note that adding an LSTM to optical

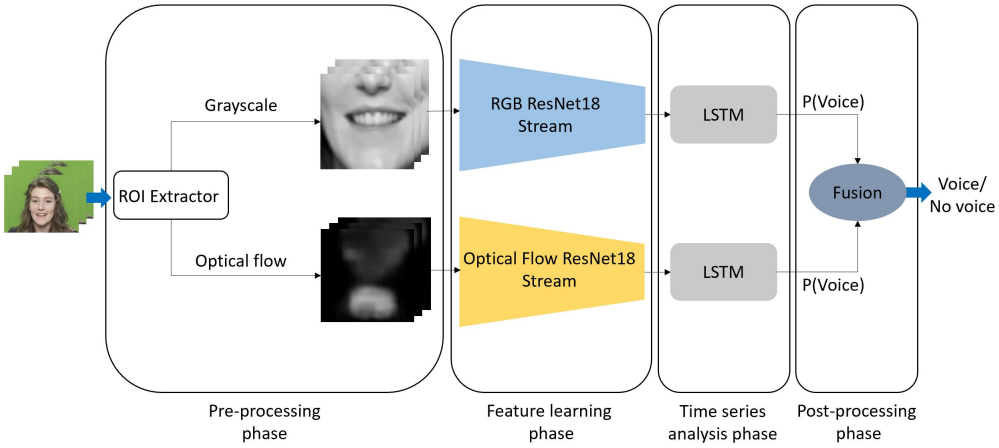

Fig. 2. Diagram of our system: from the input sequence, the lower part of the face is extracted and fed into an RGB branch (top) and an optical flow branch (bottom). In each stream a dedicated ResNet extracts static features, which are then fed into corresponding LSTM modules to exploit dynamic features. The results of both streams are fused via a probabilistic approach. The "ROI Extractor" includes: 1) nose, mouth and chin detection, 2) optical stabilization for a more robust optical flow calculation, and 3) cropping

flow, i.e. treating the optical flow as time series, is not encountered in other recent works. We also note that this setup offers the ability to infer indirectly the acceleration feature. This feature might be useful in future extensions to discriminate between language speech vs laughter or other lip gesticulations that do not involve speech.

Pre-processing: ROI Extractor The first step of our VAD pipeline is a region of interest (ROI) extractor, which consists of detecting the lower face area, stabilizing the results and cropping the image patch. For detecting the lower face area, which contains not only the mouth as in other works, but additionally nose and chin, we first use a facial landmark detector [24]. We detect landmarks for the nose, mouth and chin area and concatenate them in a list. We then compute a center point per frame by averaging all detected landmarks. The center points are averaged over multiple frames (we use the last 30 frames) to achieve optical stabilization. The crop is then created with a size of 67x67 around the optically stabilized ROI center, computed via the running average. Optical stabilization is not critical for the RGB stream, but it does help the optical flow branch achieve better results. For more difficult, in-the-wild datasets, a Kalman filter would be recommended instead of a running average. The resulting lower face crop serves as input for the RGB and the optical flow stream.

RGB stream Our RGB branch is inspired by [8]: using ResNet18, we extract RGB features of the image patch containing the lower face regions. These fea-

tures are then fed into an LSTM module. While the LSTM in [8] operates on both, audio and vision features, we pruned the audio branch for our visual-only approach and feed visual features directly to the LSTM. We modified [8] for our purposes in the following way:

- Instead of a many-to-one approach as in [8], where 15 video frames comprising 0.5 seconds are classified to a single VAD label, we use a many-to-many approach where each frame is classified to speech or non-speech depending on all the previous frames. In other words, we do not use a buffer approach where we need to accumulate a number of frames before classifying, but rather we classify in a streaming fashion, on the fly.
- The original approach determines the VAD label L as argmax of the two logits for speech (S) and non-speech (N), which result from the LSTM:

$$\mathrm{L} = \mathrm{argmax}(\mathrm{S}, \mathrm{N})\,. \tag{1}$$

  Since we need a probability $\mathrm{P_{RGB}}$ value for fusing the result of this stream with the optical flow stream, we specify:

$$\mathrm{P_{RGB}} = \mathrm{sigmoid}(\mathrm{J_{RGB}})\,, \tag{2}$$

  where $\mathrm{J_{RGB}}$ is a joint speech/non-speech logit representation which we obtain from the LSTM. We can infer the hard RGB label $\mathrm{L_{RGB}}$ as:

$$\mathrm{L_{RGB}} = \begin{cases} 0 & \text{if } \mathrm{P_{RGB}} < 0.5\,, \\ 1 & \text{otherwise}\,. \end{cases} \tag{3}$$

While the probability $\mathrm{P_{RGB}}$ is required for the whole system during test time, the hard label $\mathrm{L_{RGB}}$ is used during training, in which each stream is trained separately. Our evaluation shows that these modifications improve the training and performance of the RGB stream (see Tab. 1).

Optical flow stream The architecture of the optical flow stream is mirroring that of the RGB stream: we combine optical flow features, extracted using ResNet, with an LSTM module. We performed experiments without the LSTM module, i.e., replacing it with a multi-layer perceptron instead, but the results were worse than with the LSTM included. Hence, we keep the LSTM in the final system.

On the image patch of the lower face region, we compute a dense optical flow via the Gunnar Farneback algorithm [25]. This gives us the optical flow in Cartesian space. We then convert it to polar space and use the amplitude component to create activation maps. Mathematically: assuming we represent an image as $I(x, y, t)$, where x, y, t are the height, width and time dimensions, we can quantify the change of the image over time using Taylor series expansion and truncation of higher order terms as:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0. \tag{4}$$

Further dividing by Δt will give us the velocity, or optical flow components u and v:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0. \tag{5}$$

Each pixel will therefore have an associated (u, v) optical flow in the Cartesian space, which we represent in polar coordinates as:

$$\vec{p} = Ae^{i\phi}, \tag{6}$$

where $\vec{p}$ is the pixel vector associated with (u, v) in Cartesian space, A and ϕ are the corresponding amplitude and angle of the optical flow in polar space. We then take the magnitude component A to create activation maps and feed these into a ResNet-18 for feature extraction. The resulting features are then fed into an LSTM module, which, after applying a sigmoid function, outputs the probability of speech $\mathrm{P_{OF}}$ obtained from the optical flow sequence:

$$\mathrm{P_{OF}} = \mathrm{sigmoid}(\mathrm{J_{OF}})\,, \tag{7}$$

where $\mathrm{J_{OF}}$ is the logit computed by the optical flow LSTM module. To obtain the hard optical flow label $\mathrm{L_{OF}}$ we will threshold just like we did in the RGB case:

$$\mathrm{L_{OF}} = \begin{cases} 0 & \text{if } \mathrm{P_{OF}} < 0.5\,, \\ 1 & \text{otherwise}\,. \end{cases} \tag{8}$$

In analogy to the RGB stream, the hard label $\mathrm{L_{OF}}$ is used for training the optical flow stream separately.

Late fusion For the final fusion stage, we opt for a probabilistic strategy that is implemented only at test time. We compute the average of probabilities of the two streams and then threshold the average at 0.5 to determine the final VAD label:

$$\mathrm{F} = \begin{cases} 0 & \text{if } \frac{\mathrm{P_{RGB}}+\mathrm{P_{OF}}}{2} < 0.5\,, \\ 1 & \text{otherwise}\,, \end{cases} \tag{9}$$

where F is the final label after fusion, $\mathrm{P_{RGB}}$ is the probability of speech as encoded by the RGB branch and $\mathrm{P_{OF}}$ is the probability of speech as encoded by the optical flow branch as defined above.

The intuition behind this approach is that when both streams agree that speech is present, the fused label will be resolved easily by unanimous vote. However, when the networks disagree, we enter a gray zone, and then system certainty is quantified in the form of averaging both probabilities and thresholding to resolve the conflict and come up with the final label. In our system we did not fuse via deep learning, because the current probabilistic fusion allows for

a simple, yet effective explicit control over the probability threshold. This is a useful feature to have if we would like to manually adjust in favor of one metric over another without retraining.

4 Experimental Evaluation

In this section, we outline details about the dataset, the training, and the results we obtain.

4.1 Dataset and Training

For this work we made use of the TCD-TIMIT [9] dataset, which depicts 59 speakers in front of a green-screen and is already split into 42 train, 9 test and 8 validation speakers. Each speaker utters 99 different sentences. We resized the frames from the original 1920x1080 size to 224x224, because using originally-sized images causes landmark detection to run slow, whereas having images which are too small causes it to decrease accuracy.

The original videos were recorded with 30 frames per second (FPS), however we use a digital differential analyzer (DDA) interpolation algorithm [26] to achieve 62 FPS. This has the aim to synchronize the video frames with the audio frames that were used to generate the ground truth labels. Having a higher frequency of the audio frames allowed us to obtain more reliable labels, and hence it was natural to increase also the number of visual frames. Additionally, using a DDA can be thought of as data augmentation. We also augmented during training: we rotated frames by a random angle in the interval of $[-10°, 10°]$ and translated by a random amount in the interval of $[-0.1 \cdot \mathrm{H},\ 0.1 \cdot \mathrm{H}]$ and $[-0.1 \cdot \mathrm{W},\ 0.1 \cdot \mathrm{W}]$ respectively, where H is the image height and W is the image width.

Our approach is audio-visual from the perspective of label computation, i.e., we derive labels automatically from audio, unlike other works which derive labels from subtitle text [19] or create manual labels [8]. Since we have clean audio available, we implemented a sound energy-threshold detector. The alternative for noisy audio could be to use WebRTC [27]. We note that having 10% errors in the ground truth labels, provided they are random, i.e. not systemic, can be overcome by neural net approaches [17], which is another reason in favor of using deep learning for the VAD task.

4.2 Results

In this section, we show how our proposed approach improves the visual component for VAD systems. Our evaluation consists of three parts: first, we show that our modifications of the RGB stream from [8] are useful; second, we show the improvements we gain from adding optical flow to the pipeline; and third, we show that image crops of the lower face region improve the results and are

Table 1. Comparison of baseline model [8] to our modified RGB, optical flow/OF and fused systems. All presented results are for the lower face, i.e. nose/mouth/chin regions extracted from the TCD-TIMIT dataset by our pre-processing routine.

Score type	Baseline RGB method [8]	Our modified RGB stream	Our OF stream	Our Fusion
F1 ↑	87.6	89.4	90.9	**91.7**
TNR ↑	80.1	85.7	85.4	**87.2**
TPR ↑	92.7	93.3	**96.7**	96.5
Balanced Accuracy ↑	86.4	89.5	91.1	**91.9**

additionally more convenient from a practical point of view than precise mouth crops.

Table 1 shows the results for the first two parts of the evaluation. The first two columns of the table compare the original RGB stream from the baseline with our modified RGB stream. We can see some clear improvements, especially with respect to the true negative rate. Note however that the system by Ariav & Cohen [8] is an audio-visual system, in which the focus is on the fusion part of the network. Thus, we do not claim to outperform [8] in general, but we show that optimizing the visual stream alone results in clear performance gains, which will then most likely also be useful for complete audio-visual systems.

In column 4, we show the performance for the optical flow branch alone. We can see that it performs very similar to our RGB branch in terms of true negative rate. The true positive rate however is consistently better for the optical flow branch. When fusing both streams with our probabilistic approach (col. 5), we observe an overall improvement, especially with regard to the true negative rate. This indicates that the streams contain complementary information and profit from each other.

In Table 2, we show that focusing on the lower part of the face, i.e. nose, mouth and chin area does give better results than considering the full face. Moreover, it is better than using mouth-only images. Other works have considered just the mouth region, but including the nose and especially the chin area offers another speech specific cue, since they exhibit much movement for the optical flow branch during speech (note the visualization of the OF activation maps in Fig. 2, which shows clear activation in the chin area). This is also beneficial from a practical point of view for real-world applications, e.g. on a commodity household robot, because generating precisely cropped and well-aligned mouth regions in real time is harder than to approximate lower face crops.

Altogether, we show how the visual branch for voice activity detection can be improved by optimizing the RGB stream, adding optical flow, and focusing on image crops of the lower face region. The performance improves especially with respect to the detection of frames without voice activity. This will most likely also be useful for complete audio-visual systems.

Table 2. Fusion scores on full face, lower face, as well as precise mouth-only input crops. Using lower face images, i.e. nose, mouth and chin area, results in the best performance overall.

Score type	Full face	Mouth-only	Lower face
F1 ↑	90.1	90.4	**91.7**
TNR ↑	86.3	86.5	**87.2**
TPR ↑	94.2	94.6	**96.5**
Balanced Accuracy ↑	90.3	90.6	**91.9**

5 Conclusion

In conclusion, we have presented an approach to improve the true negative rate for voice activity detection using unbalanced speech/non-speech datasets. Our method exploits the complementary advantages of RGB and optical flow and computes static as well as dynamic features from both streams. Additionally, we show that operating on crops of the lower face not only facilitates the processing, but also results in better performance, since especially the chin area contains useful information for VAD. In future work, we will integrate our visual-only VAD into an audio-visual framework.

References

1. P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.
2. E. Verteletskaya and K. Sakhnov, "Voice activity detection for speech enhancement applications," *Acta Polytechnica*, vol. 50, no. 4, 2010.
3. E. Vincent, T. Virtanen, and S. Gannot, eds., *Audio Source Separation and Speech Enhancement*. Hoboken, NJ: John Wiley & Sons, 2018.
4. Q. Liu and W. Wang, "Blind source separation and visual voice activity detection for target speech extraction," in *2011 3rd International Conference on Awareness Science and Technology (iCAST)*, pp. 457–460, IEEE, 2011.
5. J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust speech recognition and understanding*, vol. 6, no. 9, pp. 1–22, 2007.
6. P. Bratoszewski, G. Szwoch, and A. Czyżewski, "Comparison of acoustic and visual voice activity detection for noisy speech recognition," in *2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp. 287–291, IEEE, 2016.
7. F. Tao, J. H. Hansen, and C. Busso, "An unsupervised visual-only voice activity detection approach using temporal orofacial features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
8. I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 265–274, 2019.
9. N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

10. J.-C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on speech and audio processing*, vol. 2, no. 3, pp. 406–412, 1994.
11. J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.
12. T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, 2008.
13. X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2012. Publisher: IEEE.
14. S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *2015 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 121–125, 2015.
15. B. Joosten, E. Postma, and E. Krahmer, "Visual voice activity detection at different speeds," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.
16. P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *European Conference on Computer Vision*, pp. 285–301, Springer, 2016.
17. S. Guy, S. Lathuilière, P. Mesejo, and R. Horaud, "Learning Visual Voice Activity Detection with an Automatically Annotated Dataset," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4851–4856, IEEE, 2021.
18. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *2002 IEEE International conference on acoustics, speech, and signal processing*, vol. 2, pp. II–2017, IEEE, 2002.
19. R. Sharma, K. Somandepalli, and S. Narayanan, "Toward visual voice activity detection for unconstrained videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2991–2995, IEEE, 2019.
20. M. Shahid, C. Beyan, and V. Murino, "S-VVAD: Visual Voice Activity Detection by Motion Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2332–2341, 2021.
21. T. Petsatodis, A. Pnevmatikakis, and C. Boukis, "Voice activity detection using audio-visual information," in *2009 16th International Conference on Digital Signal Processing*, pp. 1–5, IEEE, 2009.
22. F. Tao, J. H. Hansen, and C. Busso, "Improving Boundary Estimation in Audiovisual Speech Activity Detection Using Bayesian Information Criterion.," in *INTERSPEECH*, pp. 2130–2134, 2016.
23. Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 317–326, 2016.
24. D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009. Publisher: JMLR. org.
25. G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*, pp. 363–370, Springer, 2003.
26. J. E. Bresenham, "Algorithm for computer control of a digital plotter," *IBM Systems journal*, vol. 4, no. 1, pp. 25–30, 1965. Publisher: IBM.
27. S. Salishev, A. Barabanov, D. Kocharov, P. Skrelin, and M. Moiseev, "Voice activity detector (VAD) based on long-term mel frequency band features," in *International Conference on Text, Speech, and Dialogue*, pp. 352–358, Springer, 2016.