Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 00 (2020) 1–12

Speech Communi-cation

End-to-end Audiovisual Speech Activity Detection with Bimodal Recurrent Neural Models

Fei Tao, Carlos Busso

The University of Texas at Dallas

Abstract

Speech activity detection (SAD) plays an important role in current speech processing systems, including *automatic speech recognition* (ASR). SAD is particularly difficult in environments with acoustic noise. A practical solution is to incorporate visual information, increasing the robustness of the SAD approach. An audiovisual system has the advantage of being robust to different speech modes (e.g., whisper speech) or background noise. Recent advances in audiovisual speech processing using deep learning have opened opportunities to capture in a principled way the temporal relationships between acoustic and visual features. This study explores this idea proposing a *bimodal recurrent neural network* (BRNN) framework for SAD. The approach models the temporal dynamic of the sequential audiovisual data, improving the accuracy and robustness of the proposed SAD system. Instead of estimating hand-crafted features, the study investigates an end-to-end training approach, where acoustic and visual features are directly learned from the raw data during training. The experimental evaluation considers a large audiovisual corpus with over 60.8 hours of recordings, collected from 105 speakers. The results demonstrate that the proposed framework leads to absolute improvements up to 1.2% under practical scenarios over a VAD baseline using only audio implemented with *deep neural network* (DNN). The proposed approach achieves 92.7% F1-score when it is evaluated using the sensors from a portable tablet under noisy acoustic environment, which is only 1.0% lower than the performance obtained under ideal conditions (e.g., clean speech obtained with a high definition camera and a close-talking microphone).

Keywords: Audiovisual speech activity detection, end-to-end speech framework, deep learning, recurrent neural network

1. Introduction

The success of voice assistant products including Siri, Google Assistant, and Cortana has made the use of speech technology more widespread in our life. These interfaces rely on several speech processing tasks, including *speech activity detection* (SAD). SAD is a very important pre-processing step, especially for interfaces without push-to-talk button. The accuracy of a SAD system directly affects the performance of other speech processing technologies including *automatic speech recognition* (ASR) [1], speaker verification and identification [2], speech enhancement, speech emotion recognition [3, 4], speech separation [5] and speaker tracking [6, 7, 8, 9]. A key challenge for SAD is the environmental noise observed in real world applications, which can greatly affect the performance of the speech interface, especially if the SAD models are built with energy-based features [10, 11].

An appealing way to increase the robustness of a SAD system against acoustic noise is to include visual features [12, 13, 14, 15], mimicking the multimodal nature of speech perception used during daily human interactions [16, 17]. While this solution is not always practical for all applications, the use of portable devices with camera and the advances of *human-robot interaction* (HRI) make an *audiovisual speech activity detection* (AV-SAD) system a suitable solution. Noisy environment leads speakers to affect their articulatory production to increase their speech intelligibility, a phenomenon known as Lombard effect. While studies have reported differences in visual features between clean and noisy environments, these variations are not as pronounced as the differences in acoustic features [18]. Therefore, visual feature are more robust to acoustic noise. For example, Tao et al. [19] showed that a *visual speech activity detection* (V-SAD) system can achieve robust performance under clean and noisy conditions using the camera of a portable device.

Conventional approaches to integrate acoustic and visual information in SAD tasks have relied on ad-hoc fusion schemes such as logic operation, feature concatenation or pre-defined rules [20, 21, 22, 23]. These approaches oversimplify the relationship between audio and visual modalities, which may lead

to rigid models that cannot capture the temporal dynamic between these modalities. Recent advances on *deep neural network* (DNN) have provided new data-driven frameworks to appropriately model sequential data [24, 25]. These models avoid defining predefined rules or making unnecessary assumptions by directly learning relationships and distributions from the data [26]. Recent studies on audiovisual speech processing have demonstrated the potential of *deep learning* (DL) in this area [13, 27, 28, 29, 30]. A straight forward extension from conventional approaches is concatenating audiovisual features as the input for a DNN [13, 26]. Another way is to rely on auto-encoder to extract bottleneck audiovisual representations [31]. However, these methods do not directly capture the temporal relationship between acoustic and visual features. Furthermore, the systems still rely on hand-crafted features, which may not lead to optimal systems.

This study proposes an end-to-end framework for AV-SAD that explicitly captures the temporal dynamic between acoustic and visual features. The approach builds upon the framework presented in our preliminary work [32, 33], which relies on *recurrent neural networks* (RNNs). Our approach, referred to as *bimodal recurrent neural network* (BRNN), consists of three subsystems. The first two subsystems independently process the modalities using RNNs, creating an acoustic RNN and a visual RNN. These subsystems are implemented with *long short-term memory* (LSTM) layers, and their objective is to capture the temporal relationship within each modality that are discriminative for speech activity. These subsystems provide high level representations for the modalities, which are concatenated and fed as an input vector to a third subsystem. This system, also implemented with LSTMs, predicts the speech/non-speech label for each frame, capturing the temporal information across the modalities.

An important contribution of this study is that the acoustic and visual features are directly learned from the data. Recent advances in DNN for speech processing tasks have shown the benefits of learning discriminative features as part of the training process, using *convolutional neural network* (CNN) and sequence modeling with RNN [34, 35]. We can learn end-to-end system with this approach, which has led to performance improvements over hand-crafted features in many cases [29, 36]. Furthermore, we can capture the characteristics of the raw input data and extract discriminative representation for a target task [37, 38]. These observations motivate us to learn discriminate features from the data. The inputs of the BRNN framework are the raw image around the orofacial area as visual features, and the Mel-filterbank as acoustic features. For the visual input, we use three 2D convolutional layers to extract high-level representation from the raw image around the mouth area. On top of the convolutional layers, we use LSTM layers to model temporal information. For the acoustic input, we use *fully connected* (FC) layers that are connected to LSTM layers to model the temporal evolution of the data, similar to the visual part. The proposed approach is jointly trained learning discriminative features from the data, creating an effective and robust end-to-end AV-SAD system.

We evaluate our framework on a subset of the CRSS-4English-14 corpus consisting of over 60h of recordings from 105 speakers. The corpus includes multiple sensors, which allows us to evaluate the proposed approach under ideal channels (i.e., close-taking microphone, high definition camera) or under more practical channels (i.e., camera and microphone from a portable tablet). The corpus also has noisy sessions where different types of noise were played during the recordings. The various conditions can mimic practical scenarios for speech-based interfaces. We replicate state-of-the-art supervised SAD approaches proposed in previous studies to demonstrate the superior performance of the proposed approach. The experimental evaluation shows that our end-to-end BRNN approach achieves the best performance under all conditions. The proposed approach can achieve at least 0.6% absolute improvement compared to the state-of-the-art A-VAD system. Among the AV-SAD systems, the proposed approach outperforms the best baseline by 1.0% in the most challenging scenario corresponding to sensors from a portable device under noisy environment. This result for this condition is 1.2% higher than an A-SAD system, providing clear benefits of the proposed audiovisual solution for SAD.

The paper is organized as following. Section 2 reviews previous studies on AV-SAD, describing the differences with our approach, and highlighting our contributions. Section 3 describes the CRSS-4English-14 corpus and the post-processing steps to use the recordings. Section 4 introduces our proposed end-to-end BRNN framework. Section 5 presents the experimental evaluations that demonstrate the benefits of our approach. The paper concludes with Section 6, which summarizes our study and discusses potential future directions in this area.

2. Related Work

A successful SAD system can have a direct impact on ASR performance by correctly identifying speech segments. While speech-based SAD systems have been extensively investigated, SAD systems based on visual features are still under development.

Visual information describing lip motion provides valuable cues to determine the presence or absence of speech. Several studies have relied on visual features to detect speech activity in speech-based interfaces [14, 39, 40]. These methods either rely exclusively on visual features (V-SAD) [15, 19, 41, 42, 43, 44, 45], or complement acoustic features with visual cues [20, 21, 46]. As an emerging research topic, several methods have been proposed. Early studies relied on *Gaussian mixture models* (GMM) [41, 47], *hidden Markov models* (HMMs) [45, 20], or static classifiers such as *support vector machine* (SVM) [42]. Recent studies have demonstrated the potential of using deep learning for V-SAD and AV-SAD [31].

The use of deep learning offers better alternatives to fuse audiovisual modalities. Early studies relied on simple fusion schemes, including concatenating audiovisual features [21], combining the individual SAD decisions using basic "AND" or "OR" operations [22, 20], and creating hierarchical decision rules to assess which systems to use [23]. In contrast, DL solutions can incorporate in a principled manner the two modalities.

DL techniques can be used to build powerful data-driven frameworks, relying on the input data rather than rigid assumptions or rules [24, 25]. DL-based approaches provide better solutions for AV-SAD task compared with conventional approaches, increasing the flexibility of the systems.

The pioneer work of Ngiam et al. [26] demonstrated that DL can be a powerful tool for *audiovisual automatic speech recognition* (AV-ASR). For AV-SAD, however, there are only few studies using DL approaches. One exception is the approach proposed by Ariav et al. [31]. They used an autoencoder to create an audiovisual bottleneck representation. The acoustic and visual features were concatenated and used as input of the autoencoder. The bottleneck features from the autoencoder were used as input of a RNN, which aimed to detect speech activity. This approach separated the fusion of audiovisual features (autoencoder) from the classifier (RNN), which may lead to a suboptimal system where the bottleneck features are not optimized for the SAD task. To globally optimize the fusion and temporal modeling, Tao and Busso [32, 33] proposed the *bimodal recurrent neural network* (BRNN) framework for AV-SAD task. The approach used three RNNs as subnets following the structure in Figure 2. The framework can model the temporal information within and across modalities. The results show that this structure can outperform an RNN taking concatenated audiovisual features.

2.1. Features for Speech Activity Detection

For acoustic features, studies have relied on *Mel frequency cepstral coefficients* (MFCCs) [48], spectrum energy [10] and features describing speech periodicity [49]. However, there is no standard set for visual features, where studies have proposed several hand-crafted features. For example, Navarathna et al. [41] and Almajai and Milner [21] used appearance-based features such as 2D *discrete cosine transform* (DCT) coefficients from the orofacial area. Other studies have relied on geometric features [47, 23, 45, 43]. A common approach is to use *active appearance model* (AAM) [45]. Tao and Busso [50] and Neti et al. [51] suggested that appearance based features have the disadvantage of being more speaker dependent, so using geometric features can provide representations with better generalization. Some studies have combined appearance and geometric features [50].

Instead of using hand-crafted features, an appealing idea is to learn discriminative features from the data using end-to-end systems. The benefit of this approach is that the feature extraction and task modeling are jointly learned from the data, providing flexible and robust solutions. While this approach has been used in other areas, we are not aware of previous studies on end-to-end systems for AV-SAD. We hypothesize that this approach can lead to improvements in the performance, since feature representations obtained during the learning process have been shown to be effective on other tasks. CNN was originally used to learn features from images, since CNN can learn spatial, translation invariant representations from raw pixels [52]. The spatial invariance property in CNN allows the system to learn robust high-level representations from the input data [53]. Saitoh et al. [54] used CNN to extract a visual representation from concatenated images for *visual automatic speech recognition* (V-ASR) task. Petridis et al. [29] presented an end-to-end systems for V-ASR. They used raw images and their corresponding delta information as input to recognize words, relying on *bidirectional LSTMs* (BLSTMs). Amodei et al. [37], Sercu et al. [55] used CNN to extract high-level acoustic feature representations from raw acoustic data for *audio automatic speech recognition* (A-ASR) tasks. In these studies, FC layers were stacked over the CNN, mapping the representation extracted by the CNNs into the classification task space (Soltau et al. [56] showed the benefits of adding FC layers on top of CNN layers). Inspired by these studies, this study adopts a CNN as a feature extractor for visual features to learn high-level representations that are discriminative for AV-SAD tasks.

2.2. Temporal Modeling for Speech Activity Detection

Speech is characterized by semi-periodic patterns that are distinctive from non-speech sounds such as laugh, lip-smack, and deep breath. The temporal cues are observed not only on speech features, but also on orofacial features reflecting the articulatory movements needed to produce speech. Therefore, modeling temporal information is very important for SAD.

An approach to model temporal information is to include features that convey dynamic cues. A classic approach is by concatenating contiguous frames, creating contextual feature vectors [48, 41]. However, this approach relies on a predefined context window, which will constraint the capability of static frameworks such as DNN. Temporal information can also be added by taking temporal derivatives of the features [47, 21, 43], or by relying on optical flow features [44]. For example, Sodoyer et al. [43] demonstrated that dynamic features extracted by taking derivatives are more effective than the actual values describing the lip configuration. Takeuchi et al. [20] extracted the variance of optical flow as visual features to capture dynamic information.

An alternative, but complementary, approach to model temporal information is by using frameworks that capture recurrent connections. A common approach in speech processing tasks is the use of RNNs, which rely on connections between two contiguous time steps capturing temporal dependencies in sequential signals [57, 58, 59, 60]. Ariav et al. [31] used RNNs to capture temporal information in a AV-SAD task. A popular RNN framework is the use of LSTM units, which have been successfully used for AV-SAD task, showing competitive performance [32, 33]. Our proposed approach build on the *bimodal recurrent neural network* (BRNN) framework introduced in our previous studies [32, 33], which is described in Section 4.3.

2.3. Contributions of this Study

This study extends the BRNN framework proposed in Tao and Busso [32] by directly learning discriminative audiovisual features, creating an effective end-to-end AV-SAD system. While aspects of the BRNN framework were original presented in our preliminary work [32, 33], this study provides key contributions. A key difference between this study and our previous work is the features used for the task. While our preliminary

(a) Equipments

(b) Recording setting

Figure 1. Equipments and recording setting for the collection of the CRSS-4English-14 corpus. This study uses the audio from the close-taking and tablet microphones and the videos from the HD camera and the tablet.

studies relied on hand-craft audiovisual features, our method learns discriminative features directly from the data. To the best of our knowledge, this is the first end-to-end AV-SAD system.

Other key difference with our previous work is the experimental evaluation. The framework is exhaustively evaluated with other AV-SAD methods, obtaining state-of-the-art performance on a large audiovisual database. The study also demonstrates the benefits of using Mel-filterbank over speech spectrogram in the presence of acoustic noise.

3. The CRSS-4English-14 Audiovisual Corpus

This study uses the CRSS-4English-14 audiovisual corpus, which was collected by the *center of robust speech systems* (CRSS) at the University of Texas at Dallas. The corpus was described in details in Tao and Busso [13], so this section only describes the aspects of the corpus that are relevant to this study. Figure 1 shows the settings used to collect the corpus.

3.1. Data Collection

The CRSS-4English-14 corpus was collected in a $13ft \times 13ft$ *American Speech-Language-Hearing Association* (ASHA) certified sound booth. Figure 1 shows the recording setting. The audio was collected at 48 kHz with five microphones: a close-talking microphone (Shure Beta 53), a desktop microphone (Shure MX391/S), the bottom and top microphones of a cell-phone (Samsung Galaxy SIII), and a tablet (Samsung Galaxy Tab 10.1N). This study only uses the close-talking microphone, which was placed close to subject's mouth, and the microphone from the tablet, which was placed facing the subjects about two meters from them. The illumination was controlled with two LED light panels to collect high quality videos. The videos were collected with two cameras: a high definition (HD) camera (Sony HDR-XR100) and a camera from the tablet (Samsung Galaxy Tab 10.1N). This study uses the recordings from both cameras. The HD camera has a resolution of 1440×1080 at a frame rate of 29.97 fps. The tablet camera has a resolution of 1280×720 at a frame rate of 24 fps. Both cameras were placed facing the subjects about two meters from them, capturing frontal views of the head and shoulder of the subjects. A green screen was placed behind the subjects to create a smooth background. The participants were free to move their head and body during the data collection, without any constraint.

We used a computer screen about three meters from the subjects, presenting slides with the instructions for each task. The corpus includes read speech and spontaneous speech. For the read speech, we included different tasks such as single words (e.g., "yes"), city names (e.g., "Dallas, Texas"), short phrases or commands (e.g., "change probe"), connected digit sequences (e.g., "1,2,3,4"), questions (e.g. "How tall is the Mountain Everest"), and continuous sentences (e.g., "I'd like to see an action movie tonight, any recommendation?"). For the spontaneous speech, we proposed questions to the speakers, who are required to respond using spontaneous speech. The sentences for each of the tasks are randomly selected from a pool of options created for the data collection, so the content per speaker is balanced, but not the same. The data collection starts with the *clean session* where the speaker completed all the requested tasks (about 45 minutes). The clean session includes read and spontaneous speech. Then, we collected the *noisy session* for five minutes. We played pre-recorded noise using an audio speaker (Beolit 12), which was located about three meters from the subjects. The noise recordings were recorded in four different locations: restaurant, house, office and shopping mall. Playing noise during the recording rather than adding artificial noise after the recording is a strength of this corpus, as it includes speech production variations associated with Lombard effect (speakers adapt their speech production in the presence of acoustic noise). For the noisy session, the slides were randomly selected from the ones used in the clean section. However, we only considered slides with read speech, discarding slides used for spontaneous speech.

The corpus contains recording from 442 subjects with four English Accent: American (115), Australian (103), Indian (112) and Hispanic (112). All the subjects are asked to speak in English. This study only uses the subset of the corpus collected from American speakers. During the recordings, we had problem with the data from 10 subjects, so we only use data collected from 105 subjects (55 females and 50 males). The

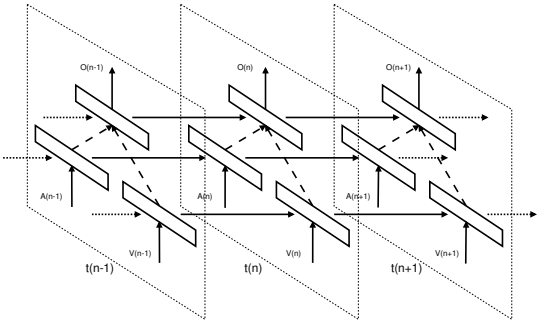


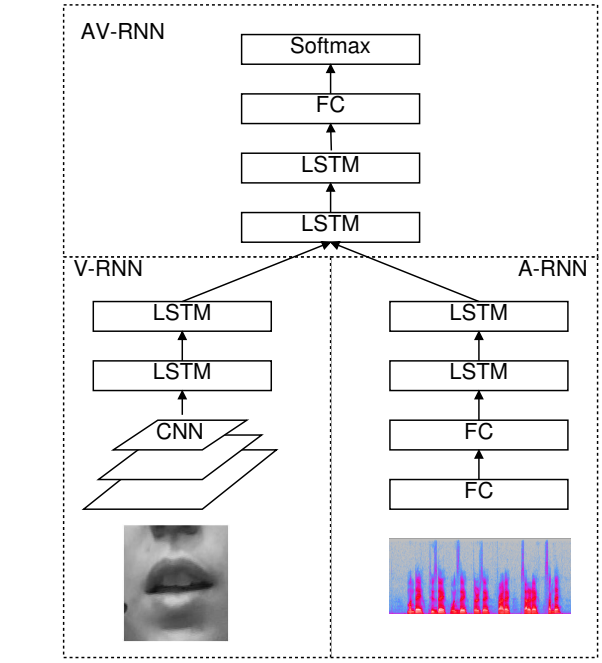
Figure 2. Diagram of the BRNN framework, which consists of three subnets implemented with RNNs. The A-RNN processes acoustic information, while the V-RNN processes visual information. The AV-RNN takes the concatenation of the outputs from the A-RNN and V-RNN as input, predicting the task label as output.

total duration of this set is 60 hours and 48 minutes. The size, variability in tasks, presence of clean and noisy sessions, and use of multiple devices make this corpus unique to evaluate our AV-SAD framework under different conditions.

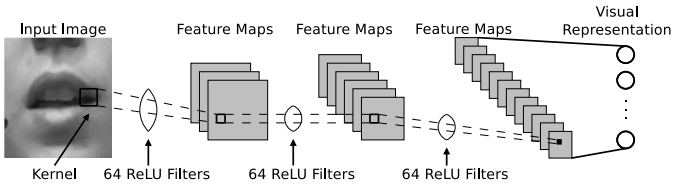
3.2. Data Processing

A bell ring was recorded as a marker every time the subjects switched slides. This signal was used to segment the recordings. We manually transcribed the corpus, using forced-alignment to identify speech and non-speech labels. For this task, we use the open-source software SAILAlign [61]. In the annotation and transcription of the speech, we annotate non-speech activities such as laughers, smacks, and coughs. We carefully remove these segments from frames labeled as 'speech'.

We resample the sampling rate of the videos collected with the tablet to match the sampling rate of the HD camera (e.g., 29.97 fps). We also resample the audio to 16kHz for both audio channels (close-talking microphone and tablet).



(a)



(b)

Figure 3. Detailed structure of the BRNN framework. (a) The structure of the proposed framework for one time frame. The A-RNN subnet includes FC and LSTM layers to process acoustic Mel-filterbank features. The V-RNN subnet has CNN extracting a visual representation from raw pixels and LSTMs to process temporal information. The AV-RNN subnet relies on FC and LSTM to process the concatenated output from the substructures A-RNN and V-RNN. (b) Detailed CNN configuration used to learn visual features.

4. Proposed Framework

In this study, we propose an end-to-end AV-SAD system building on the BRNN framework proposed in Fei and Busso [32]. Figure 2 describes the BRNN framework, which has three subnets implemented with RNN: an audio subnet, a video subnet and an audiovisual subnets. The audio and video subnets separately process each set of the features, capturing the temporal dependencies within modality that are relevant for SAD. The outputs from these two RNNs are concatenated and fed into a third subnet, fusing the information by capturing the temporal dependencies across modalities. This section describes the three subnets.

4.1. Audio Recurrent Neural Network (A-RNN)

The audio subnet corresponds to the *audio recurrent neural network* (A-RNN) and it is described in Figure 3(a). The A-RNN takes the acoustic features as input of a network consisting of static layers and dynamic layers (recurrent layers). The static layers model the input feature space, extracting the essential characteristics to determine speech activity. We rely on two *fully connected* (FC) layers. This study uses two max-out layers [62] rather than convolutional layers as static layers to reduce the computational complexity in training the models. Each layer has 512 neurons. The outputs of the FC layers are fed to dynamic layers to model the time dependencies within modality, as temporal patterns are important in SAD tasks [33]. We use two LSTM layers as dynamic layers. While bidirectional LSTMs have been used for this task [32], we only use unidirectional LSTMs to reduce the latency of the model, as our goal is to implement this approach in practical applications.

The acoustic feature used in our system are the Mel-filterbank features, which correspond to the energy in the frequency bands defined by the Mel scale filters. Therefore, it is a raw input feature that retains the main spectral information of the original speech. We use the tool *python speech features* to extract the Mel-filterbank features, using the default setting (25 ms window, 10 ms shifting step, and 26D filters in the Mel-filterbank). In this study, we concatenate 11 contiguous

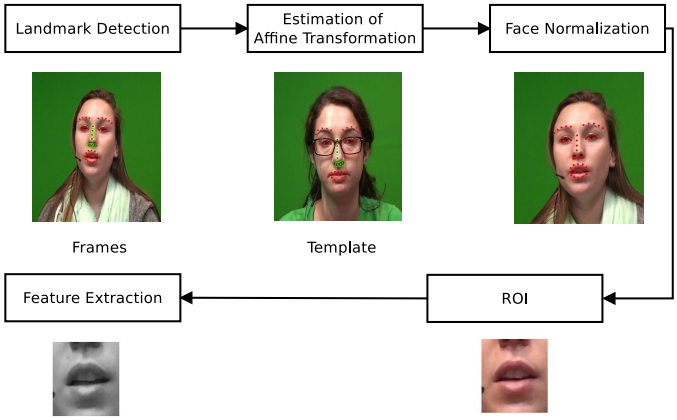


Figure 4. Flowchart to extract the ROI around the orofacial area. The solid dots are facial landmarks. The dots with circle are used to estimate the affine transformation for normalization. The ROI is determined after normalization. The final mouth image is resized to 32×32 and transformed into a gray scale.

frames as input, which includes 10 previous frames, in addition to the current frame. Concatenating previous frames improves the temporal modeling of the framework, while keeping the latency of the system low.

4.2. Visual Recurrent Neural Network (V-RNN)

The video subnet corresponds to the *video recurrent neural network* (C-RNN) and it is also described in Figure 3(a). The V-RNN takes raw images as visual feature. It extracts visual representation relying on convolutional layers. Figure 3(b) describes the details of the architecture. The convolutional layers capture visual patterns, such as edges, corners and texture from raw pixels based on local convolutions. The visual patterns can depict the mouth appearance and shape associated with speech activity. We stack three convolutional layers with *rectified linear units* (ReLUs) [63] to capture the visual patterns. Each layer has 64 filters. The kernel size is 5×5 and the stride is two (Fig. 3(b)). By using stride, we reduce the size of the feature map, so we do not need to use the pooling operation. The feature representation defined by the CNNs is a 64D feature vector. The implementation is intended to keep a compact network with lower hardware requirements and computation workload, which is ideal for practical applications. On top of the convolutional layers, we rely on two LSTM layers to further process the extracted visual representation, capturing the temporal information along time. Each layer has 64 neurons. Therefore, the V-RNN is able to directly extract both the visual patterns and temporal information from raw images that are relevant for speech articulation.

Figure 4 shows the flow chart of the visual feature extraction process used in this study. We manually pick a frame of a subject posing a frontal face as the template. We extract 49 facial landmarks from the template and each frame of the videos. The facial landmarks are extracted with the toolkit IntraFace [64]. IntraFace does not output a valid number when it fails to detect the landmarks, which occurred on some frames. We apply linear interpolation to predict these values when less than 10% of the frames of a video are missing. Otherwise, we discard the video. We apply an affine transformation to normalize the face by comparing the positions of nine facial points in the template and the current frame. This normalization compensates for the rotation and size of the face. These nine points are selected around the nose area, because they are more stable when the people are speaking (points highlighted on Fig. 4 describing the template). After face normalization, we compute the mouth centroid based on the landmarks around mouth. We downsample the *region of interest* (ROI) to 32×32 and convert it to gray scale colormap to limit the memory and computation workload.

4.3. Audiovisual Recurrent Neural Network (AV-RNN)

The high-level feature representations provided by the top layers of the A-RNN and V-RNN subnets are concatenated together and fed into the *audiovisual recurrent neural network* (AV-RNN) subnet (Fig. 3(a)). The proposed framework considers two LSTM layers to process the concatenated input. These LSTM layers aim to capture the temporal information across the modalities. On top of the LSTM layers, we include a FC layer implemented with maxout to further process the audiovisual representation. Each of the LSTM and FC layers are implemented with 512 neurons. The output is then sent to a softmax layer for classification, which determines whether the sample corresponds to a speech or non-speech segment.

The BRNN framework is designed to model the time dependency within single modality and across the modalities. The convolutional layers allow us to directly extract visual representation from raw images. The framework also directly obtain acoustic representation from Mel-filterbank features. The proposed BRNN is jointly trained, minimizing a common objective function (in this case, the cross-entropy function). This framework provides a powerful end-to-end system for SAD, as demonstrated with the experimental results.

5. Experiments and Results

5.1. Experiment Settings

We evaluate our proposed approach on the CRSS-4English-14 corpus (Sec. 3). We partition the corpus into train (data from 70 subjects), test (data from 25 subjects) and validation (data from 10 subjects) sets. All these sets are gender balanced. We use accuracy, recall rate, precision rate and F1-score as the performance metrics. The positive class to estimate precision and recall rates is *speech* (i.e., frames with speech activity). We estimate F1-score with Equation 1, combining precision and recall rates.

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

We consider the F1-score as the main metric to compare alternative methods.We separately compute the results for each of the 25 subjects in the test set, reporting the average performance across individuals. We perform one-tailed t-test on the average of the F1-scores to determine if one method is statistically better than the other, asserting significance at p-value=0.05.

The experiments were conducted with a GPU using the Nvidia GeForce GTX 1070 graphic card (8GB graphic memory).

Since all of the three baseline approaches rely on deep learning techniques, we apply the same training strategy (Sec. 5.2). We use dropout with p= 0.1 to improve the generalization of the models. We use ADAM optimizer [65], monitoring the loss on the validation set. We rely on early stopping when the validation loss stops decreasing.

5.2. Baseline Methods

We aim to compare the proposed approach with state-of-the-art methods proposed for SAD. The study considers five baselines. Three approaches rely only on acoustic features. Two of them are implemented with conventional methods (SVM and GMM), and one is implemented with deep learning solutions. By using conventional approaches, we aim to show the effectiveness of applying deep learning techniques for this problem. Two approaches rely on audiovisual features. Our proposed approach uses unidirectional LSTM instead of BLSTM to reduce the latency in the model, which is a key feature for practical applications. To make the comparison fair, we also implement the baselines using deep learning with unidirectional LSTMs.

5.2.1. Conventional non-DNN A-SAD

The first two baselines are implemented with conventional A-SAD frameworks that are not implemented with deep learning algorithms. The first system relies on SVM trained with Mel-filterbank feature. The classifier is trained with a linear kernel. We refer to this system as "SVM-SAD".

The second system is the Combo-SAD system, which is a state-of-the-art unsupervised A-SAD proposed by Sadjadi and Hansen [49]. The system takes five acoustic features as inputs: harmonicity, clarity, prediction gain, periodicity and perceptual spectral flux. These features capture key speech properties that are discriminative of speech segments such as periodicity and slow spectral fluctuations. *Harmonicity*, also called *Harmonics-to-Noise Ratio* (HNR), measures the relative value of the maximum autocorrelation peak, which produces high peaks for voiced segments. *Clarity* is defined as the relative depth of the minimum *average magnitude difference function* (AMDF) valley in the possible pitch range. This metric also leads to large values in the voiced segments. *Prediction gain* corresponds to the energy ratio between the original signal and the *linear prediction* (LP) residual signal. It will also show higher values for voiced segments. *Periodicity* is a frequency domain feature based on the *harmonic product spectrum* (HPS). *Perceptual spectral flux* captures the quasi-stationary feature of the voice activity, as the spectral properties of speech do not change as quickly as non-speech segments or noise. The details of these features are explained in Sadjadi and Hansen [49]. These five features are combined using *principle component analysis* (PCA), where the resulting signal is the first principal component referred to as the 1-D "combo" feature. A threshold is then obtained by fitting a GMM with two mixtures using expectation maximization under the assumption that one mixture represents speech and the other mixture represents silence. We name this baseline as "Sadjadi-2013".

5.2.2. A-SAD using DNN [48]

The third baseline corresponds to the A-SAD framework proposed by Ryant et al. [48], which we denote "Ryant-2013". This approach is a state-of-the-art supervised A-SAD system using DNN. The system has four fully connected layers with 256 maxout neurons per layer. On top of the four layers, the system has a 2-class softmax layer for SAD classification. This approach uses 13D MFCCs. We concatenate 11 feature frames as input to make the system comparable with the proposed approach.

5.2.3. AV-SAD system using BRNN [32]

The fourth baseline is the AV-SAD system proposed by Tao and Busso [32], which we denote "Tao-2017". This framework is a state-of-the-art AV-SAD system, relying on BRNN. The network is similar to the approach presented in this paper (Fig. 2). The key difference is the audiovisual features, which correspond to hand-crafted.

The acoustic features correspond to the five features proposed by Sadjadi and Hansen [49] for A-SAD: harmonicity, clarity, prediction gain, periodicity and perceptual spectral flux (Sec. 5.2.1). The visual features include geometric and optical flow features describing orofacial movements characteristic of speech articulation. We extract 26D visual features from the ROI shown in Figure 4. This vector is created as follows. First, we extract a 7D feature vector from the ROI (three optical flow features, and four geometric features). The optical flow features consists of the variance of the optical flow in the vertical and horizontal direction within the ROI. The third optical flow feature corresponds to the summation of the variance in both direction, which provides the overall temporal dynamic on the frame. The four geometric features include the width, height, perimeter and area of the mouth. Based on the 7D feature vector, we compute three statistics over short-term window: variance, *zero crossing rate* (ZCR) and *speech periodic characteristic* (SPC) (details are introduced in Tao et al. [22]). The short-term window is shifted one frame at a time. We set its size equal to nine frames (about 0.3s) to balance the trade off between resolution (it requires short window) and robust estimation (it requires long window). The three statistics estimated over the 7D vector results in a 21D feature vector. We append the summation of the optical flow variances and the first order derivative of the 4D geometric feature to the 21D vector, since they can also provide dynamic information. The final visual feature is, therefore, a 26D vector. All the visual features are z-normalized at the utterance level.

We concatenate 11 audio feature frames as audio input, and use 1 visual feature frame as visual input. The subnet processing the audio features has four layers, each of them implemented with 256 neurons. The first two layers are maxout neuron layers and the other two layers are LSTM layers. The subnet processing the video features has four layers, each of them implemented with 64 neurons. The first two layers are maxout neuron layers and the last two layers are LSTM layers. The hidden values from the top layers of the two subnets are concatenated and fed to the third subnet, which has four layers. The first two layers are LSTM layers with 512 neurons.

The third layer is implemented with maxout neurons with 512 neurons. The last layer is the softmax layer for classification.

5.2.4. AV-SAD System using Autoencoder [31]

The fifth baseline corresponds to the AV-SAD approach proposed by Ariav et al. [31], which relies on autoencoder (Section 2 describes this approach). We refer to this method as "Ariav-2018". 13D MFCCs are used as audio feature, and optical flow over the frame is used as visual feature. There are two stages in this approach. The first stage is the feature fusion stage which relies on an autoencoder. We implement this approach by concatenating 11 audio feature frames and one visual feature frame. The concatenated features are used as input for a five-layer autoencoder. The middle layer has 64 neurons, and the other layers have 256 neurons. The hidden values extracted from the middle layers are used as bottleneck features. All the layers use maxout neurons. The second stage uses the bottleneck features as the input of a four-layer RNN. The first two layers are LSTM layers, with 256 neurons per layer. The third layer is a maxout layer with 256 neurons. The last layer is the softmax layer for classification.

5.3. Experimental Results

While the CRSS-4English-14 corpus has several recording devices, this study only considers two combinations. The *ideal channels* use the data collected with the close-talking microphone and the HD camera, which have the best quality. The *practical channels* consider the video and audio recordings collected with the tablet. We expect that these sensors are good representations of the sensors used in practical speech-based interfaces. In addition, there are two types of environment as described in Section 3.1: clean and noisy audio recordings. Altogether, this evaluation considers four testing conditions (ideal channel+clean; ideal channel+noise; practical channel+clean; practical channel+noise). All the models are trained with the ideal channels under clear recordings, so the other conditions create train-test mismatches. We are interested in evaluating the robustness of the approaches under these channel and/or environment mismatches.

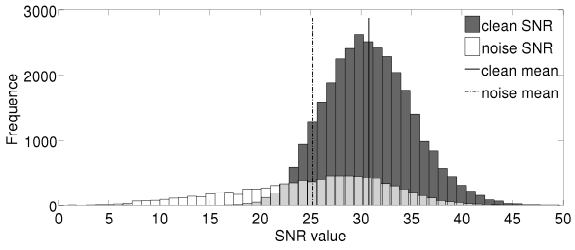
The first six rows of Table 1 shows the performance for the ideal channel under clean audio recordings. The proposed approach can outperform the baseline approaches ("SVM-SAD" by 6.0%, "Sadjadi-2013" by 7.4%, "Ryant-2013" by 0.6%; "Tao-2017" by 4.5%; "Ariav-2018" by 0.5%). The differences between our framework and "Tao-2017" are statistically significant (p-value > 0.05). "Tao-2017" used hand-crafted features. Our end-to-end BRNN system achieves better performance, which demonstrate the benefits of learning the features from the raw data. The proposed approach perform slightly better than the baselines 'Ryant-2013" and "Ariav-2018", although the differences are not statistically significant. The proposed approach is also significantly better than the conventional SAD systems that do not rely on deep learning ("SVM-SAD", and "Sadjadi-2013").

The last six rows of Table 1 shows the performance under noisy audio recordings. The proposed approach can outperform the baselines ("SVM-SAD" by 6.6%, "Sadjadi-2013" by 7.9%, "Ryant-2013" by 0.7%; "Tao-2017" by 2.3%; "Ariav-2018" by 1.0%). The differences are statistically significant (p-value<0.05) when our approach is compared with the "SVM-SAD", "Sadjadi-2013", "Tao-2017" and "Ariav-2018" systems. The classification improvement over the "Ariav-2018" system demonstrates that the BRNN framework combine better the modalities than the autoencorder framework, especially in the presence of noise. The proposed BRNN structure jointly learns how to extract the features and fuse the modalities, improving the temporal modeling of the system. Table 1 shows that the system tested with ideal channels has better performance under noisy conditions than under clean conditions. This unintuitive result is due to two reasons. First, the *signal-to-noise ratios* (SNRs) under noisy and clean conditions are very similar for the ideal channels since the microphone is close to the subject's mouth and far from the audio speaker playing the noise. Figure 5 shows the distribution of the predicted SNR, using the NIST Speech SNR Toolkit [66]. The Figure 5(a) shows an important overlap between both conditions. Second, we only have read speech in the noisy section. In addition to read speech, the clean section also has spontaneous speech, which is a more difficult task for SAD.

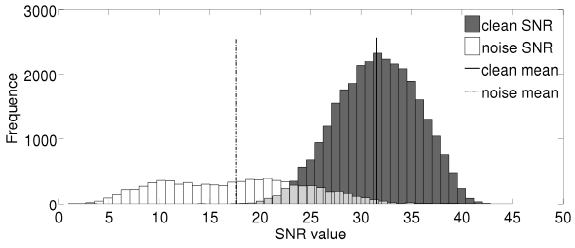
Table 1. Performance of the SAD systems for the ideal channels (close-talking microphone, HD camera). "Env" stands for testing environment ("C" is clean; "N" is noisy). "Modality" stands for modality used by the approach ("A" is A-SAD, "AV" is AV-SAD). "Approach" stands for corresponding framework ("Acc:" accuracy; "Pre:" precision rate; "Rec:" recall rate;"F:" F1-score).

Env	Modality	Approach	Acc	Pre	Rec	F
C	A	SVM-SAD	88.4	97.2	80.3	88.0
	A	Sadjadi-2013	87.2	97.6	77.8	86.6
	A	Ryant-2013	90.3	96.6	90.5	93.4
	AV	Tao-2017	90.1	94.6	84.8	89.5
	AV	Ariav-2018	93.4	95.4	91.7	93.5
	AV	**Proposed**	93.9	95.8	92.3	94.0
N	A	SVM-SAD	88.8	96.8	82.5	89.1
	A	Sadjadi-2013	87.7	97.6	79.8	87.8
	A	Ryant-2013	94.8	96.4	93.8	95.0
	AV	Tao-2017	93.3	93.1	94.0	93.4
	AV	Ariav-2018	94.4	95.4	94.1	94.7
	AV	**Proposed**	95.3	96.2	95.2	95.7

Table 2 presents the results for the practical channels, which shows that our approach also achieves better performance than the baseline methods across conditions. For clean audio recordings, the proposed approach can significantly outperforms all the baselines ("SVM-SAD" by 6.6%, "Sadjadi-2013" by 7.9%,"Ryant-2013" by 0.8%; "Tao-2017" by 4.3%; "Ariav-2018" by 0.8%). For noisy audio recordings, we observe that the performances drop across conditions compared to the results obtained under clean recordings. The microphone of the tablet is closer to the audio speaker playing the noise, so the SNR is lower (Fig. 5(b)). The proposed approach can maintain a 92.7% F1-score performance, outperforming all the baseline frameworks ("SVM-SAD" by 12.7%, "Sadjadi-2013" by 11.8%, "Ryant-2013" by 1.2%; "Tao-2017" by 6.7%; "Ariav-2018" by 1.0%). The differences are statistically significant for



(a) Ideal channels



(b) Practical channels

Figure 5. Distributions of the SNR predictions for the ideal and practical channels. The SNR prediction are estimated with the NIST Speech SNR Toolkit [66](CRSS-4English-14 corpus). For the noisy audio recordings, the microphone in the tablet was closer to the audio speaker playing the noise, so the microphone of the practical channels is more affected by the noise.

all the baselines. This result shows that the proposed end-to-end BRNN framework can extract audiovisual feature representations that are robust against noisy audio recordings.

It can also be noticed that the DNN-based A-SAD ("Ryant-2013") is better than the conventional non-DNN approaches ("SVM-SAD" and "Sadjadi-2013") under all conditions. The DNN-based A-SAD ("Ryant-2013") outperforms the best conventional approach ("Sadjadi-2013") with absolute differences in F1-score as large as 10.6%. The results of conventional non-DNN baselines are also significantly worse than our proposed audiovisual approach. These results confirm that the DNN-based approaches have better robustness and modeling capabilities for SAD problems. In the rest of the evaluation, we only consider approaches relying on DNN.

5.4. BRNN Implemented with Different Acoustic Features

We also re-implement the proposed approach with alternatives acoustic features to demonstrate the benefits of using Mel-filterbank features. The first acoustic features considered in this section is the spectrogram features without using the Mel filters. We extract 320D features using a Turkey filter with uniform bins between 0-8KHz. The second acoustic features correspond to the 5D hand-crafted acoustic features proposed by Sadjadi and Hansen [49], which we describe in Section 5.2.1. In both cases, we concatenate 10 previous frames to the current frame to create a contextual window, following the approach used for the Mel-filterbank features. For the spectrogram, the A-RNN subnet has four layers, each of them implemented with 4,096 neurons. For the 5D hand-craft features, the A-RNN subnet has four layers, each of them implemented with 256 neurons. The configuration for the rest of the framework is consistent with the proposed approach, including the A-RNN and AV-RNN.

Table 2. Performance of the SAD systems for the practical channels (microphone and camera from the tablet). "Env" stands for testing environment ("C" is clean; "N" is noisy). "Modality" stands for modality used by the approach ("A" is A-SAD, "AV" is AV-SAD). "Approach" stands for corresponding framework ("Acc:" accuracy; "Pre:" precision rate; "Rec:" recall rate;"F:" F1-score).

Env	Modality	Approach	Acc	Pre	Rec	F
C	A	SVM-SAD	87.5	97.1	78.9	87.1
	A	Sadjadi-2013	86.3	97.2	76.4	85.6
	A	Ryant-2013	92.7	94.3	91.6	92.9
	AV	Tao-2017	90.0	91.9	87.3	89.4
	AV	Ariav-2018	92.8	95.2	90.8	92.9
	AV	**Proposed**	93.4	95.4	92.0	93.7
N	A	SVM-SAD	78.9	84.1	76.3	80.0
	A	Sadjadi-2013	80.5	88.2	74.7	80.9
	A	Ryant-2013	90.8	90.6	92.5	91.5
	AV	Tao-2017	83.3	77.5	96.7	86.0
	AV	Ariav-2018	91.2	92.9	90.6	91.7
	AV	**Proposed**	92.1	92.9	92.6	92.7

Table 3. Performance of the BRNN framework implemented with different acoustic features. "CH" stands for channel. "Env" stands for testing environment ("C" is clean; "N" is noisy). "Feature" stands for acoustic feature used in the evaluation ("Acc:" accuracy; "Pre:" precision rate; "Rec:" recall rate;"F:" F1-score).

CH	Env	Feature	Acc	Pre	Rec	F
Ideal	C	Mel-filterbank	93.8	95.8	92.3	94.0
		Spectrogram	93.4	94.8	93.1	93.9
		Hand-crafted [49]	92.2	94.0	90.4	92.2
Practical	N	Mel-filterbank	92.1	92.9	92.6	92.7
		Spectrogram	76.8	71.3	97.4	82.2
		Hand-crafted [49]	66.9	64.3	88.1	74.3

The evaluation only considers two conditions: ideal channels with clean audio recordings, and practical channels with noisy audio recordings. These two conditions represent the easiest and hardest settings considered in this study, respectively.

Table 3 presents the results. For the ideal channels under clean audio recordings, using a feature representation learnt from Mel-filterbank is slightly better than using a representation learnt from the spectrogram. Both of these feature representations lead to significantly better performance than the system trained with hand-crafted features. Learning flexible feature representations from the raw data lead to better performance than using hand-crafted features, as they are not constrained by pre-defined rules or assumptions. For the practical channels under noisy audio recordings, the model trained with hand-crafted features achieve the worse performance. The feature representation learnt from the Mel-filterbank is able to significantly outperforms the representation learnt from the spectrogram by a large margin (10.5%). The feature representation learnt from the spectrogram is more sensitive to acoustic noise.

The experiments in this section show that learning feature representations from Mel-filterbank leads to better performance across conditions, showing competitive results under clear and noisy speech.

Table 4. Performance for unimodal SAD systems and the bimodal SAD system. 'CH" stands for channel. "Env" stands for testing environment ("C" is clean; "N" is noisy). "Modality" stands for modality used by the approach ("Acc:" accuracy; "Pre:" precision rate; "Rec:" recall rate;"F:" F1-score).

CH	Env	Modality	Acc	Pre	Rec	F
Ideal	C	Bimodal	93.8	95.8	92.3	94.0
		Audio	92.7	94.5	91.4	92.8
		Video	60.0	65.3	50.9	57.2
Practical	N	Bimodal	92.1	92.9	92.6	92.7
		Audio	90.3	89.2	93.9	91.5
		Video	65.5	69.3	68.5	68.9

5.5. Performance of Unimodal Systems

We also explore the performance of SAD systems trained with unimodal features to highlight the benefits of using audiovisual information. The experimental setup uses the V-RNN and A-RNN modules of the BRNN framework (Fig. 3(a)). For the audio-based system, we use the pre-trained A-RNN models. The weights of this subnet are not modified. On top of the A-RNN, we implement the same structure used in the BRNN consisting of two LSTM layers, one FC layer, and a softmax layer (Fig. 3(a)). We train these four layers from scratch, using the same training scheme used to train the BRNN network (i.e., dropout, ADAM, early stopping). The visual-based system is trained using the same strategy, starting with the V-RNN subnet. Similar to Section 5.4, the evaluation only considers the ideal channels with clean audio recordings, and practical channels with noisy audio recordings.

Table 4 shows the performance for the audio-based and visual-based systems. For comparison, we also include the proposed audiovisual BRNN approach. For the ideal channels under clean audio recordings, the bimodal system can outperform the unimodal systems, showing the benefits of using audiovisual features. The result from the audio-based system is 35.6% (absolute) better than the result from the video-based systems. This result is consistent with findings from previous study [32, 22, 19]. In spite of the lower performance of the visual-based system, the addition of orofacial features lead to clear improvements in the BRNN system. For noisy channels under noisy audio recordings, the bimodal system still achieves the best performance, outperforming the unimodal systems where the differences are statistically significant. The audio-based system achieves better results than the visual-based system. The performance for the visual-based system using noisy audio recordings is higher than the results obtained with clean audio recordings. This result is explained due to two reasons: (1) the visual features are not greatly affected by the background acoustic noise, and (2) the data for noisy audio recordings does not contain spontaneous speech, as explained in Section 5.3. If we include only read sentences recorded in both noisy and clear recordings, the performance of the visual-based system trained with the ideal channels under clean audio recordings is 69.2%. This result is slightly higher than the value reported in Table 4 for visual-based system under noisy audio recordings.

The comparison between bimodal and unimodal inputs shows the benefit of using bimodal features. It highlights that our proposed BRNN approach can achieve better performance than state-of-the-art unimodal SAD systems.

6. Conclusion and Future Work

This study proposed an end-to-end AV-SAD framework where the acoustic and visual features are directly learnt during the training process. The proposed approach relies on LSTM layers to capture temporal dependencies within and across modalities. This objective is achieved with three subnets. The first two subnets separately learn visual and acoustic representations that are discriminative for SAD tasks. The visual subnet uses CNNs to learn features directly from images of the orofacial area. The audio subnet extracts acoustic representation directly from Mel-filterbank features. The outputs of these subnets are concatenated and used as input of a third subnet, which models the temporal dependencies across modalities. Instead of using BLSTM, the proposed framework relies on unidirectional LSTM, reducing the latency, and, therefore, increasing the usability of the system in real applications. To the best of our knowledge, this is the first end-to-end AV-SAD system.

We evaluated the proposed approach on a set of the CRSS-4English-14 corpus (105 speakers), which is a large audiovisual corpus. The proposed approach outperformed alternative state-of-the-art A-SAD and AV-SAD systems. We observed consistent improvements across conditions. The proposed end-to-end BRNN framework maintained good performance in the presence of different noise and channel conditions. The system also achieved better performance than an implementation of the BRNN system using audiovisual hand-crafted features. These results demonstrated the benefits of learning feature representations during the training process. This approach provides an appealing solutions for practical applications.

There are several research directions to extend the proposed approach. This study only focused on acoustic noise. In the future, we will evaluate the framework in the presence of visual artifacts (e.g., blurred images, occlusions). Likewise, the proposed approach assumes that the audiovisual modalities are available. We are exploring alternative solutions to address missing information. Finally, we leave as a future work to learn acoustic representations with CNNs. This direction was not pursued on this study due to the high computational cost and memory requirements needed to train the models.

Acknowledgment

This study was funded by the National Science Foundation (NSF) grants IIS-1718944 and IIS-1453781 (CAREER).

References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine 29 (6) (2012) 82–97, doi:\bibinfo{doi}{10.1109/MSP.2012.2205597}.

[2] G. Liu, Q. Qian, Z. Wang, Q. Zhao, T. Wang, H. Li, J. Xue, S. Zhu, R. Jin, T. Zhao, The Opensesame NIST 2016 Speaker Recognition Evaluation System, in: Interspeech 2017, Stockholm, Sweden, 2854–2858, doi:\bibinfo{doi}{10.21437/Interspeech.2017-997}, 2017.

[3] S. Parthasarathy, C. Busso, Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning, in: Interspeech 2017, Stockholm, Sweden, 1103–1107, doi:\bibinfo{doi}{10.21437/Interspeech.2017-1494}, 2017.

[4] F. Tao, G. Liu, 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, 6209–6213, doi:\bibinfo{doi}{10.1109/ICASSP.2018.8461617}, 2018.

[5] B. Rivet, W. Wang, S. M. Naqvi, J. A. Chambers, Audiovisual speech source separation: An overview of key methodologies, IEEE Signal Processing Magazine 31 (3) (2014) 125–134, doi:\bibinfo{doi}{10.1109/MSP.2013.2296173}.

[6] Q. Liu, W. Wang, T. de Campos, P. J. B. Jackson, A. Hilton, Multiple Speaker Tracking in Spatial Audio via PHD Filtering and Depth-Audio Fusion, IEEE Transactions on Multimedia 20 (7) (2018) 1767–1780, doi:\bibinfo{doi}{10.1109/TMM.2017.2777671}.

[7] V. Kılıç, M. Barnard, W. Wang, A. Hilton, J. Kittler, Mean-shift and sparse sampling-based SMC-PHD filtering for audio informed visual speaker tracking, IEEE Transactions on Multimedia 18 (12) (2016) 2417–2431, doi:\bibinfo{doi}{10.1109/TMM.2016.2599150}.

[8] V. Kılıç, M. Barnard, W. Wang, J. Kittler, Audio assisted robust visual tracking with adaptive particle filtering, IEEE Transactions on Multimedia 17 (2) (2015) 186–200, doi:\bibinfo{doi}{10.1109/TMM.2014.2377515}.

[9] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. Naqvi, J. Chambers, Robust multi-speaker tracking via dictionary learning and identity modeling, IEEE Transactions on Multimedia 16 (3) (2014) 864–880, doi:\bibinfo{doi}{10.1109/TMM.2014.2301977}.

[10] J. Pang, Spectrum energy based voice activity detection, in: IEEE Annual Computing and Communication Workshop and Conference (CCWC 2017), Las Vegas, NV, USA, 1–5, doi:\bibinfo{doi}{10.1109/CCWC.2017.7868454}, 2017.

[11] M. Moattar, M. Homayounpour, A simple but efficient real-time voice activity detection algorithm, in: European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, 2549–2553, 2009.

[12] G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, S. Thermos, Audio and visual modality combination in speech processing applications, in: S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, A. Krüger (Eds.), The Handbook of Multimodal-Multisensor Interfaces, volume 1, vol. 1, ACM Books, ISBN 978-1970001648, 489–543, doi:\bibinfo{doi}{10.1145/3015783.3015797}, 2017.

[13] F. Tao, C. Busso, Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (7) (2018) 1286–1298, doi:\bibinfo{doi}{10.1109/TASLP.2018.2815268}.

[14] Q. Liu, A. J. Aubrey, W. Wang, Interference Reduction in Reverberant Speech Separation With Visual Voice Activity Detection, IEEE Transactions on Multimedia 16 (6) (2014) 1610–1623, doi:\bibinfo{doi}{10.1109/TMM.2014.2322824}.

[15] Q. Liu, W. Wang, P. Jackson, A visual voice activity detection method with Adaboosting, in: Sensor Signal Processing for Defence (SSPD 2011), London, UK, 1–5, doi:\bibinfo{doi}{10.1049/ic.2011.0145}, 2011.

[16] J. Keil, N. Müller, N. Ihssen, N. Weisz, On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states, Cerebral Cortex 22 (1) (2012) 221–231, doi:\bibinfo{doi}{10.1093/cercor/bhr125}.

[17] K. Van Engen, Z. Xie, B. Chandrasekaran, Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect, Attention, Perception, & Psychophysics 79 (2) (2017) 396–403, doi:\bibinfo{doi}{10.3758/s13414-016-1238-9}.

[18] T. Tran, S. Mariooryad, C. Busso, Audiovisual corpus to analyze whisper speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), Vancouver, BC, Canada, 8101–8105, doi:\bibinfo{doi}{10.1109/ICASSP.2013.6639243}, 2013.

[19] F. Tao, J. Hansen, C. Busso, An Unsupervised Visual-only Voice Activity Detection Approach Using Temporal Orofacial Features, in: Interspeech 2015, Dresden, Germany, 2302–2306, 2015.

[20] S. Takeuchi, T. Hashiba, S. Tamura, 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, 151–154, 2009.

[21] I. Almajai, 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, 1–5, 2008.

[22] F. Tao, J. L. Hansen, C. Busso, Improving Boundary Estimation in Audiovisual Speech Activity Detection Using Bayesian Information Criterion, in: Interspeech 2016, San Francisco, CA, USA, 2130–2134, doi:\bibinfo{doi}{10.21437/Interspeech.2016-406}, 2016.

[23] T. Petsatodis, A. Pnevmatikakis, C. Boukis, Voice activity detection using audio-visual information, in: International Conference on Digital Signal Processing (ICDSP 2009), Santorini, Greece, 1–5, doi:\bibinfo{doi}{10.1109/ICDSP.2009.5201171}, 2009.

[24] G. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Computation 18 (7) (2006) 1527–1554, doi:\bibinfo{doi}{10.1162/neco.2006.18.7.1527}.

[25] Y. Bengio, Learning deep architectures for AI, Foundations and trends® in Machine Learning 2 (1) (2009) 1–127, doi:\bibinfo{doi}{10.1561/2200000006}.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, Multimodal deep learning, in: International conference on machine learning (ICML2011), Bellevue, WA, USA, 689–696, 2011.

[27] F. Tao, C. Busso, Aligning Audiovisual Features for Audiovisual Speech Recognition, in: IEEE International Conference on Multimedia and Expo (ICME 2018), San Diego, CA, USA, 1–6, doi:\bibinfo{doi}{10.1109/ICME.2018.8486455}, 2018.

[28] S. Petridis, M. Pantic, Deep complementary bottleneck features for visual speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, 2304–2308, doi:\bibinfo{doi}{10.1109/ICASSP.2016.7472088}, 2016.

[29] S. Petridis, Z. Li, M. Pantic, End-to-end visual speech recognition with LSTMs, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA, 2592–2596, doi:\bibinfo{doi}{10.1109/ICASSP.2017.7952625}, 2017.

[30] J. Chung, A. Senior, O. Vinyals, A. Zisserman, Lip reading sentences in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 3444–3453, doi:\bibinfo{doi}{10.1109/CVPR.2017.367}, 2017.

[31] I. Ariav, D. Dov, I. Cohen, A deep architecture for audio-visual voice activity detection in the presence of transients, Signal Processing 142 (2018) 69–74, doi:\bibinfo{doi}{10.1016/j.sigpro.2017.07.006}.

[32] F. Tao, C. Busso, Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection, in: Interspeech 2017, Stockholm, Sweden, 1938–1942, doi:\bibinfo{doi}{10.21437/Interspeech.2017-1573}, 2017.

[33] F. Tao, C. Busso, Audiovisual Speech Activity Detection with Advanced Long Short-Term Memory, in: Interspeech 2018, Hyderabad, India, 1244–1248, doi:\bibinfo{doi}{10.21437/Interspeech.2018-2490}, 2018.

[34] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, T. Ogata, Lipreading using convolutional neural network, in: Interspeech 2014, Singapore, 1149–1153, 2014.

[35] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: International Conference on Machine Learning (ICML 2006), Pittsburgh,PA, USA, 369–376, doi:\bibinfo{doi}{10.1145/1143844.1143891}, 2006.

[36] Y. Zhang, W. Chan, N. Jaitly, Very deep convolutional networks for end-to-end speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA, 4845–4849, doi:\bibinfo{doi}{10.1109/ICASSP.2017.7953077}, 2017.

[37] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, J. Engel, W. Fang, L. Fan, C. Fougner, L. Gao, C. Gong, A. Hannun, T. Han, L. Johannes, B. Jiang, C. Ju, B. Jun, P. LeGresley, L. Lin, J. Liu, Y. Liu, W. Li, X. Li, D. Ma, S. Narang, A. Ng, S. Ozair, Y. Peng, R. Prenger, S. Qian, Z. Quan, J. Raiman, V. Rao, S. Satheesh, D. See-

tapun, S. Sengupta, K. Srinet, A. Sriram, H. Tang, L. Tang, C. Wang, J. Wang, K. Wang, Y. Wang, Z. Wang, Z. Wang, S. Wu, L. Wei, B. Xiao, W. Xie, Y. Xie, D. Yogatama, B. Yuan, J. Zhan, Z. Zhu, Deep speech 2: End-to-end speech recognition in english and mandarin, in: International Conference on Machine Learning (ICML 2016), New York, NY, USA, 173–182, 2016.

[38] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, A. Ng, Deep Speech: Scaling up end-to-end speech recognition, CoRR abs/1412.5567.

[39] B. Rivet, L. Girin, C. Jutten, Visual voice activity detection as a help for speech source separation from convolutive mixtures, Speech Communication 49 (7) (2007) 667–677, doi:\bibinfo{doi}{10.1016/j.specom.2007.04.008}.

[40] P. De Cuetos, C. Neti, A. Senior, Audio-visual intent-to-speak detection for human-computer interaction, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), vol. 4, Istanbul, Turkey, 2373–2376, doi:\bibinfo{doi}{10.1109/ICASSP.2000.859318}, 2000.

[41] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, P. Lucey, Visual Voice Activity Detection Using Frontal versus Profile Views, in: International Conference on Digital Image Computing Techniques and Applications (DICTA 2011), Noosa, Queensland, Australia, 134–139, doi:\bibinfo{doi}{10.1109/DICTA.2011.29}, 2011.

[42] B. Joosten, E. Postma, E. Krahmer, Visual Voice Activity Detection at different Speeds, in: International Conference on Auditory-Visual Speech Processing (AVSP 2013), Annecy, France, 187–190, 2013.

[43] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, C. Jutten, An Analysis of Visual Speech Information Applied to Voice Activity Detection, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), vol. 1, Toulouse, France, 601–604, doi:\bibinfo{doi}{10.1109/ICASSP.2006.1660092}, 2006.

[44] A. Aubrey, Y. Hicks, J. Chambers, Visual voice activity detection with optical flow, IET Image Processing 4 (6) (2009) 463–472, doi:\bibinfo{doi}{10.1049/iet-ipr.2009.0042}.

[45] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, C. Jutten, Two novel visual voice activity detectors based on appearance models and retinal filltering, in: European Signal Processing Conference (EUSIPCO 2007), Poznań, Poland, 2409–2413, 2007.

[46] R. Ahmad, S. Raza, H. Malik, Unsupervised multimodal VAD using sequential hierarchy, in: IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2013), Singapore, 174–177, doi:\bibinfo{doi}{10.1109/CIDM.2013.6597233}, 2013.

[47] P. Liu, 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, 609–612, doi:\bibinfo{doi}{10.1109/ICASSP.2004.1326059}, 2004.

[48] N. Ryant, M. Liberman, J. Yuan, Speech activity detection on Youtube using deep neural networks, in: Interspeech 2013, Lyon, France, 728–731, 2013.

[49] S. Sadjadi, J. H. L. Hansen, Unsupervised speech activity detection using voicing measures and perceptual spectral flux, IEEE Signal Processing Letters 20 (3) (2013) 197–200, doi:\bibinfo{doi}{10.1109/LSP.2013.2237903}.

[50] F. Tao, C. Busso, Lipreading Approach for Isolated Digits Recognition under Whisper and Neutral Speech, in: Interspeech 2014, Singapore, 1154–1158, 2014.

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

[52] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: M. Arbib (Ed.), The handbook of brain theory and neural networks, The MIT Press, ISBN 978-0262511025, 255–258, 1998.

[53] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems (NIPS 2012), vol. 25, Lake Tahoe, CA, USA, 1097–1105, 2012.

[54] T. Saitoh, Z. Zhou, G. Zhao, M. Pietikäinen, Concatenated frame image based CNN for visual speech recognition, in: C. Chen, J. Lu, K. Ma (Eds.), Asian Conference on Computer Vision (ACCV) - Workshops, vol. 10117 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Taipei, Taiwan, ISBN 978-3-319-54426-7, 277–289, doi:\bibinfo{doi}{10.1007/978-3-319-54427-4_21}, 2016.

[55] T. Sercu, C. Puhrsch, B. Kingsbury, Y. LeCun, Very deep multilingual convolutional neural networks for LVCSR, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Shanghai, China, 4955–4959, doi:\bibinfo{doi}{10.1109/ICASSP.2016.7472620}, 2016.

[56] H. Soltau, G. Saon, T. Sainath, Joint training of convolutional and non-convolutional neural networks, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), Florence, Italy, 5572–5576, doi:\bibinfo{doi}{10.1109/ICASSP.2014.6854669}, 2014.

[57] R. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, Neural computation 1 (2) (1989) 270–280, doi:\bibinfo{doi}{10.1162/neco.1989.1.2.270}.

[58] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, S. Khudanpur, Recurrent neural network based language model, in: Interspeech 2010, Makuhari, Japan, 1045–1048, 2010.

[59] A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), Vancouver, BC, Canada, 6645–6649, doi:\bibinfo{doi}{10.1109/ICASSP.2013.6638947}, 2013.

[60] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, 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, 4945–4949, doi:\bibinfo{doi}{10.1109/ICASSP.2016.7472618}, 2016.

[61] A. Katsamanis, M. P. Black, P. Georgiou, L. Goldstein, S. Narayanan, SailAlign: Robust long speech-text alignment, in: Workshop on New Tools and Methods for Very-Large Scale Phonetics Research, Philadelphia, PA, USA, 1–4, 2011.

[62] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems (NIPS 2014), vol. 27, Montreal, Canada, 2672–2680, 2014.

[63] V. Nair, G. Hinton, Rectified linear units improve restricted Boltzmann machines, in: International Conference on Machine Learning (ICML 2010), Haifa, Israel, 807–814, 2010.

[64] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), Portland, OR, USA, 532–539, doi:\bibinfo{doi}{10.1109/CVPR.2013.75}, 2013.

[65] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: International Conference on Learning Representations, San Diego, CA, USA, 1–13, 2015.

[66] V. M. Stanford, NIST Speech SNR Tool, https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements, 2005.