INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection

Fei Tao, Carlos Busso
Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
fxt120230@utdallas.edu, busso@utdallas.edu

Abstract

Voice activity detection (VAD) is an important preprocessing step in speech-based systems, especially for emerging hand-free intelligent assistants. Conventional VAD systems relying on audio-only features are normally impaired by noise in the environment. An alternative approach to address this problem is *audiovisual VAD* (AV-VAD) systems. Modeling timing dependencies between acoustic and visual features is a challenge in AV-VAD. This study proposes a bimodal *recurrent neural network* (RNN) which combines audiovisual features in a principled, unified framework, capturing the timing dependency within modalities and across modalities. Each modality is modeled with separate *bidirectional long short-term memory* (BLSTM) networks. The output layers are used as input of another BLSTM network. The experimental evaluation considers a large audiovisual corpus with clean and noisy recordings to assess the robustness of the approach. The proposed approach outperforms audio-only VAD by 7.9% (absolute) under clean/ideal conditions (i.e., *high definition* (HD) camera, close-talk microphone). The proposed solution outperforms the audio-only VAD system by 18.5% (absolute) when the conditions are more challenging (i.e., camera and microphone from a tablet with noise in the environment). The proposed approach shows the best performance and robustness across a varieties of conditions, demonstrating its potential for real-world applications.

Index Terms: voice activity detection, multimodal signal processing, deep learning, recurrent neural network.

1. Introduction

Voice activity detection (VAD) is a key preprocessing step in speech-based interfaces. Emerging technologies in hand-free intelligent assistants such as Amazon's Alexa and Google's voice search have increased the needs for robust VAD in real scenarios, avoiding the use of push-to-talk interfaces. Conventional VAD systems rely on acoustic features (i.e., *audio-based voice activity detection* (A-VAD)), where the performance may be severely impaired with noisy environment. If the VAD algorithm fails, the rest of the components in the speech-based system will be affected (e.g., *automatic speech recognition* (ASR) systems). Robust VAD solutions are necessary for practical application in real world.

Introducing visual information is an appealing alternative to increase the robustness of VAD. Previous studies have demonstrated that adding *visual-based voice activity detection* (V-VAD) systems relying on facial features can improve the performance in noisy environments [1, 2]. We also demonstrated that V-VAD can improve robustness against different speech modes (e.g., whispering speech) [3]. Fusing audiovisual information is also feasible in many applications, since current smart devices have cameras. However, conventional *audiovisual-based voice activity detection* (AV-VAD) approaches use specific rules to fuse modalities, providing suboptimal frameworks that lack flexibility in capturing the relation between speech and facial features. Advances on deep learning provide an opportunity to address this problem in a principled way. They can model more complicated relationship between modalities by stacking non-linear layers, mapping the distribution in a manifold space [4]. In particular, *recurrent neural network* (RNN) can model the timing dependencies between facial and acoustic features, providing an ideal framework for real-time AV-VAD applications.

This study proposes a bimodal RNN for AV-VAD. Each modality is modeled with separate *bidirectional long short-term memory* (BLSTM) networks. This step models the temporal relationship within modalities. The output layers are then used as inputs of another BLSTM network which captures the relationship across modalities. The evaluation considers a large audiovisual corpus with 37.2 hours collected from 74 subjects. The evaluation considers different microphones and noise conditions. The proposed AV-VAD system outperforms the baseline A-VAD system by 7.9% under clear conditions. Under noisy conditions, the improvement is even larger, achieving F-score of 88.2% using the microphone and camera of a commercial tablet. To the best knowledge of the authors, this is the first AV-VAD system using RNN.

2. Background

2.1. Audio-Visual Voice Activity Detection

V-VAD is an emerging research area. Previous studies have shown that introducing visual cues in speech-based system, including VAD [1–3, 5], is an effective approach to improve robustness against noisy environments and different speech modes [6–9].

Liu and Wang [10] developed a V-VAD relying on *Gaussian mixture model* (GMM). They extracted hand-craft features from the mouth area and augmented them with their first order derivative. They reduced the feature dimension using *principle component analysis* (PCA). Several visual features have been used for this task. Petsatodis et al. [11] extracted geometric features describing the vertical distance of the mouth opening, and Almajai and Milner [12] used appearance features relying on 2D *discrete cosine transform* (DCT). Both of these studies applied first order derivatives as dynamic information, which is important to detect lip movements associated with speech. Aubrey et al. [13] extracted static features including *active appearance model* (AAM). The dynamic aspect was introduced at the model level with *hidden Markov model* (HMM). More recently, Joosten et al. [14] proposed *spatiotemporal Gabor filters* (SGFs) for this task.

V-VAD systems are generally fused with speech-based systems creating AV-VAD solutions. Takeuchi et al. [15] applied the logical operations "AND" and "OR" to combine the decision boundaries of V-VAD and A-VAD. Almajai and Milner [12] concatenated acoustic and visual features integrating

This work was funded by NSF CAREER award IIS-1453781.

http://dx.doi.org/10.21437/Interspeech.2017-1573

(a) (b)

Figure 1: *Sample images showing the data collection settings to collect the UT-CRSS-4EnglishAccent corpus.*

the modalities at the feature level. Petsatodis et al. [11] designed a fusion rule, where V-VAD would be considered only when the lips were detected. However, these approaches assigned equal weights to each modality which may not provide the best performance. One exception is the work of Buchbinder et al. [2], where they explored dynamic weights relying on *signal-to-noise-ratio* (SNR) estimation.

2.2. Deep Learning Techniques in VAD

Solutions based on deep learning have been widely used in many areas of speech processing, including A-VAD. Ryant et al. [16] extracted *Mel frequency cepstral coefficients* (MFCCs) as features to train a *deep neural network* (DNN) as classifier. The system was evaluated with the HAVIC corpus [17], which consists of audio from YouTube videos. The study reported 19.6% frame error rate. Zhang et al. [18–20] developed an ensemble of DNNs, where each DNN was trained with cochleagram features at different time resolutions.

DNN does not capture temporal relationship which is important for VAD. To address this problem, studies have proposed VAD frameworks based on RNN, which have shown improved performance on ASR [21]. Eyben et al. [22] and Hughes et al. [23] relied on RNNs to model timing dependency for VAD tasks, using *perceptual linear predictive* (PLP) features. Zazo et al. [24] proposed a end-to-end framework that combines *convolutional neural network* (CNN) and RNN. CNN takes raw speech as input extracting relevant features that are fed into the RNN, to get a silence/speech class label per frame. They achieved 4.2% *false alarm rates* (FAR) under noisy conditions, after fixing the system at 2% *false reject rate* (FRR).

Most of previous work with deep learning focused on read speech, which is an easier task than spontaneous speech. To the best knowledge of the authors, there is not study using RNN for AV-VAD. This study proposes an AV-VAD that provides competitive performance even with spontaneous speech.

3. Database and Audiovisual Features

3.1. The UT-CRSS-4EnglishAccent corpus

This study uses the UT-CRSS-4EnglishAccent corpus collected by the *Center for Robust Speech Systems* (CRSS) at *The University of Texas at Dallas* (UTD). This corpus contains 442 subjects with four English accents: American (115), Australian (103), Indian (112) and Hispanic (112).

The data was collected in a $13ft \times 13ft$ sound booth, which is ASHA certified (Fig. 1). The data collection includes five microphones corresponding to a close-talk microphone, a desktop microphone, two microphones from the Samsung Galaxy SIII cellphone, and a microphone from a Samsung Galaxy tablet. The close-talk microphone was placed close to the subjects' mouth. The desktop microphone was placed on the desk in front of the subjects. The Samsung Galaxy SIII cellphone was on the desk. The Samsung Galaxy tablet was placed on a mount holder about 1.5 meters in front of the subjects. The five channels were converted into 16KHz. We used two cameras during the recordings: a Samsung Galaxy tablet (that also recorded the audio) and Sony *high definition* (HD) camera. The tablet recorded at 24 *frames per second* (fps) with a resolution of 1280×720. The HD camera, placed next to the tablet, recorded at 29.97 fps with a resolution of 1440×1080.

A computer monitor was placed two meters in front of the subjects. They read the text displayed on the monitor, which included isolated words, short phrases, and sentences. They also recorded spontaneous speech, where they answered questions displayed on the monitor. The data collection has clean and noisy sections. After recording the session without noise, a subset of the slides with prompted speech were repeated (i.e., no spontaneous speech). We use a loud speaker to play four types of noises which were prerecorded (restaurant, shopping mall, office and house). The loud speaker was placed 2.5 meters from the subjects, close to the tablet. The subjects recorded prompted speech for five minutes under different noisy environments. A key feature of the corpus is that the noise was played during the recordings, and not artificially added after the data collection. The data was manually transcribed. The speech/silence labels were obtained with forced alignment using the data from the close-talk microphone and the transcriptions.

The evaluation considers two settings. The *ideal* condition uses the close-talk microphone and the HD camera. The *tablet* condition uses the camera and microphone from the tablet. Other audio channels were not used in this study. We manually synchronized the audio and video streams using a clapboard. During early recordings, we lost some of the videos collected with either the HD camera or the camera from the tablet. This study only uses data from 74 subjects with American accent, from whom we have all the modalities. This dataset has 37.2 hours.

3.2. Acoustic and Visual Features

This study uses the audiovisual features used in our previous work [3, 5]. The acoustic features correspond to the 5D feature vector proposed by Sadjadi and Hansen et al. [25] for A-VAD. It includes harmonicity, clarity, prediction gain, periodicity and perceptual spectral flux. Harmonicity refers to *Harmonics-to-Noise* (HNR), and it is used to measure the quality of periodic signals. Clarity is defined as the relative depth of the minimum *average magnitude difference function* (AMDF) valley in the assumed pitch range. Prediction gain is defined as the energy ratio of the original signal to the *linear prediction* (LP) residual signal. Perceptual spectral flux capture the quasi-stationary feature of voice activity.

The visual features capture lip activity associated with speech. We detect facial landmarks using Intraface [26], defining the *region of interest* (ROI) around the lips. We estimate the variance of the optical flow within the ROI, including vertical and horizontal directions, and their summation. We also estimate geometric features, including the width, height, perimeter and area of the mouth. This process forms a 7D feature vector. Dynamic information is very important to discriminate speech from non-speech related lip movements. We estimate three statistics over overlapped windows that are shifted one frame at a time: variance, *zero crossing rate* (ZCR) and *speech periodic characteristic* (SPC) (details of these statistics are given in Tao et al. [5]). To balance the tradeoff between resolution (requiring

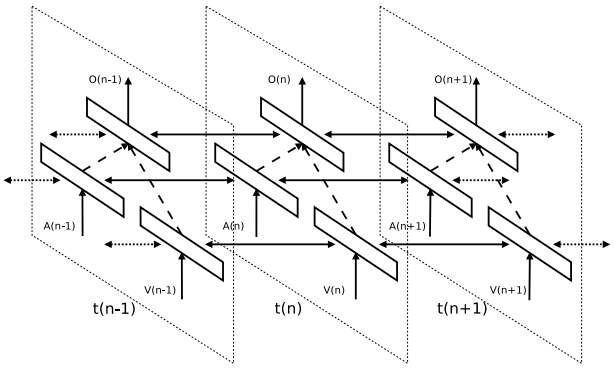


Figure 2: *The proposed bimodal RNN. $A(n)$ and $V(n)$ are acoustic and visual features at time $t(n)$. $O(n)$ is the corresponding output. Dashed arrows represents the concatenation of hidden values from audio and visual modalities. Bidirectional arrows represent the BLSTM.*

short window) and robust estimation of the statistics (requiring long window), we set the window size to 9 frames (approximately 0.3s). We estimate the variance, ZCR and SPC of each of the 7D feature vector, creating a 21D feature vector. In addition, we estimate the first order derivatives only on the four geometric features. We did not include the first order derivative of the optical flow feature because the resulting values were not informative about speech activity [5]. A 25D feature vector is formed by concatenating the vectors from the statistics and first order derivatives. Finally, we append the summation of the optical flow variance in two directions to form a 26D feature vector.

We z-normalized the acoustic and visual features at the utterance level to preserve similar feature ranges across speakers.

4. Approach

The proposed AV-VAD relies on deep learning. By deploying varieties of nonlinear functions within layer and stacking multiple layers, deep learning forms powerful feature representations to model different distributions in a manifold space [4]. This framework does not need any assumption about the distribution of the data, or the relationship between the input modalities. In addition, the weights between layers are adjusted by error propagation, which is totally data driven. This feature is appealing for audiovisual processing, since it provides dynamic weighting capability to model the relationship between modalities [27]. The flexibility of deep learning structures motivates our team to propose a principled framework for AV-VAD.

We propose a bimodal RNN implemented with BLSTM. The structure is a hierarchal network with separate subnetworks for speech and visual modalities. The unimodal substructures are regular feedforward layers capturing the dependencies within the modalities. Some of the layers are implemented with BLSTM. The outputs of these sub-networks are concatenated (80D) to form the input of another RNN implemented with BLSTM. It is expected that the feedforward layers will process the input features, and the BLSTM will capture the temporal dependency between modalities. By using bidirectional structures, we also expect to leverage information from previous and future frames.

Figure 2 shows the proposed bimodal RNN, where we unrolled the network. For simplicity, the figure only shows the recurrent connections. Each rectangle denotes a deep network. All the sub-networks are jointly trained, where the weighs for the modalities are learned from the data.

The implementation of the network is as follows. The network for each modality has four layers, where the first two layers are maxout layer and the last two layers are BLSTM. The network for the acoustic features has 16 neurons per layer, and the network for the visual features has 64 neurons per layer. The network that combines the modalities also has four layers. The first two layers are BLSTM, the third layer is a maxout layer, and the last layer is the softmax layer. Each layer has 128 neurons, except the softmax layer. The softmax layer is added to classify the frame as speech or silence. All the maxout neurons used in this study have 3 units.

5. Experiment and Results

We performed subjects independent evaluation in the experiments, where the training set consists of data from 67 randomly selected subjects (32.5 hours). The testing set includes data from the remaining seven subjects (4.7 hours). We train the data with features extracted from the close-talk microphone and HD camera without noise (e.g., *ideal* conditions). We evaluate the proposed solution with clean and noisy recordings under *ideal* and *tablet* conditions. Since we have five minute of noisy recording per subject, its duration in the testing set is only 35 minutes. The rest of of testing set corresponds to recordings without noise (4.1 hours). In addition to accuracy, we also report the precision and recall rates, from which we estimate the F1-score (the target class is speech).

5.1. Baselines

We evaluate the performance of the proposed system with unimodal systems. We train A-VAD and V-VAD systems implemented with DNN. We build these DNNs using six layers, where the first five layers are maxout layers and the last layer is a softmax layer. For A-VAD, we use 16 neurons per layer. For V-VAD, we use 64 neurons.

We also evaluate the proposed approach with two competitive AV-VAD alternatives. First, we replace all the BLSTM layers with maxout layers. This structure has the same number of layers and neurons as the proposed system. However, it does not have recurrent connections, where the speech/ non-speech decision is independently determined for each frame. We refer to this structure as *AV-VAD_DNN*. Second, we implement a system where the audiovisual features are concatenated and used as inputs of a RNN. We refer to this system as *Concatenated RNN*. This system is implemented with six layers, where the first two are maxout layers, the next two are BLSTM layers, the fifth one is maxout layer, and the last one is softmax layer. All the layers have 128 neurons, except the softmax layer. For all the networks in this paper, we apply 0.1 dropout and use the Adam learning approach [28].

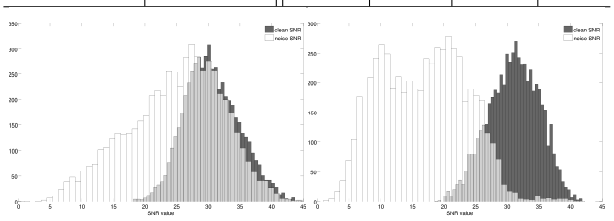
5.2. Results on Recordings without Noise

Table 1 shows the results on recordings without noise. A-VAD system outperforms the V-VAD system, showing the benefits of using acoustic features for this task. However, the three AV-VAD systems evaluated in this study outperform the system based on single modalities. Adding visual cues provides complementary information improving the VAD performance even in clean recordings.

The table also demonstrates the benefits of recurrent connections. Both approaches using RNN (Concatenated RNN, Bimodal RNN) outperform the AV-VAD_DNN with gains between 4.7% to 5.6% in F1-score. RNN models the timing dependency between frames, which is informative to detect voice

Table 1: *Performance of VAD systems with clean recordings. We report the results for ideal (close-talk microphone, HD camera) and tablet (microphone and camera from the tablet) conditions [Acc: accuracy, Pre: precision, Rec: recall, F: F1-score].*

Approach	Test Condition	Acc[%]	Pre[%]	Rec[%]	F[%]
A-VAD	Ideal	85.59	93.19	78.57	85.59
	Tablet	86.88	91.79	82.59	85.95
V-VAD	Ideal	79.98	78.94	84.89	81.80
	Tablet	78.62	79.63	80.06	79.84
AV-VAD_DNN	Ideal	87.75	92.44	83.76	87.88
	Tablet	88.39	92.95	84.46	88.50
Concatenated RNN	Ideal	93.19	94.57	92.49	**93.51**
	Tablet	92.78	92.63	93.81	93.22
Bimodal RNN	Ideal	93.18	94.94	92.07	93.48
	Tablet	92.86	92.91	93.64	**93.27**



(a) Close-talk microphone (b) Tablet's microphone

Figure 3: *The SNR distributions from clean and noisy recordings. The tablet's microphone is more affected by the noise.*

activity. For clear recordings, the performance of the *Concatenated RNN* system and the proposed *Bimodal RNN* system is very similar. Notice that the proposed approach outperforms the A-VAD baseline, showing gains between 7.3% and 7.9%.

When we compare the *ideal* condition (close-talk microphone, HD camera) with *tablet* condition (microphone and camera from the tablet), we observe that the performances are similar. Without noise, the quality of the audio from both microphones is good enough for VAD. The close performance between conditions indicates that the extracted features are robust and the proposed deep learning architecture is able to handle the channel mismatch.

5.3. Results on Recordings with Noise

This section presents the results on recordings with noise, which are listed in Table 2. An important remark is that spontaneous speech was not included during noisy recordings (Sec. 3.1). It only includes prompted speech which is an easier problem for VAD. For example, the F1-scores for the V-VAD system in Table 2 (noisy recordings) are higher than the one reported in Table 1 (clean recordings). Notice that the visual features are not affected by acoustic noise (although lombard speech may affect lips movement), so the gain in performance depends on the task. Therefore, it is not straightforward to compare the results from both tables.

A second important note is that the *signal to noise ratio* (SNR) for the close-talk microphone (*ideal* condition) is higher than the SNR in the tablet's microphone (*tablet* condition), since the loud speaker was placed close to the tablet. We estimate the SNR on the sentences from the clean and noisy recordings using the NIST speech SNR toolkit. Figure 3 shows their distributions. For the close-talk microphone, the SNR distributions for clean and noisy recordings overlap, suggesting that the noise did not significantly affect the speech signal, since the microphone was placed close to the subjects' mouth. For the tablet's microphone, however, the SNR distributions are different. The lower SNR is the primary reason of the 9.3% drop in F1-score in the A-VAD system when evaluated in the *ideal* (78.9%) and *tablet* (69.6%) conditions (Table 2).

Table 2: *Performance of VAD systems with noisy recordings. We report the results for ideal (close-talk microphone, HD camera) and tablet (microphone and camera from the tablet) conditions [Acc: accuracy, Pre: precision, Rec: recall, F: F1-score]*

Approach	Test Condition	Acc[%]	Pre[%]	Rec[%]	F[%]
A-VAD	Ideal	75.38	74.11	84.40	78.92
	Tablet	58.75	58.06	86.97	69.64
V-VAD	Ideal	82.20	80.50	88.96	84.52
	Tablet	78.93	79.89	81.86	80.85
AV-VAD_DNN	Ideal	84.76	83.11	90.47	84.76
	Tablet	72.41	68.47	91.32	78.26
Concatenated RNN	Ideal	93.18	92.85	94.82	93.82
	Tablet	84.50	80.24	94.86	86.94
Bimodal RNN	Ideal	93.82	93.79	94.97	**94.38**
	Tablet	85.84	80.86	96.90	**88.16**

Table 2 shows that using audiovisual solutions with RNN improves the performance of the system over A-VAD and V-VAD systems. The table confirms the importance of modeling temporal dynamic using recurrent connections (the AV-VAD_DNN system has similar performance than V-VAD). In the presence of noise, the best performance is achieved with the proposed bimodal RNN, achieving 94.4% and 88.2% F1-scores for *ideal* and *tablet* conditions, respectively. These F1-scores are higher than the one observed for the concatenated RNN system. These results show that the performance of the proposed system is competitive under clean and noise conditions, providing an appealing solution for practical applications.

6. Conclusion and Future Work

This paper proposed a bimodal RNN for AV-VAD. The approach takes audiovisual features as inputs deriving reliable VAD decisions for clean and noisy recordings. The framework models each modality with RNN, capturing temporal dependencies within modalities. The output layers of each modality are concatenated, and fed as input of another RNN modeling the temporal dependencies across modalities. The results showed that the proposed approach outperformed the A-VAD system by 7.9% when it was tested with clean data under *ideal* conditions (close-talk microphone, HD camera). With noisy recordings, the proposed approach achieved the best result, showing important improvements over other unimodal systems and audiovisual solutions. The results show the benefits of the proposed solution.

The evaluation of the approach considered a subset of the UT-CRSS-4EnglishAccent corpus which includes 37.2 hours of audiovisual recordings using multiple microphones and cameras. Our data contains spontaneous speech, where recordings material are close to real-world applications. The corpus also includes a commercial tablet. The results with the camera and microphone from this tablet (i.e., *tablet* condition) demonstrated that the proposed solution can be effectively used in practical applications. Although the evaluation considered offline processing, the proposed approach can be implemented in real-time.

This study used hand-craft features, which were pre-defined following our experience from previous studies. Our goal is to create an end-to-end system, which takes raw audiovisual inputs. Another future direction is to deal with missing information from features from one modality. We are investigating deep learning techniques to address this problem.

7. References

[1] 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, April 2015.

[2] M. Buchbinder, Y. Buchris, and I. Cohen, "Adaptive weighting parameter in audio-visual voice activity detection," in *IEEE International Conference on the Science of Electrical Engineering (ICSEE 2016)*, Eilat, Israel, November 2016, pp. 1–5.

[3] F. Tao, J. Hansen, and C. Busso, "An unsupervised visual-only voice activity detection approach using temporal orofacial features," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2302–2306.

[4] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, January 2009.

[5] F. Tao, J. L. Hansen, and C. Busso, "Improving boundary estimation in audiovisual speech activity detection using Bayesian information criterion," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2130–2134.

[6] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Workshop 2000 Final Report, Technical Report 764, October 2000.

[7] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, March 2009.

[8] X. Fan, C. Busso, and J. Hansen, "Audio-visual isolated digit recognition for whispered speech," in *European Signal Processing Conference (EUSIPCO-2011)*, Barcelona, Spain, August-September 2011, pp. 1500–1503.

[9] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *Interspeech 2014*, Singapore, September 2014, pp. 1154–1158.

[10] P. Liu and Z. Wang, "Voice activity detection using visual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. 609–612.

[11] T. Petsatodis, A. Pnevmatikakis, and C. Boukis, "Voice activity detection using audio-visual information," in *International Conference on Digital Signal Processing (ICDSP 2009)*, Santorini, Greece, July 2009, pp. 1–5.

[12] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *European Signal Processing Conference (EUSIPCO 2008)*, Switzerland, Lausanne, August 2008.

[13] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filltering," in *European Signal Processing Conference (EUSIPCO 2007)*, Poznań, Poland, September 2007.

[14] B. Joosten, E. Postma, and E. Krahmer, "Visual voice activity detection at different speeds," in *International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, Annecy, France, August-September 2013.

[15] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," in *International Conference on Audio-Visual Speech Processing (AVSP 2009)*, Norwich, United Kingdom, September 2009, pp. 151–154.

[16] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on Youtube using deep neural networks," in *Interspeech 2013*, Lyon, France, August 2013, pp. 728–731.

[17] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel, "Creating HAVIC: Heterogeneous audio visual internet collection," in *International conference on Language Resources and Evaluation (LREC)*, vol. 4, Istanbul, Turkey, May 2012, pp. 2573–2577.

[18] X. 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, April 2013.

[19] X. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Interspeech 2014*, Singapore, September 2014, pp. 1534–1538.

[20] ——, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, February 2016.

[21] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 4945–4949.

[22] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 483–487.

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

[24] R. Zazo, T. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3668–3672.

[25] S. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, March 2013.

[26] F. De la Torre, W. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *International conference on machine learning (ICML2011)*, Bellevue, WA, USA, June-July 2011, pp. 689–696.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv e-prints (arXiv:1412.6980)*, December 2014.