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

Attention-based cross-modal fusion for audio-visual voice activity detection in musical video streams

Yuanbo Hou[1*], *Zhesong Yu*[2], *Xia Liang*[2], *Xingjian Du*[2], *Bilei Zhu*[2], *Zejun Ma*[2], *Dick Botteldooren*[1]

[1]Ghent University, Belgium
[2]ByteDance AI Lab, China

{yuanbo.hou, Dick.Botteldooren}@UGent.be
{yuzhesong, liangxia.21, duxingjian.real, zhubilei, mazejun}@bytedance.com

Abstract

Many previous audio-visual voice-related works focus on speech, ignoring the singing voice in the growing number of musical video streams on the Internet. For processing diverse musical video data, voice activity detection is a necessary step. This paper attempts to detect the speech and singing voices of target performers in musical video streams using audio-visual information. To integrate information of audio and visual modalities, a multi-branch network is proposed to learn audio and image representations, and the representations are fused by attention based on semantic similarity to shape the acoustic representations through the probability of anchor vocalization. Experiments show the proposed audio-visual multi-branch network far outperforms the audio-only model in challenging acoustic environments, indicating the cross-modal information fusion based on semantic correlation is sensible and successful.
Index Terms: Audio-visual voice activity detection, cross-modal fusion, attention, multimedia signal processing

1. Introduction

With the popularity of musical videos on social platforms, a wide variety of musical videos have been uploaded to the Internet. To recognize speech and singing voices in these videos, voice activity detection (VAD) is a necessary preprocessing to identify the start and end time of human voice activities. VAD has attracted many interests due to its wide applications such as speech [1, 2] and music information processing [3].

In scenes of musical video streams, usually one or more anchors (performers) sing or talk to the audience in front of the camera, while music is being played in the background, containing voices of other people and accompaniments. Besides, there may be other sounds such as applause, cheers, and screams from audiences. This paper aims to detect the singing voice and speech of an anchor (the target performer) in musical videos, which have a challenging acoustic environment like a cocktail party. In such noisy environments, audio-only VAD methods [4, 5, 6, 7] are difficult to work accurately and effectively because the sound of various musical instruments, cheers and applause from audiences, and other non-target sound signals, will interfere with the audio-only VAD. Besides, voices from other non-target people in the background music can be mistaken as active voices. It is difficult to distinguish the target voice from various non-target sounds using only audio information, and it is more difficult to further identify the target singing voice and speech when they are embedded in transient interferences and highly non-stationary noises [8] from the background. That is, audio-only VAD methods do not work well in musical videos.

The visual information from video is more robust than the audio information from noisy acoustic environments. The facial information of an anchor can directly reflect whether the anchor is vocalizing. So, audio-visual VAD (AVVAD) is introduced to utilize the visual modality to make up for deficiencies of audio-only VAD in complex acoustic environments. Visual features are used to identify natures of lips in AVVAD [9] for speech processing. By analyzing lip shapes during speech and non-speech, an appropriate visual parameter [10] is used for detecting sections of voice activity in speech embedded in non-stationary noise. Even in clean acoustic conditions using visual channels in addition to speech results in significantly improved classification performance [11]. However, the above audio-visual works mainly focus on speech but not on general sounds such as music and singing voice. This paper aims to detect the anchor's speech and singing voice in musical video streams, which is more challenging because there are not only a lot of speech-like noises but also other people's voices in audio streams.

To ignore the interference of acoustic noise and pay attention to detect voices of the anchor (target performer) in musical videos, this paper uses visual information that is not affected by acoustic noises to assist the model to more accurately judge voices source. Therefore, how to fuse information between two modalities to achieve a better combination effect is the core challenge of this work. Audio-visual integration strategies in previous works can be divided into three categories: feature fusion (*FF*) [12], decision fusion (*DF*) [13], intermediate fusion (*IF*) [14]. *FF* is simple splicing of audio and visual features to form a new feature set and modeling it. Based on the modeling of audio and image stream respectively, *DF* controls the final decision result by stream weight. *IF* attempts to model the fusion of intermediate representations of audio and visual features. Compared with *IF*, *DF* cannot take advantage of the temporal and semantic correlation between audio and visual features. To exploit the correlation between audio and visual features in musical videos, this paper uses *IF* to integrate audio and visual vectors and make comprehensive decisions, that is, high-level representations of acoustic and image features are fused with the attention mechanism based on the semantic similarity.

In this paper, the correlation in the semantic space between the voice representations and the anchor vocalization representations is used to determine whether the voice comes from the anchor. The voice representations are adjusted according to the corresponding correlation coefficient based on attention. This paper attempts to explore the possibility of cross-modal fusion based on semantic similarity between different modalities to help the model actively learn how to fuse cross-modal information, let the model decide "how" to combine given multi-modal information most optimally, rather than based on artificial rules.

[*] Work performed as an intern at Bytedance AI Lab.

http://dx.doi.org/10.21437/Interspeech.2021-37

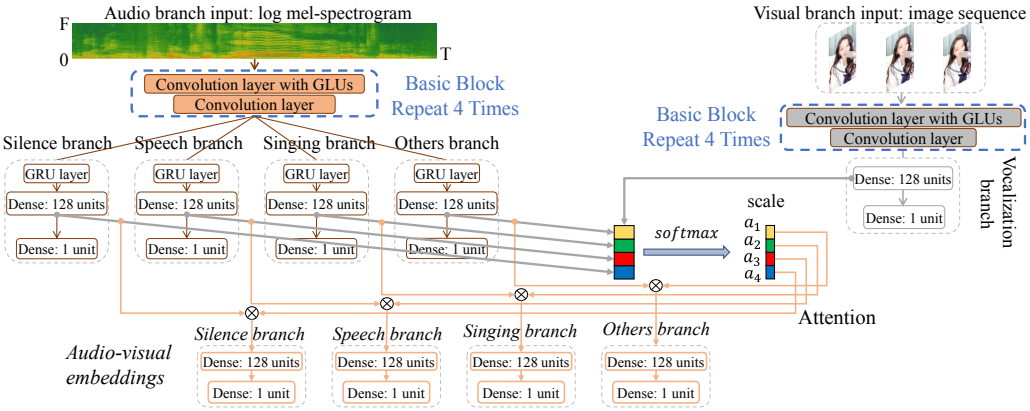


Figure 1: *The proposed attention-based AVVAD (ATT-AVVAD) framework.*

The main contributions of this paper are: 1) a multi-branch network is proposed for AVVAD to learn the high-level representations of different target events, and fuse the cross-modal information based on the semantic correlation by attention; 2) the possibility of detecting both the speech and singing voice of the target performer in challenging noisy acoustic environments is explored; 3) the intermediate representations of the proposed AVVAD model are visually analyzed to further investigate the performance of the model. This paper is organized as follows, Section 2 shows the attention-based audio-visual framework. Section 3 describes the dataset, baseline, experimental setup, and analyzes the results. Section 4 gives conclusions.

2. Multi-branch ATT-AVVAD framework

The proposed attention-based AVVAD (ATT-AVVAD) framework in Figure 1 consists of the audio-based module, image-based module, and attention-based fusion module. The audio-based module produces acoustic representation vectors for four target audio events: Speech of the anchor, Singing voice of the anchor, Silence, and Others. The image-based module aims to obtain the possibility of anchor vocalization based on facial parameters. Finally, an attention-based module fuses audio-visual information to comprehensively consider the bi-modal information to make final decisions at the audio-visual level.

2.1. The audio-based module (audio branch)

The goal of the audio-based module is to predict the probability of four target event classes (Silence, Speech, Singing, Others) at the audio level, wherein Singing and Speech only refer to the singing voice and speech of the anchor, rather than those from other people. Different from typical sound event detection (SED) [15] models, the audio branch in Figure 1 attempts to generate high-level core acoustic representations for each target event class at the end of each output. In Figure 1, a multi-output convolutional recurrent neural network (CRNN) is used to extract the core representations of different event classes. The log mel-spectrogram [16] is extracted from the audio clip and input into the network. Then there are four blocks and each block contains gated linear units (GLUs) [17], a convolutional layer, a batch normalization layer [18], and a ReLU [19]. GLUs can be used to effectively learn local shift-invariant patterns and acoustic representations of target events from the spectrogram [20].

To obtain a separate acoustic representation vector for each target event, after the blocks there are four independent embedding layers to extract core representation. Each embedding layer includes a GRU layer to capture temporal information, followed by two fully connected layers. The output of the first fully connected layer is regarded as the core representation vector of the corresponding event, and will be used to combine with the visual embedding vector. The second fully connected layer with one unit is a binary classification with sigmoid [21] to predict the probability of the corresponding event in the current audio branch input. Please visit the source code on our homepage (https://github.com/Yuanbo2020/Attention-based-AV-VAD) for the specific parameters and real video detection demos.

2.2. The image-based module (visual branch)

In musical videos, there are both a lot of speech-like instrumental accompaniment sounds and other people's voices. These interferences result in the poor performance of relying on the audio signal to detect the anchor's speech and singing voice because it is difficult to determine the source of voices with audio alone. Hence, it is necessary to combine the anchor's visual information, such as the change of eye and lip contours, to distinguish whether voices are from the anchor or other people.

The visual branch aims to obtain the core representation vector of the anchor vocalization and to assist in judging whether voices are from the anchor. To obtain the probability of the anchor vocalization, a fixed-length image sequence corresponding to the input time of the audio branch is input to the network. The visual branch and audio branch have a similar structure with different parameters. Experiments [22] show convolutional layers are more effective than GRU layers in image feature extraction, so there are no GRU layers in the visual branch. Similar to the structure of the audio branch, there are two fully connected layers after four blocks in the visual branch. The output of the dense layer with 128 units is regarded as the core representation vector of the anchor vocalization, the final classification layer with one unit is a binary classification with sigmoid to predict the probability of the anchor vocalization.

2.3. Attention-based fusion module

In videos, visual events usually occur together with acoustic events and they are coordinated. The facial information can reflect whether the anchor is vocalizing or not. To explore the correlation between the anchor's voice activity and face parameters, this paper attempts to fuse the audio and visual high-level representation vectors based on the semantic similarity by attention mechanism to make comprehensive decisions for detecting the four target event classes at the audio-visual level.

In Figure 1, the subbranches in the audio branch output the high-level representation vector of Silence, Speech, Singing, and Others, respectively. The visual branch outputs the representation vector to represent the anchor vocalization. By the

attention module, the correlation between acoustic representation vectors and visual vocalization vector in the semantic space is calculated to strengthen audio-level representations, then the network can pay more attention to sound events related to the anchor. Given $\{\boldsymbol{q_1}, \boldsymbol{q_2}, \boldsymbol{q_3}, \boldsymbol{q_4}\} \in \mathbb{R}^{128\times1}$ denote the acoustic embedding vector of Silence, Speech, Singing and Others in the audio branch, respectively. $\boldsymbol{Q} = [\boldsymbol{q_1}, \boldsymbol{q_2}, \boldsymbol{q_3}, \boldsymbol{q_4}]$. $\boldsymbol{K} \in \mathbb{R}^{128\times1}$ denotes the visual vocalization vector. The attention (ATT) [23] can be defined as:

$$ATT = Softmax(Q^T K/\sqrt{d_k}) \tag{1}$$

where d_k is the dimension of $\boldsymbol{K}$. ATT is a 4×1 vector containing the scaling factors to be applied to the representation vector of the corresponding acoustic event. After multiplying the acoustic event vector with the corresponding attention scale factor, the audio-visual event vector is obtained.

Figure 2 shows the core idea of the attention-based fusion, when the audio branch detects the speech or singing voice, and the visual branch predicts the anchor is vocalizing at this time, the corresponding acoustic event vector will be given more attention and transmitted to the audio-visual module. With the scaling effect of the attention factor, we try to train the model to find such a relationship: when the audio branch detects speech or singing voice, and the visual branch indicates the anchor is vocalizing, then representations of the speech or singing voice will be relatively enhanced, that is, representations of the audio branch is confirmed in the audio-visual module. Representations given by the audio branch are corrected, and errors of the audio-visual module are reduced. That is, only when representations of the audio and visual branch are consistent in the semantic space, the corresponding acoustic representations will be relatively enhanced, while the importance of other acoustic event representations will be relatively weakened.

To consider the information of audio, visual, and audio-visual modality at the same time in the training phase, the losses of different modules are calculated together. The final loss function of the ATT-AVVAD model is:

$$\begin{aligned} L =& \lambda_1 L_{a-sil} + \lambda_2 L_{a-spe} + \lambda_3 L_{a-sin} + \lambda_4 L_{a-oth} \\ &+ \lambda_5 L_{v-voc} + \lambda_6 L_{av-sil} + \\ &\lambda_7 L_{av-spe} + \lambda_8 L_{av-sin} + \lambda_9 L_{av-oth} \end{aligned} \tag{2}$$

where L_a, L_v and L_{av} denote the loss of audio branch, visual branch and audio-visual module; sil, spe, sin and oth denote silence, speech, singing and others; voc denotes vocalizing. λ_i is the scale factor of each loss function, the size of λ_i determines the importance of each loss function in training.

3. Experiments and results

3.1. Dataset, Baseline, and Evaluation metrics

To train the ATT-AVVAD model to detect both target speech and singing voice in challenging acoustic environments, a 500-minute video dataset with frame-level labels is used. The duration of the dataset for training, validation, and testing is 360 mins, 40 mins, and 100 mins, respectively. To prevent the model bias caused by the unbalanced number of male and female anchors in the training phase, the total duration of live broadcasts of male and female anchors is close in the dataset.

In training, log mel-bank energy with 64 banks [24] of the input audio stream is used in the audio branch, which is extracted by STFT with Hamming window length of 44 ms and overlap of 50% between the window. To comprehensively consider the contextual information, the input of audio branch is a moving feature block whose time length is consistent with

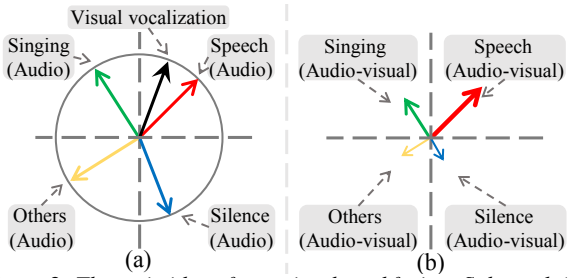


Figure 2: *The main idea of attention-based fusion. Subgraph (a) shows the audio and visual representation vectors in the same semantic space. Subgraph (b) is the final audio-visual representation vector of target events after the attention-based fusion.*

that represented by the image sequence of the visual branch. ATT-AVVAD model is expected to be used online, a short latency is required. So a lightweight face detection algorithm [25] is used to pre-mark the face in visual inputs, to reduce the computational burden of the model and help it to find focal areas faster and more accurately. The output of each branch in ATT-AVVAD is binary classification, hence Adam optimizer [26] with a learning rate of 0.001 is used to minimize the binary cross-entropy. Dropout [27] is used to prevent overfitting.

To compare the performance of our proposed method comprehensively, two baseline systems are considered. A common and typical multi-modal recurrent neural model [28] with two branches similar to the proposed ATT-AVVAD is used as the audio-visual baseline (*Base-AV*). A compound convolutional recurrent neural network [3] trained by transfer learning is used as the audio-only baseline (*Base-A*) to compare the performance of the ATT-AVVAD from more perspectives.

For evaluation metrics, event-based precision (*P*), recall (*R*), *F-score* and Error rate (*ER*) [29] are used to analyze the performance of the model. Compared with segment-based metrics used in previous studies [30], event-based metrics are more rigorous and accurate to measure the location of events. Higher *P*, *R*, *F* and lower *ER* indicate a better performance.

3.2. Results and Analysis

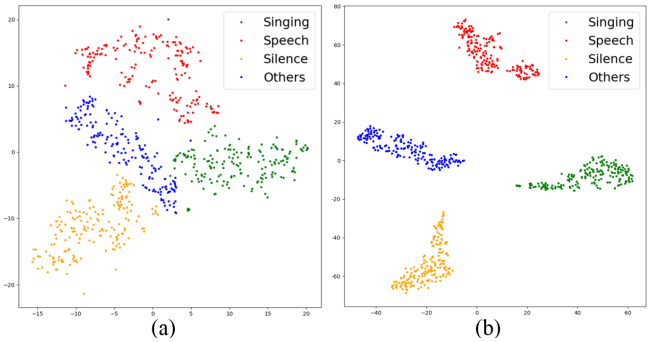
This section tries to analyze the performance of the proposed method based on the following **R**esearch **Q**uestions (**RQ**):

• **RQ1**: λ_i in the final loss function determines the importance of each loss function in training [31]. What effect do different values of λ_i have on the performance of the model?

The four classes of target events in our task are equally important, so λ_i related to the event sub-branch in the audio branch are the same in Table 1. The same goes for audio-visual branch. In Table 1, given λ_1 is 1 and λ_5 is 0.5, it is equivalent to attaching importance to the audio silence branch and reducing the importance of the visual vocalization branch. Different values of λ_i represent the difference in importance between audio information, visual information, and audio-visual information. Table 1 shows that for the VAD task, the joint audio-visual information is more important than the audio information, and the audio information is more important than the visual information.

Table 1: *Results of different values of λ_i on the test dataset.*

$\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$	$\{\lambda_5\}$	$\{\lambda_6, \lambda_7, \lambda_8, \lambda_9\}$	*F-score (%)*
{*1, 1, 1, 1*}	{1}	{0.5, 0.5, 0.5, 0.5}	76.38
{*1, 1, 1, 1*}	{0.5}	{1, 1, 1, 1}	77.01
{*0.5, 0.5, 0.5, 0.5*}	{1}	{1, 1, 1, 1}	76.73
{*1, 1, 1, 1*}	{1}	{1, 1, 1, 1}	77.81
{*0.5, 0.5, 0.5, 0.5*}	{0.5}	{1, 1, 1, 1}	77.90
{*0.5, 0.5, 0.5, 0.5*}	{0.5}	{0.5, 0.5, 0.5, 0.5}	77.12



(a) (b)

Figure 3: *Visualization of core representation vectors distribution from a test sample using t-SNE [32]. The vectors in subgraph (a) are from the audio branch, vectors in subgraph (b) are from audio-visual modules after attention-based fusion.*

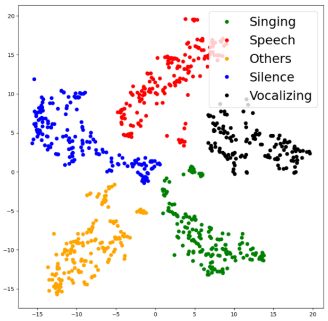


Figure 4: *Visualization of acoustic representation vectors and visual vocalization vector distribution from a test sample using t-SNE. The vector (black dots) representing the vocalizing of the anchor is distributed on the side representing the voices of the anchor (green dots for singing, red dots for speech).*

• **RQ2**: Does the proposed ATT-AVVAD in this paper perform better than the audio-visual baseline *Base-AV*? Are the detection results of ATT-AVVAD better than audio-only VAD baseline *Base-A*, and how much improvement is there?

To compare the performance of the proposed method, Table 2 shows the detailed results of the proposed ATT-AVVAD, *Base-AV* and *Base-A*. The last classification layer of *Base-AV* uses the Softmax activation function to make decisions, which means that each event class in the classification layer is equally important. As mentioned before, λ_i is the scale factor of each loss function. For a fair comparison between the proposed ATT-AVVAD model and audio-visual baseline *Base-AV*, all λ_i in L are equal to 1 in the following results.

The results in Table 2 show that: 1) in challenging noisy environments like musical video, the performance of anchor's voices detection based on audio-visual information (*Base-AV* and *ATT-AVVAD*) far outperforms that of the audio only approach (*Base-A*). Compared with the audio-only *Base-A* with an *ER* of 0.86 and *F-score* of 40.93%, the *ER* and *F-score* of the ATT-AVVAD in this paper are 0.39 and 77.81%. 2) Even though both are based on audio-visual information, the cross-modal learning method based on attention fusion proposed in this paper has a lower *ER* and more accurate detection results than the typical audio-visual model *Base-AV*, which means the bi-modal framework proposed in this paper is effective, and the attention-based fusion mechanism is helpful.

• **RQ3**: What changes have taken place in the model learning before and after the attention-based fusion?

To gain deeper insights into the effect of attention-based fusion on the model, the distribution of the core representation vector representing four target event classes before and after the fusion is visualized. The representation vectors of target events in the audio branch can reflect the decision tendency of the model before the fusion, and the corresponding audio-visual representation vectors can reflect modified results of the model after the attention fusion. As shown in Figure 3, before fusion, the model can roughly divide different target classes in the audio branch, but the classification boundary interval between each class is not obvious, and each cluster is not compact. After focusing on the visual vector based on attention, there is a clear classification boundary between different classes, and each class is more tightly clustered. This indicates the proposed attention-based fusion does play a regulating role and makes the final joint classification in the audio-visual module easier. Based on attention fusion, the distribution of acoustic core representation vectors and visual vocalization vector in semantic space tends to be aligned as shown in Figure 4. The visual vocalization vector is distributed on the side of the vectors of speech and singing voice, and is away from the event vectors without the anchor's voices, which means the semantics of audio and visual vectors are consistent and the cross-modal information fusion based on the correlation between acoustic embeddings and visual vocalization vectors is reasonable.

Table 2: *Event-based evaluation of detection results.*

	ER	*P (%)*	*R (%)*	*F-score (%)*
Base-A	0.86	47.93	35.72	40.93
Base-AV	0.72	66.17	53.41	59.11
ATT-AVVAD	**0.39**	**85.24**	**71.57**	**77.81**

The bi-modal ATT-AVVAD is effective, but how big is the effect of different modal branches? The ablation studies [33] in Table 3 show the results of ATT-AVVAD after removing certain structures. The detection result based on pure visual information is the worst, perhaps because the opening and closing of the mouth of the anchor are binary, it is difficult to use the binary information to detect four types of target events. Compared with the audio-only detection results, the bi-modal detection results have been improved, indicating that it is useful to use visual information to correct audio information to assist judgment.

4. Conclusion

To detect the speech and singing voice of the anchor in musical video streams, a multi-branch ATT-AVVAD framework is proposed with attention-based fusion by semantic similarity, which performs well in noisy environments. Experiments show that: 1) the performance of the audio-visual model far outperforms that of the audio-only model in challenging acoustic environments; 2) the multi-branch network that can produce core representation vectors for target events, and attention-based fusion are effective; 3) the cross-modal information fusion based on semantic similarity is successful.

5. Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

Table 3: *Ablation experiments of the ATT-AVVAD model.*

Audio module	*Visual module*	*Fusion module*	*F-score (%)*
✔	✗	✗	68.59
✗	✔	✗	33.92
✔	✔	✔	77.81

6. References

[1] M. W. Hoffman, Z. Li, and D. Khataniar, "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 175–178, 2001.

[2] J. Ramírez, J. Segura, J. M. Górriz, and L. García, "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.

[3] Y. Hou, F. K. Soong, J. Luan, and S. Li, "Transfer learning for improving singing-voice detection in polyphonic instrumental music," in *Proceedings of INTERSPEECH*, 2020, pp. 1236–1240.

[4] 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.

[5] A. L. Berenzweig and D. P. Ellis, "Locating singing voice segments within music signals," in *Proceedings of IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 119–122.

[6] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[7] K. Lee, K. Choi, and J. Nam, "Revisiting singing voice detection: A quantitative review and the future outlook," in *Proceedings of International Society for Music Information Retrieval (ISMIR)*, 2018, pp. 506–513.

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

[9] P. Liu and Z. Wang, "Audio-visual voice activity detection," *Frontiers of Electrical and Electronic Engineering in China*, vol. 1, no. 4, pp. 425–430, 2006.

[10] D. Sodoyer, B. Rivet, L. Girin, J. . Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006, pp. I–I.

[11] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2130–2134.

[12] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7596–7599.

[13] P. Teissier, J. Robert-Ribes, J.-L. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 629–642, 1999.

[14] S. Receveur, R. Weiß, and T. Fingscheidt, "Turbo automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 846–862, 2016.

[15] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 121–125.

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[17] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70. PMLR, 2017, pp. 933–941.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37. PMLR, 2015, pp. 448–456.

[19] K. Eckle and J. Schmidt-Hieber, "A comparison of deep networks with relu activation function and linear spline-type methods," *Neural Networks*, vol. 110, pp. 232–242, 2019.

[20] Y. Hou, Q. Kong, S. Li, and M. D. Plumbley, "Sound event detection with sequentially labelled data based on connectionist temporal classification and unsupervised clustering," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 46–50.

[21] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International Workshop on Artificial Neural Networks*. Springer, 1995, pp. 195–201.

[22] N. Singhal, N. Singhal, and Srishti, "Comparing cnn and rnn for prediction of judgement in video interview based on facial gestures," in *5th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2018, pp. 175–180.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[24] A. Bala, A. Kumar, and N. Birla, "Voice command recognition system based on mfcc and dtw," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335–7342, 2010.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of 3rd International Conference on Learning Representations (ICLR)*, 2015.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[28] F. Tao and C. Busso, "End-to-end audiovisual speech activity detection with bimodal recurrent neural models," *Speech Communication*, vol. 113, pp. 25–35, 2019.

[29] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[30] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[31] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[32] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[33] C. Fawcett and H. H. Hoos, "Analysing differences between algorithm configurations through ablation," *Journal of Heuristics*, vol. 22, no. 4, pp. 431–458, 2016.