



Noise-Tolerant Self-Supervised Learning for Audio-Visual Voice Activity Detection

Ui-Hyun Kim

Corporate Research and Development Center, Toshiba Corporation, Japan

uihyun.kim@toshiba.co.jp

Abstract

Recent audio-visual voice activity detectors based on supervised learning require large amounts of labeled training data with manual mouth-region cropping in videos, and the performance is sensitive to a mismatch between the training and testing noise conditions. This paper introduces contrastive self-supervised learning for audio-visual voice activity detection as a possible solution to such problems. In addition, a novel self-supervised learning framework is proposed to improve overall training efficiency and testing performance on noise-corrupted datasets, as in real-world scenarios. This framework includes a branched audio encoder and a noise-tolerant loss function to cope with the uncertainty of speech and noise feature separation in a self-supervised manner. Experimental results, particularly under mismatched noise conditions, demonstrate the improved performance compared with a self-supervised learning baseline and a supervised learning framework.

Index Terms: audio-visual, voice activity detection, noise-tolerant, contrastive learning, self-supervised learning

1. Introduction

Voice activity detection (VAD) refers to the preprocessing task of determining whether a signal contains human speech during speech processing, such as in automatic speech recognition, speaker verification, and speech coding [1]. Conventional approaches are based on energy level detection [2], zero-crossing rate [3], periodicity estimation [4], cepstral distance [5], and spectral entropy [6] with digital signal processing techniques in the time and frequency domains. Approaches using statistical models, such as Gaussian mixture models and hidden Markov models, have been widely adopted to date and studied to improve performance, especially in noisy conditions with a low signal-to-noise ratio (SNR) [7–11]. Recently, deep learning approaches have been successfully applied and have outperformed earlier approaches. They offer better modeling capabilities and improved detection with fully connected deep neural networks [12, 13], convolutional neural networks [14, 15], recurrent neural networks [16], long short-term memory networks [17], and attention-based neural networks [18].

However, VAD remains a challenging task in noisy real-world environments, which commonly include highly nonstationary noise and various transient interferences. Several deep learning approaches cope with this problem by using other modalities, such as visual information, which is invariant in the acoustic environment, but they inevitably require preparatory work of manual mouth-region cropping to better learn visual features [19–21]. In addition, they also have the drawbacks of a supervised method: large amounts of

labeled training data are required, and the performance is sensitive to a mismatch between the training and testing conditions [22, 23]. As potential studies that can overcome these drawbacks, contrastive self-supervised learning frameworks based on the similarity between acoustic and visual representations have been proposed for sound source localization, source separation, and active speaker detection in recent years [24–28].

Although the VAD task has traditionally been developed and evaluated to be robust to acoustic noise environments, recently, neural network models based on self-supervised learning are being trained and tested on clean datasets for their powerful representations and best results. However, this paper deals with contrastive self-supervised learning for audio-visual VAD (AV-VAD), which was conducted on a dataset corrupted by acoustic noise, as in a real-world scenario. The main contributions of this work are summarized as follows:

- A contrastive self-supervised learning baseline for AV-VAD is introduced.
- A novel self-supervised learning framework, which includes a branched audio encoder and a noise-tolerant loss function, is proposed for overall training efficiency and testing performance improvement.
- Experimental performance comparison between supervised, baseline, and proposed learning frameworks, which were trained on noise-free or noise-corrupted datasets and evaluated in matched or mismatched noise conditions, is provided.

2. Baseline Method

In videos, images and their corresponding sounds are synchronized, and the visible mouth movements and audible speech of a speaker in a video are closely correlated with each other. Based on this fact, for AV-VAD, we assume that the higher the correlation between audio and video embedding features, the higher the probability that speech is present. In other words, the absence of speech or mouth movements means that either or both of audio and video embedding features are weak, resulting in a low correlation.

2.1. Overall architecture

The baseline neural network architecture for AV-VAD, depicted in Fig. 1, includes an audio encoder consisting of five 1D convolutional layers and a video encoder consisting of four 3D convolutional layers with one 3D max-pooling layer. Every convolutional layer is followed by batch normalization [29] and rectified linear unit activation [30]. Given an input video that has monaural sound, the audio encoder maps a log-compressed Mel spectrogram $A \in \mathbb{R}^{T \times F}$, which has the number of frames T and the number of Mel-frequency bins F ,

The number of parameters: 1,245,792

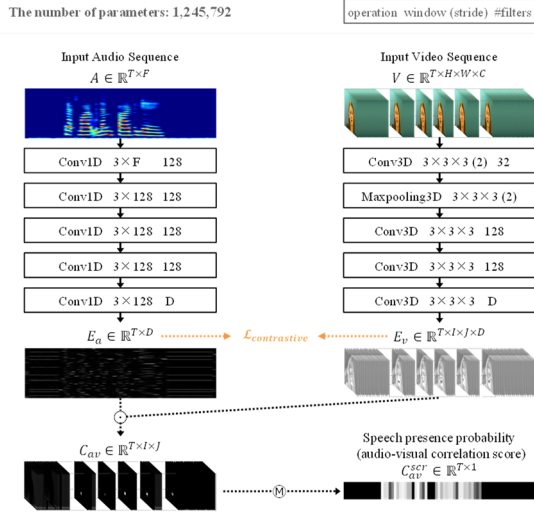


Figure 1: Baseline neural network architecture.

into a D -dimensional latent representation $E_a \in \mathbb{R}^{T \times D}$ for every frame of the input audio sequence. In a similar way, the video encoder maps a size-scaled video $V \in \mathbb{R}^{T \times H \times W \times C}$, which has a height of H pixels, width of W pixels, and C channels (e.g., 3 channels for RGB colors), into a D -dimensional latent representation $E_v \in \mathbb{R}^{T \times I \times J \times D}$ with spatially embedded dimensions $I \times J$ for every frame of the input video sequence. These two latent representations are used as an audio embedding vector $E_a(t)$ with a frame index $t = (1, \dots, T)$ and a video embedding vector $E_v(t, i, j)$ with spatial dimension indexes $i = (1, \dots, I)$ and $j = (1, \dots, J)$ to exploit the inherent correlation between the audio and visual modalities.

2.2. Cross-modal contrastive learning

The time-spatial correlation between audio and video embedding vectors is calculated with cosine similarity as follows:

$$C_{av}(t, i, j) = \frac{E_a(t)^T \cdot E_v(t, i, j)}{\|E_a(t)\| \cdot \|E_v(t, i, j)\|}, \quad (1)$$

where C_{sv} is commonly regarded as the AV attention map [24–28], which represents the (i, j) spatial coordinate of the video embedding map for every corresponding audio frame t . To make the model learn fine-grained correlations, we also apply the correlation scoring function based on the maximum spatial response following common practice in multi-instance learning [27]:

$$C_{av}^{scr}(t) = \max_{i,j} C_{av}(t, i, j). \quad (2)$$

We aim to maximize the correlation between audio and video embedding vectors from time-synchronized inputs as the probability of speech presence while minimizing their correlation from unsynchronized inputs for the absence of speech or mouth movements. To achieve this, we treat AV synchronized inputs as positive pairs, while unsynchronized inputs are treated as negative pairs for the cross-modal contrastive loss function as follows [28]:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(C_{av}^{scr}(A, V))}{\exp(C_{av}^{scr}(A, V)) + \sum_{u=1}^U \exp(C_{av}^{scr}(A_u^{unsync}, V))}, \quad (3)$$

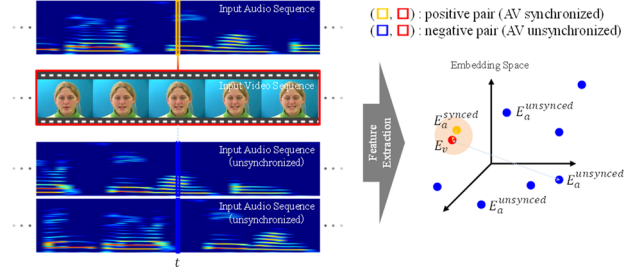


Figure 2: Training pair sampling scheme with positive and negative pairs illustrated.

where A_u^{unsync} denotes U other input audio sequences augmented by shifting the corresponding audio frames, which are unsynchronized with the input video sequence. As shown in Fig. 2, the cross-modal contrastive loss encourages intra-similarity between synchronized inputs of audio and video in embedding spaces in a self-supervised manner.

2.3. Speech presence probability estimation

We consider the AV correlation score from Eq. (2) to be the predicted probability of speech presence in the trained model. Each frame is determined to be “speech-present” or “speech-absent” by using a decision procedure [31] with a threshold η :

$$\begin{aligned} \text{if } C_{av}^{scr}(t) > \eta & \text{ then } t = \text{speech present,} \\ \text{else} & \text{ } t = \text{speech absent.} \end{aligned} \quad (4)$$

3. Proposed Method

The baseline method predicts the probability of speech presence from the model trained on the inherent relationship between the audio and visual modalities by solving a self-supervised task based on AV synchronization that depends on the simultaneous occurrence of audio and video embedding features. This means that the baseline method cannot extract good embedding features that are generally helpful for AV-VAD tasks under continuous noise conditions, and that training on noisy datasets causes significant performance degradation. VAD tasks have been traditionally studied and evaluated in acoustic noise environments. To cope with this problem, we propose an audio encoder branched into two convolutional layers to extract speech and noise embedding features separately and a loss function to solve the uncertainty of speech and noise feature separation in a self-supervised manner.

3.1. Branched audio encoder

The neural network architecture proposed to cope with noisy input conditions is depicted in Fig. 3. The main difference compared with the baseline neural network architecture is that the last convolutional layer of the audio encoder branches into two convolutional layers to allow for extracting speech and noise embedding features individually from a noisy input audio sequence. We configure this branched audio encoder to have the same number of parameters as the baseline neural network architecture for fair comparison and evaluation. The proposed audio encoder encourages the extraction of a D -dimensional speech embedding vector $E_s(t)$ and a D -dimensional noise embedding vector $E_n(t)$ separately as a speech separation task.

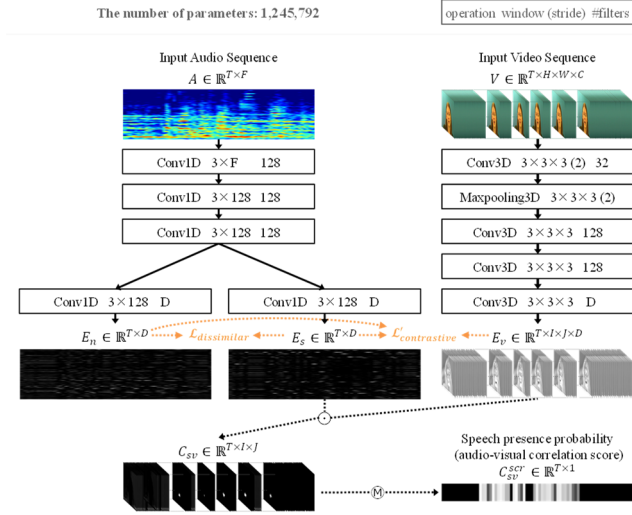


Figure 3: Proposed neural network architecture.

3.2. Noise-tolerant contrastive learning

The key idea for training a robust model that can cope with a noise-corrupted dataset depends on solving the uncertainty of speech and noise feature separation by the branched audio encoder in a self-supervised manner. To achieve this, we first use the correlation scoring function between noise and video embedding vectors from time-synchronized inputs as an additional negative pair for the cross-modal contrastive learning (Eq. (3)). In the same form as Eqs. (1) and (2), correlation scoring functions between speech, noise, and video embedding vectors can be written using cosine similarity and the maximum spatial response as follows:

$$C_{sv}^{scr}(t) = \max_{i,j} \left(\frac{E_s(t)^T \cdot E_v(t,i,j)}{\|E_s(t)\| \cdot \|E_v(t,i,j)\|} \right), \quad (5)$$

$$C_{nv}^{scr}(t) = \max_{i,j} \left(\frac{E_n(t)^T \cdot E_v(t,i,j)}{\|E_n(t)\| \cdot \|E_v(t,i,j)\|} \right), \quad (6)$$

where $C_{sv}^{scr}(t)$ denotes the correlation score between speech and video embedding vectors and $C_{nv}^{scr}(t)$ denotes the correlation score between noise and video embedding vectors, for every corresponding time step t . Then, the proposed contrastive loss function is defined as

$$\mathcal{L}'_{contrastive} = -\log \frac{\exp(C_{sv}^{scr}(A,V))}{\exp(C_{sv}^{scr}(A,V)) + \sum_{u=1}^U \exp(C_{sv}^{scr}(A_u^{unsync}, V)) + \exp(C_{nv}^{scr}(A,V))}, \quad (7)$$

where the correlation between noise and video embedding vectors from time-synchronized inputs are encouraged to be minimized, under the assumption that noise is less correlated to key visual information, such as mouth movement. We next apply a dissimilar loss function to enable the branched audio encoder to reliably distinguish and extract speech and noise embedding features:

$$\mathcal{L}_{dissimilar} = -\log \frac{1}{1 + \exp\left(\frac{E_s(t)^T \cdot E_n(t)}{\|E_s(t)\| \cdot \|E_n(t)\|}\right)}. \quad (8)$$

The proposed dissimilar loss function encourages the minimization of the inter-similarity between the speech and noise embedding features extracted from noisy inputs in a self-supervised manner based on speech and noise having different energy distributions over the frequency range [32, 33]. Third, we jointly train the model using two proposed loss functions

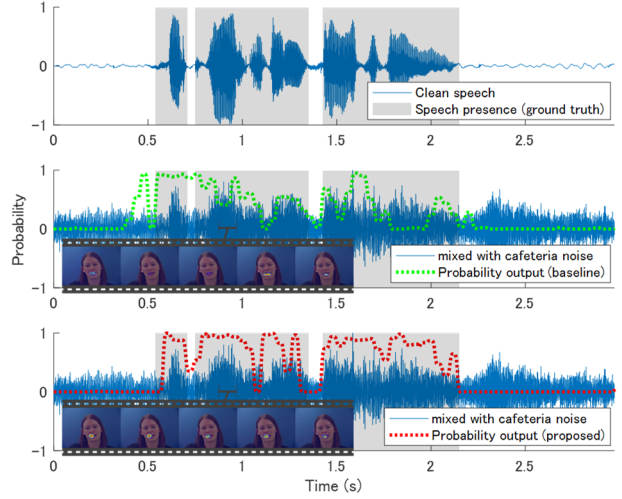


Figure 4: Speech presence probability estimation from *swwm5a.mpg* (GRID). (Top) with visualized speech occurrence (gray boxes); (Center and Bottom) mixed with *BGD_150203_010_CAF.CH3.wav* (CHiME4) at a SNR of -5 dB and heat map visualization of audio-visual attention maps for input video frames.

as multi-objective optimization [34]. The noise-tolerant loss function is defined as follows:

$$\mathcal{L}_{noise-tolerant} = \mathcal{L}'_{contrastive} + \mathcal{L}_{dissimilar}. \quad (9)$$

For the AV-VAD decision procedure in the proposed method, C_{sv}^{scr} is used instead of C_{av}^{scr} in Eq. (4).

4. Experiments

4.1. Supervised learning

The proposed method is based on self-supervised learning with completely arbitrary unlabeled data that are not related to the problem of distinguishing speech segments from background noise. To compare the performance with the supervised learning method, the baseline neural network architecture was used with a simple modification for the supervised learning settings. By adding a supervised loss component instead of the cross-modal contrastive loss function (Eq. (3)), the baseline neural network architecture can be converted to an architecture that can work as supervised learning:

$$\mathcal{L}_{supervised} = \sum_{t=1}^T (y(t) - C_{av}^{scr}(t))^2, \quad (10)$$

where $y(t) \in [0, 1]$ is the label. We assigned 1 to an input audio frame containing speech and 0 to a frame containing silence or only noise. The label was produced by applying a common VAD [9] to corresponding clean speech data and the ground truth for the evaluation was produced in the same way.

4.2. Datasets

The supervised model, the baseline model, and the proposed model were trained and evaluated on the GRID AV sentence corpus [35], which consists of time-synchronized AV recordings of 1000 sentences spoken by each of 34 speakers (18 males and 16 females). To verify the performance degradation by training on a noise-corrupted dataset, background noise provided by the 4th CHiME challenge [36] was randomly selected and mixed into audio recordings of the

Table 1: Area under the receiver operating characteristic curve for each evaluation condition with the GRID dataset. The best results are highlighted in bold, and the second-best results are underlined.

METHOD	Noise Condition : matched / mismatched		Signal-to-Noise Ratio							Relative Improvement	
	Pre-training	Evaluation	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Average	Baseline	Supervis.
Supervised	none	East Anglia	91.92%	88.61%	83.66%	76.91%	68.32%	60.92%	78.39%		
Baseline	none	East Anglia	84.64%	84.45%	83.33%	<u>83.34%</u>	<u>80.60%</u>	<u>78.32%</u>	<u>82.45%</u>		
Proposed	none	East Anglia	<u>86.62%</u>	<u>86.43%</u>	<u>86.32%</u>	83.56%	81.84%	78.36%	83.86%	1.71%	6.97%
Supervised	CHiME4	East Anglia	95.98%	94.20%	90.62%	82.25%	72.30%	64.27%	83.27%		
Baseline	CHiME4	East Anglia	76.31%	74.32%	73.40%	72.97%	<u>72.78%</u>	<u>72.59%</u>	73.73%		
Proposed	CHiME4	East Anglia	<u>87.83%</u>	<u>87.77%</u>	<u>87.57%</u>	85.20%	81.51%	78.14%	84.67%	14.84%	1.68%
Supervised	CHiME4	CHiME4	97.98%	97.76%	97.32%	96.57%	94.45%	89.42%	95.58%		
Baseline	CHiME4	CHiME4	77.10%	76.98%	76.57%	75.11%	73.52%	73.07%	75.39%		
Proposed	CHiME4	CHiME4	<u>87.92%</u>	<u>87.73%</u>	<u>87.47%</u>	<u>86.90%</u>	<u>86.35%</u>	<u>85.32%</u>	<u>86.95%</u>	15.33%	-

GRID corpus with SNRs from -5 to 20 dB. The CHiME4 noises were recorded using a tablet device at four different locations: bus, cafeteria, pedestrian area, and street junction. In a similar manner, to evaluate the performance under mismatched noise conditions, 10 non-visualized noises of the East Anglia dataset [37] were randomly mixed into all the testing audio recordings with SNRs from -5 to 20 dB in 5-dB steps. The East Anglia dataset provides environmental noise collected from 10 different locations: bar, beach, bus, car, football match, launderette, lecture, office, rail station, and street. Moreover, for the evaluation in matched noise conditions, the CHiME4 noises were mixed again into all the testing audio recordings with SNRs from -5 to 20 dB in 5-dB steps.

4.3. Setting

The recordings of each speaker of the GRID corpus were partitioned into training, validation, and testing sets in the ratio 6:2:2. The validation set was used to identify the right set of hyperparameters for each model and each experimental condition empirically. For the input audio sequence, all audio recordings were resampled with a 16-kHz sampling rate and a 128-dimensional log-compressed Mel spectrogram [38] was computed by the short-time Fourier transform [39] with a window size of 1280 samples (80 ms) and hop length of 640 samples (40 ms). For the input video sequence, all video recordings were converted to 25 frames per second (40 ms) with a resolution of $H \times W = 224 \times 224$ pixels. The dimensions of spatial and feature embedding $I \times J \times D$ were $56 \times 56 \times 128$. The number of unsynchronized audio sequences U was set at 30 and augmented by shifting the corresponding frames of the input audio sequence in the range of ± 640 ms at each iteration for the negative pairs. All models were trained by an Adam optimizer [40] with gradient clipping [41] for 70 epochs to observe loss convergence. The learning rate was initialized at $1e-3$ and halved every 10 epochs, and the batch size was set to 1 [42].

4.4. Quantitative results

The area under the receiver operating characteristic curve (AUROC) [43] was used as a metric for quantitative performance evaluation of AV-VAD. The experimental results for each experimental condition are summarized in Table 1. Among the experimental results, the proposed method with the branched audio encoder and noise-tolerant contrastive learning significantly outperformed the baseline method: the relative improvements were 1.71%, 14.84%, and 15.33% for each

experimental condition. Particularly, in the case that CHiME4 noise was added into training datasets, the average AUROC by the baseline method significantly deteriorated by 10.57% from 82.45% to 73.73% in mismatched noise conditions (East Anglia), while the average AUROC of the proposed method improved by 0.97% from 83.86% to 84.67% under the same mismatched noise conditions. This result demonstrates that the baseline method is highly vulnerable to noisy training datasets and easily overfitted to noise, resulting in performance degradation. Meanwhile, the proposed method could distinguish and isolate noise features in a self-supervised manner both for clean and synthetic noise datasets. In performance comparison with the supervised method (Eq. (10)) in matched noise conditions (CHiME4), the supervised method outperformed both the baseline and proposed methods, which are based on self-supervised learning. However, the proposed method could produce better results at lower SNRs (-5 to 5 dB) in mismatched noise conditions, resulting in a higher average AUROC: the relative improvements were 6.97% and 1.68% from training on noise-free and noise-corrupted datasets, respectively.

Figure 4 illustrates an example of performance improvement by the proposed method compared with the baseline method under matched noise conditions. The proposed method improved overall accuracy for predicting the presence of speech in both speech-present and speech-absent segments. The method also produced better AV correlations, as can be seen in the heat map visualization of the learned AV attention map.

5. Conclusions

This paper proposed a new noise-tolerant self-supervised learning framework for AV-VAD on noise-corrupted training datasets that approximate real-world scenarios. The proposed learning framework includes an audio encoder branched to extract speech and noise embedding features individually and a loss function to solve the uncertainty of speech and noise feature separation in a self-supervised manner based on the facts that speech and noise have different energy distributions over the frequency range and that noise tends to be less correlated to key visual information, such as mouth movement. Among the experimental results, including matched and mismatched noise conditions, the proposed learning framework significantly outperformed a baseline learning framework by up to 15.33% for the average AUROC. It also produced better results, by up to 6.97%, for mismatched noise conditions, compared with a supervised learning framework.

6. References

- [1] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system Robustness," *Robust Speech Recognition and Understanding, Book*, pp. 1–22, 2007.
- [2] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.
- [3] R. G. Bachu, S. Koppurthi, B. Adapa and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," *Advanced Techniques in Computing Sciences and Software Engineering*, 2010
- [4] R. Tucker, "Voice activity detection using a periodicity measure", in *Proceedings of the Institution of Electrical Engineers*, vol. 139, no. 4, pp. 377–380, Aug. 1992.
- [5] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*, 1993, pp. 321–324.
- [6] J. Shen, J. Hung, and L. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proceedings of International Conference on Spoken Language Processing*, pp. 232–235, 1998.
- [7] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proceedings INTERSPEECH 2012*.
- [8] R. Sarikaya, and J. Hansen, "Robust detection of speech activity in the presence of noise," in *Proceedings of International Conference on Spoken Language Processing*, pp. 1455–1458, 1998
- [9] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.
- [10] J. Ramirez, J. Segura, M. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, pp. 689–692, 2005.
- [11] J. Chang, N. Kim, and S. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.
- [12] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," in *Proceedings INTERSPEECH 2013*.
- [13] X. L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proceedings INTERSPEECH 2014*.
- [14] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proceedings of ICASSP2014*.
- [15] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.
- [16] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proceedings ICASSP 2013*.
- [17] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proceedings ICASSP 2013*.
- [18] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," in *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181–1185, Aug. 2018.
- [19] 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, April 2015.
- [20] T. L. Cornu and B. Milner, "Voicing classification of visual speech using convolutional neural networks," in *Proceedings of the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*, pp. 103–108, 2015.
- [21] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," *Signal Processing*, vol. 142, pp. 69–74, Jan. 2018,
- [22] S. O. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, pp. 197–200, 2013.
- [23] F. Germain, D. Sun, and G. Mysore, "Speaker and Noise Independent Voice Activity Detection," in *Proceedings INTERSPEECH 2013*.
- [24] R. Arandjelović, A. Zisserman, "Objects that Sound," in *Proceedings ECCV 2018*.
- [25] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings ECCV 2018*.
- [26] A. Senocak, T. -H. Oh, J. Kim, M. -H. Yang, I. S. Kweon, "Learning to Localize Sound Source in Visual Scenes," in *Proceedings CVPR 2018*.
- [27] D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *Computer Vision – ECCV 2018*.
- [28] T. Afouras, A. Owens, J. -S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proceedings ECCV 2020*.
- [29] S. Ioffe and C. Szegedy, "Batchnormalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings ICML 2015*.
- [30] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings ICML 2010*.
- [31] U. -H. Kim and H. G. Okuno, "Improved binaural sound localization and tracking for unknown time-varying number of speakers," *Advanced Robotics*, vol. 27, pp. 1161–1173, 2013.
- [32] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
- [33] T. Kim, T. Eltoft and T. -W. Lee, "Independent vector analysis: an extension of ICA to multivariate components" *Independent Component Analysis and Blind Signal Separation*, vol. 3889, pp. 165–172, 2006.
- [34] J. J. Liang, C. T. Yue and B. Y. Qu, "Multimodal multi-objective optimization: A preliminary study," in *Proceedings of IEEE Congress on Evolutionary Computation*, pp. 2454–2461, 2016.
- [35] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Acoustical Society of America*, vol. 120, pp. 421–2424, 2006.
- [36] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "The 4th CHiME speech separation and recognition challenge," http://spandh.dcs.shef.ac.uk/chime_challenge/CHiME4/.
- [37] L. Ma, D. Smith, and B. Milner, "Context awareness using environmental noise classification," in *Proceedings EUROSPEECH 2003*.
- [38] M. Sahidullah, and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, pp. 543–565, May 2012.
- [39] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 25, pp. 235–238, Jun. 1977.
- [40] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proceedings ICLR 2015*.
- [41] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *Proceedings ICLR 2020*.
- [42] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima." in *Proceedings ICLR 2017*.
- [43] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29–36, 1982.