Interspeech 2018
2-6 September 2018, Hyderabad

Audiovisual Speech Activity Detection with Advanced Long Short-Term Memory

Fei Tao and Carlos Busso

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

fxt120230@utdallas.edu, busso@utdallas.edu

Abstract

Speech activity detection (SAD) is a key pre-processing step for a speech-based system. The performance of conventional *audio-only SAD* (A-SAD) systems is impaired by acoustic noise when they are used in practical applications. An alternative approach to address this problem is to include visual information, creating *audiovisual speech activity detection* (AV-SAD) solutions. In our previous work, we proposed to build an AV-SAD system using *bimodal recurrent neural network* (BRNN). This framework was able to capture the task-related characteristics in the audio and visual inputs, and model the temporal information within and across modalities. The approach relied on *long short-term memory* (LSTM). Although LSTM can model longer temporal dependencies with the cells, the effective memory of the units is limited to a few frames, since the recurrent connection only considers the previous frame. For SAD systems, it is important to model longer temporal dependencies to capture the semi-periodic nature of speech conveyed in acoustic and orofacial features. This study proposes to implement a BRNN-based AV-SAD system with *advanced LSTMs* (A-LSTMs), which overcomes this limitation by including multiple connections to frames in the past. The results show that the proposed framework can significantly outperform the BRNN system trained with the original LSTM layers.

Index Terms: speech activity activation, advanced LSTM, bimodal RNN, audiovisual speech processing, deep learning.

1. Introduction

Speech activity detection (SAD) systems aim to discriminate between speech and non-speech segments. It is usually used as a pre-processing step in speech-based *human-computer interface* (HCI) or *artificial intelligence* (AI) products (e.g., in intelligent personal assistant with speech capability such as Siri and Alexa). SAD performance is important, since mis-detection of speech segments will lead to mistakes in the following speech processing steps, such as *automatic speech recognition* (ASR). Background acoustic noise in real world will impair the performance of the *audio-only SAD* (A-SAD). Introducing visual information to a SAD system can lead to improvements in speech detection performance and robustness against environmental conditions [1, 2]. Conventional *audiovisual SAD* (AV-SAD) systems rely on pre-defined rules to fuse the two modalities, such as logical operations [3]. These rules provide a rigid framework without the flexibility to model the temporal information within and across modalities. Recent advances in deep learning techniques provide appealing data-driven frameworks to address this problem. The frameworks are flexible since the models depend on the input data rather than pre-defined rules.

Tao and Busso [4] proposed a *bimodal recurrent neural network* (BRNN) for modeling three important aspects: (1) the characteristics of the input feature space; (2) the relationship between audio and visual modalities; and (3) the temporal information within the input sequences. The BRNN consists of three *recurrent neural networks* (RNNs). The first two RNNs process the input modalities, one for the audio features and one for the visual features. These RNNs model the feature space and the temporal dependencies within each modality. The hidden values from the top layers of these two RNNs are concatenated and used as input of a third RNN. The third RNN is able to capture the relationship between modalities. The RNNs in the BRNN framework relied on *long short-term memory* (LSTM) to model temporal information. LSTM takes inputs from the lower layer and current layer at the previous frame. While the cells in the model are expected to capture the long and short term dependencies, the explicit temporal link on only the previous frame reduces the capability of a LSTM to model temporal dependencies beyond 100 frames [5]. An appealing solution to improve the temporal modeling of the network is by considering temporal information from multiple frames in the past. Previous studies have showed that a model can improve its performance when the higher layer have access to longer contextual information [5, 6]. Tao and Liu [7] proposed an *advanced long short-term memory* (A-LSTM) layer, which takes inputs from the lower layer and the current layer at several frames in the past. The A-LSTM layer, therefore, offers better temporal modeling capability than a standard LSTM layer [8].

For AV-SAD, extending the temporal modeling is important as longer periodicity in the acoustic and visual features is crucial to distinguish between speech and non-speech segments. This study proposes to explore the A-LSTM layers in the BRNN system to improve the temporal dependency of our AV-SAD system. We evaluate the proposed approach on the CRSS-4ENGLISH-14 corpus under different conditions. The proposed framework significantly improves the performance of the AV-SAD system, with absolute gain of 1.1% on noisy data recorded with mismatched channel conditions over a BRNN system implemented with the original LSTMs.

2. Related Work

Introducing visual information can improve the performance and robustness of speech-based systems [9–11]. This section summarizes studies on AV-SAD systems, focusing on their modeling frameworks.

Different schemes have been explored for audiovisual fusion in the context of AV-SAD. Takeuchi et al. [3] proposed the logical operations "AND" and "OR" to combine the decisions from a A-SAD system and a *visual-only SAD* (V-SAD) system. Almajai and Milner [12] used concatenated audio and visual features to fuse the two modalities. Petsatodis et al. [13] defined a rule for fusion, adding visual features only when the lips were detected. Otherwise, their system relied only on the acoustic features. These fusion schemes provide rigid rules, which

This work was funded by NSF awards IIS-1718944 and IIS-1453781 (CAREER).

10.21437/Interspeech.2018-2490

(a) (b)

Figure 1: *Data collection setting with the equipments used in the recordings of the CRSS-4ENGLISH-14 corpus [22].*

are not flexible to address different environmental conditions. Buchbinder et al. [14] proposed to combine audio and visual modalities relying on dynamic weights based on the *signal-to-noise-ratio* (SNR) estimation. This scheme can provide flexibility under different conditions. For example, when the speech was noisy, the visual modality received a higher weight. However, this scheme relies on the accuracy of the SNR estimation algorithm.

Deep learning techniques have shown advantages in many areas of speech processing, since they provide powerful data-driven frameworks to learn relevant patterns [15, 16]. Several studies have explored SAD with DNN. Ryant et al. [17] fed *Mel frequency cepstral coefficients* (MFCCs) to a *deep neural network* (DNN). The result showed that a DNN can outperform a conventional approach relying on *Gaussian mixture model* (GMM). A DNN is a static model, which may not be suitable for modeling dynamic information. *Recurrent neural network* (RNN) was proposed to model the temporal information, which is more suitable for a SAD task [18, 19]. Recently, end-to-end frameworks relying on *convolutional neural network* (CNN) were proposed to better capture high-level representations from the raw input features. Zazo et al. [20] showed that the end-to-end framework can outperform a DNN or RNN system trained with handcrafted features. However, end-to-end frameworks are computational expensive.

There are few studies on AV-SAD using deep learning frameworks. Ariav et al. [21] relied on *deep auto-encoder* (DAE) to fuse audio and visual modalities. However, the fusion and classification systems were separately trained, where the parameters may not be totally optimized. Tao and Busso [4] proposed the BRNN framework which is able to model the feature space of each modality, to capture temporal dependency of the sequence inputs, and learn the fusion scheme from the data (Sec. 4.1). Our study builds upon this work, increasing the temporal modeling capability of the BRNN system, which is important to capture longer semi-periodic patterns in the audiovisual features that are characteristic of speech segments.

3. Database and Audiovisual Features

3.1. The CRSS-4ENGLISH-14 Corpus

The study relies on the CRSS-4ENGLISH-14 corpus [22]. The corpus was collected in a ASHA certified sound booth by the *Center for Robust Speech Systems* (CRSS) at *The University of Texas at Dallas* (UTD). During the collection, the booth was illuminated by two LED light panels providing uniform illumination (Fig. 1(a)). The corpus includes recordings from 442 subjects (217 females and 225 males), from speakers with the following English accents: Australian (103), Indian (112), Hispanic (112) and American (115).

The data was recorded with multiple microphones and cameras (Fig. 1(b)). This study uses the audios from two channels: (1) a close-talking microphone (Shure Beta 53) placed close to the subject's mouth, and (2) a microphone in a tablet placed about 150 centimeters from the subject (Samsung Galaxy Tab 10.1N). The recordings were collected at 44.1 kHz. This study uses the videos from two channels: (1) a *high-definition* (HD) camera (Sony HDR-XR100) placed beside the tablet, and (2) the camera of the tablet. The HD camera has 1440×1080 resolution with 29.97 fps. The tablet camera has 1280×720 resolution with 24 fps. The subjects were asked to use a clapping board at the beginning of the recordings to synchronize all the channels. A monitor was placed in front of the speakers to display the instructions for the subjects.

There were two sessions during the recording: clean and noisy sessions. The clean session contains read and spontaneous speech of the subjects completing multiple tasks (see Tao and Busso [22] for details). In the noisy session, a subset of the prompted text used for the read speech was randomly selected, and the subject was asked to read them again (no spontaneous speech). This time, we played noise through a loudspeaker (Beolit 12) inside the sound booth. There were four types of noise: home, office, shopping mall and restaurant. Playing noise in the background is more realistic than artificially adding noise to clean speech, as speech production changes in the presence of noise to improve speech intelligibility (Lombard effect). All the recordings were manually transcribed.

This study only uses the American speaker set to eliminate the variability due to accent. We use data from 105 subjects (55 females and 50 males), since some videos were lost during the recordings. The total duration of the data is 60 hours and 48 minutes. We split the data into three sets: train (70 subjects), validation (10 subjects) and test (25 subjects) partitions. This partition tries to keep the balance in the gender for each set.

3.2. Audiovisual Features

We adopt the audiovisual features used in our previous work [4]. The acoustic feature was proposed by Sadjadi and Hansen [23] for their A-SAD system. It consists of five speech features: harmonicity, clarity, prediction gain, periodicity and perceptual spectral flux. Harmonicity measures harmonics-to-noise ratio, and it is estimated by finding the relative height of the autocorrelation peak in a fixed range. Clarity is computed as the relative depth of the minimum *average magnitude difference function* (AMDF) valley in the plausible pitch range, which provides high values for speech segments. Prediction gain is the energy ratio between the original signal and the linear prediction residual signal. Periodicity is computed from the *harmonic product spectrum* (HPS). It is the product of the frequency-compressed copies of the original spectrum. This feature captures the periodic characteristic which is an important indicator for speech activity. Perceptual spectral flux captures the quasi-stationary characteristic in speech.

For the visual feature extraction, we obtain the *region of interest* (ROI), which corresponds to the mouth area. We rely on the toolkit IntraFace [24] to obtain the landmarks for the lips. We estimate the width, height, perimeter and area of the mouth, which are referred to as geometric features. We also calculate optical flow features on the ROI between two consecutive frames. We estimate the variance of the optical flow in the horizontal and vertical directions. The overall movement corresponds to the summation of the optical flow variances in both directions. We referred to these three features as optical flow features. We combine the geometric and optical flow features forming a 7D feature vector. Since dynamic information is closely related to speech activity, we compute three short-term functionals based on the 7D original feature vector to capture dynamic information: *zero crossing rate* (ZCR), variance

and *speech periodic characteristic* (SPC) (the details are described in Tao et al. [2]). The short-term functionals are computed within a window of nine frames, which is about 0.3 seconds. The window size is determined to balance the tradeoff between the resolution (short window) and estimation accuracy (long window). We also consider the first order derivative of the geometric features, which also provides dynamic information. We append the derivative of the geometric features and the overall optical flow to form a 26D visual feature (21D short-term feature, 4D derivative, and 1D overall optical flow).

All the audiovisual features are z-normalized at the utterance level to scale them into a similar range.

4. Proposed Approach

4.1. Bimodal Recurrent Neural Network

We adopt the BRNN framework proposed by Tao and Busso [4] as a back-end for this study (Fig. 2(a)). The BRNN consists of three RNNs. Two RNNs process the inputs, one for the acoustic features and one for the visual features. The two RNNs model the feature space of each modality, relying on deep learning techniques. The recurrent layers in these RNNs can model the temporal information within modalities. The hidden values from the top layers of these two RNNs provide a high-level representation for each modality (including dynamic and static information), capturing the discriminative patterns related to the SAD task. The hidden values are concatenated together and fed into a third RNN. The third RNN models the relationship across audiovisual features. By adjusting the parameters of the RNNs, the framework automatically tunes the weights associated with each modality learning their temporal relations from the data.

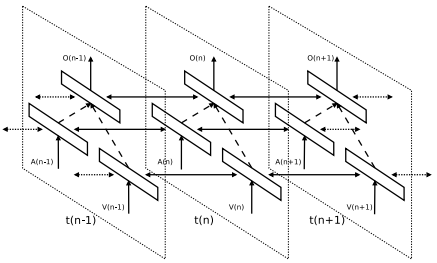
4.2. Advanced LSTM

RNNs are commonly implemented with LSTMs [25]. A conventional LSTM relies on cells to storage previous information over time, where the gates control the information flow. Equation 1 shows that the candidate information of the cell, $\widetilde{C}_t$, is computed based on the output from the lower layer, x_t, and the hidden value of the current layer at the previous frame, h_{t-1}. W_C and b_C are the trainable weight matrix and the bias vector, respectively. The cell information will be updated based on the candidate information shown in Equation 2, where C_{t-1} is the cell information from the previous frame, f_t is the forget gate vector, and i_t is the input gate vector. The gate vector is a sigmoid function with values close to either 1 or 0. The symbol $\odot$ represents element-wise multiplication (Hadamard product). The cell is updated element by element, by replacing, keeping or combining the previous information. This mechanism only has a direct link between the current and previous frames. As a result, the temporal modeling capability is limited, especially for some tasks that have longer time dependencies such as SAD.

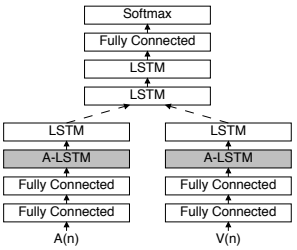
$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{1}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \tag{2}$$

To extend the temporal dependency modeled by a LSTM, Tao and Liu [7] proposed the *advanced LSTM* (A-LSTM), where C_{t-1} in Equation 2 is substituted with C', computed following Equation 3. The parameter C' is the linear combination of the cell values at several frames in the past, including $\widetilde{C}_t$ in Equation 1. C_T has the cell values at the selected time points T. The variable w_{C_t} is a scalar computed with Equation 4, where W is a trainable parameter. The variable w_{C_t} works as an attention model, weighting the cell values from previous frames considered in the model. By combining the cell values



(a) BRNN framework [4]



(b) Proposed BRNN with A-LSTM

Figure 2: *The proposed bimodal RNN with A-LSTM. $A(n)$ and $V(n)$ are acoustic and visual features at time $t(n)$. $O(n)$ is the corresponding output. The dashed arrows represent the concatenation of hidden values from acoustic and visual modalities. Figure (b) presents the implementation with A-LSTM (temporal links are omitted).*

at different frames, the current cell can model information from multiple previous frames, improving the temporal dependency in the models. Since W is a trainable parameter, the temporal dependencies are automatically learned from the data.

$$C' = \sum_T w_{C_T} \times C_T \tag{3}$$

$$w_{C_T} = \frac{\exp(W \cdot C_T)}{\sum_T \exp(W \cdot C_T)} \tag{4}$$

4.3. BRNN with A-LSTM

This study extends the BRNN framework implemented with A-LSTM layers. We denote this model *advance bimodal recurrent neural network* (A-BRNN). Figure 2(b) shows the implementation of the proposed models, omitting the temporal links. Tao and Liu [7] observed that training an A-LSTM layer requires high computational resources, since it has multiple connections to previous frames. Therefore, we do not replace all the LSTM layers with A-LSTM layers. The A-BRNN has three sub-networks. The sub-network for the acoustic features has four layers. The first two are maxout layers [26] with 16 neurons. The third layer is the A-LSTM layer with 16 neurons. The forth layer is a LSTM layer with 16 neurons. The sub-network for the visual features has a similar structure, except it has 64 neurons per layer. The third sub-network also has four layers. The first two layers are LSTM layers with 128 neurons. The third layer is a maxout layer with 128 neurons. The last layer is the softmax layer for the binary SAD task. We set T in Equation 3 equals to $\{t-1, t-6\}$, which implies that the A-LSTM layer connects each cell with two frames: the previous frame ($T = t-1$) and six frames in the past ($T = t-6$). We select $T = t-6$ since it represents a frame 200ms in the past (30 fps), which is a proper range for speech activity detection.

5. Experiments and Results

We consider two channel conditions: *high definition* (HD) and *tablet* (TG). HD data consists of the ideal scenario with audio from the close-talking microphone and the video from the HD camera. The TG data consists of a more realistic scenario with the audio and video collected by the tablet. The audio data is down-sampled to match the sampling rate of the video. While it is possible to handle the asynchrony between audio and visual features [27], we do not attempt to synchronize the inputs, leaving the temporal modeling to the A-BRNN framework. For each channel, we have clean and noisy sessions. As a result, we have four conditions in total, considering different channels and recording scenarios: clean HD, noisy HD, clean TG and noisy TG. The model is always trained with clean data using the HD channel. We test the model with the four conditions, evaluating the robustness of the proposed model to train-test mismatches. We use the two-tailed z-test to assert whether differences in performance are significant with $p - value < 0.05$.

The model proposed in Tao and Busso [4] was implemented with *bidirectional long short-term memory* (BLSTM). The use of BLSTM requires a delay to compute the backward LSTM, which is not appealing for real-time applications. Therefore, we reimplement the approach with LSTM, using this system as our baseline. The baseline BRNN has the same network structure described in Figure 2(b), except that all the recurrent layers are LSTM layers. The parameters of the BRNN correspond to the settings presented in Tao and Busso [4]. All the experiments are implemented with Keras [28] using a 8G graphical card (Nvidia GTX 1070). The optimizer is Adam [29], and the dropout rate is set to $p = 0.1$.

5.1. Speech Activity Detection

The evaluation mainly focuses on challenging conditions for SAD. We consider two datasets: (1) spontaneous speech in clean environment (first four rows in Table 1), and (2) data under noisy environment (last four rows in Table 1).

For spontaneous speech under clean condition, the proposed framework outperforms the BRNN framework for the HD and TG channels (0.3% and 0.5% absolute improvements in F-score, respectively). The differences are statistically significant. The comparison between the HD and TG channels shows no significant difference (0.1% gap), indicating that our audiovisual features are robust to channel mismatch.

The F-scores for the noisy condition are shown in the last four rows of Table 1. As discussed on Section 3.1, this set only has read speech. For the TG channel, the *signal to noise ratio* (SNR) is low under the noisy condition, since the microphone in the tablet is closer to the speaker playing the noise. The NIST speech SNR tool [30] estimates that the average SNR for this condition is 16.8 dB. The distribution of the SNR is provided in Tao and Busso [31]. The performance of the baseline BRNN model drops from 89.3% to 86.1%. The performance of an A-SAD system using BLSTM was 69.6% in the TG noisy condition, which shows the benefits of using visual information [4]. The performance of the proposed A-BRNN model drops from 89.8% to 87.2%. The table shows that the A-BRNN framework can outperform the BRNN by 1.1% (absolute), which is statistically significant. These results show the advantage of the proposed approach. For the HD channel, the results for the A-BRNN model is slightly better than the BRNN model, but the differences are not statistically different. Notice that the microphone is close to the subject's mouth, about two meters from the loudspeaker playing the noise. Therefore, the SNR is higher than in the TG channel. Also, the lack of spontaneous speech in the noisy condition makes the SAD task easier than in the clean condition, which only has spontaneous speech. Due to these reasons, the F-scores for both methods are higher for the noisy condition than for the clean condition.

Table 1: *Performance of AV-SAD systems in terms of accuracy (Acc), precision (Pre), recall (Rec) and F1-score (F) ('HD': close-talk microphone + HD camera; 'TG': sensors from tablet; 'C': clean sessions; 'N': noisy session).*

Env	Approach	Test Condition	Acc[%]	Pre[%]	Rec[%]	F[%]
C	BRNN	HD	90.1	94.5	84.8	89.4
		TG	90.1	91.9	87.0	89.3
	A-BRNN	HD	90.6	94.5	85.3	89.7
		TG	90.4	92.3	87.4	89.8
N	BRNN	HD	93.3	93.0	94.1	93.5
		TG	83.1	77.7	96.6	86.1
	A-BRNN	HD	93.1	92.7	94.8	93.7
		TG	84.6	79.4	96.8	87.2

Table 2: *Specificity rate on non-speech segments with lip motion ('HD': close-talk microphone + HD camera; 'TG': sensors from tablet; 'C': clean sessions; 'N': noisy session).*

Approach	HD		TG	
	C	N	C	N
BRNN	94.2	93.1	94.7	89.0
A-BRNN	94.7	93.4	94.7	89.4

5.2. Non-speech Segments with Active Lip Motion

We also evaluate the robustness of the proposed approach to lip movements that are not associated to speech (smiles, lip-smack, deep breath). These segments are challenging scenarios since the lip movements are not related to speech activity. We manually identified 7,397 frames across speakers. We use the specificity rate, defined as true negative divided by the condition negative. The true negative represents the number of frames that is correctly classified among the selected non-speech frames. The condition negative represents the total number of non-speech frames selected by the models. Table 2 reports the results, which show that the A-BRNN model achieves the best results under all conditions.

6. Conclusions

This study extended the bimodal recurrent neural network using A-LSTM layers. The proposed framework takes advantage of the BRNN model, capturing the temporal dependencies within and across modalities. The addition of A-LSTM layers in the recurrent networks improves the temporal modeling of the models, which is important for speech activity detection. The model effectively captures longer periodic patterns associated with speech activity in the acoustic and visual features. We evaluate the proposed framework on the CRSS-4ENGLISH-14 corpus under difference channels and environmental conditions. The results on the SAD tasks show that the proposed approach can significantly outperform the BRNN implemented with the original LSTM layer for challenging conditions such as spontaneous speech or noisy conditions. We also evaluated the robustness of the proposed approach for non-speech segments with lip movements. The A-BRNN model achieves the best performance under all conditions.

The current implementation of our AV-SAD system only uses the A-LSTMs in one of the layers of the audio and visual RNNs (Fig. 2(b)). We expect that other configurations may lead to better performance. For example, we can replace all the LSTM layers for A-LSTM layers. We can consider more frames in the past (Equation 3). We can consider bidirectional A-LSTM. These alternative frameworks will require powerful resources to train the AV-SAD system.

7. References

[1] 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.

[2] 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.

[3] 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.

[4] F. Tao and C. Busso, "Bimodal recurrent neural network for audiovisual voice activity detection," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1938–1942.

[5] Y. Wang and F. Tian, "Recurrent residual learning for sequence classification," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, TX, USA, November 2016, pp. 938–943.

[6] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 3214–3218.

[7] F. Tao and G. Liu, "Advanced LSTM: a study about better time dependency modeling in emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 2906–2910.

[8] F. Tao, G. Liu, and Q. Zhao, "An ensemble framework of voice-based emotion recognition system for films and TV programs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 6209–6213.

[9] 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.

[10] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[11] 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.

[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] 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.

[14] 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.

[15] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[16] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[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, January 2018.

[22] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. to appear, 2018.

[23] 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.

[24] 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.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[26] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *International Conference on Machine Learning (ICML 2013)*, Atlanta, GA, USA, June 2013, pp. 1–9.

[27] F. Tao and C. Busso, "Aligning audiovisual features for audiovisual speech recognition," in *IEEE International Conference on Multimedia and Expo (ICME 2018)*, San Diego, CA, USA, July 2018.

[28] F. Chollet, "Keras: Deep learning library for Theano and TensorFlow," https://keras.io/, April 2017. [Online]. Available: https://github.com/fchollet/keras

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[30] V. M. Stanford, "NIST speech SNR tool," https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements, December 2005.

[31] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1286–1298, July 2018.