Signal Processing 142 (2018) 69–74

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Short communication

A deep architecture for audio-visual voice activity detection in the presence of transients☆

Ido Ariav*, David Dov, Israel Cohen

The authors are with the Andrew and Erna Viterbi Faculty of Electrical Engineering, The Technion-Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 9 April 2017
Revised 5 July 2017
Accepted 11 July 2017
Available online 12 July 2017

Keywords:
Audio-visual speech processing
Voice activity detection
Auto-encoder
Recurrent neural networks

ABSTRACT

We address the problem of voice activity detection in difficult acoustic environments including high levels of noise and transients, which are common in real life scenarios. We consider a multimodal setting, in which the speech signal is captured by a microphone, and a video camera is pointed at the face of the desired speaker. Accordingly, speech detection translates to the question of how to properly fuse the audio and video signals, which we address within the framework of deep learning. Specifically, we present a neural network architecture based on a variant of auto-encoders, which combines the two modalities, and provides a new representation of the signal, in which the effect of interferences is reduced. To further encode differences between the dynamics of speech and interfering transients, the signal, in this new representation, is fed into a recurrent neural network, which is trained in a supervised manner for speech detection. Experimental results demonstrate improved performance of the proposed deep architecture compared to competing multimodal detectors.

1. Introduction

Voice activity detection is a segmentation problem of a given speech signal into sections that contain speech and sections that contain only noise and interferences. It constitutes an essential part in many modern speech-based systems such as those for speech and speaker recognition, speech enhancement, emotion recognition and dominant speaker identification. We consider a multimodal setting, in which speech is captured by a microphone, and a video camera is pointed at the face of the desired speaker. The multimodal setting is especially useful in difficult acoustic environments, where the audio signal is measured in the presence of high levels of acoustic noise and transient interferences, such as keyboard tapping and hammering [1,2]. The video signal is completely invariant to the acoustic environment, and nowadays, it is widely available in devices such as smart-phones and laptops. Therefore, proper incorporation of the video signal significantly improves voice detection, as we show in this paper.

In silent acoustic environments, speech segments in a given signal are successfully distinguished from the silence segments using methods based on simple acoustic features such as zero-crossing rate and energy values in short time intervals [3–5]. However, the performances of these methods significantly deteriorate in the presence of noise even with moderate levels of signal-to-noise ratios (SNR). Another group of methods assumes statistical models for the noisy signal, focusing on estimation of the model parameters. For example, the variances of speech and noise can be estimated by tracking the variations of the noisy signal over time [6–9]. The main drawback of such methods is that they cannot properly model highly non-stationary noise and transient interferences, which are in the main scope of this study. The spectrum of transients often rapidly varies over time, as does the spectrum of speech, and as a result, they are not properly distinguished [2].

More recent studies address the problem of voice activity detection from a machine learning point of view, in which the goal is to classify segments of the noisy signal into speech and non-speech classes [10,11]. Learning-based methods learn implicit models from training data instead of assuming explicit distributions for the noisy signal. A particular school of models, relevant to this paper, is deep neural networks, which have gained popularity in recent years in a variety of machine learning tasks. These models utilize multiple hidden layers for useful signal representations, and their potential for voice activity detection has been partially exploited in recent studies. Zhang and Wu [12] proposed using a deep-belief network to learn an underlying representation of a speech signal from predefined acoustic features. The new representation is then fed into a linear classifier for speech detection. Mendelev et al. [13] introduced a multi-layer perceptron network for speech detection, and proposed to improve its robustness to

☆ This research was supported by the Israel Science Foundation (grant no. 576/16).
* Corresponding author.
E-mail addresses: idoariav@tx.technion.ac.il, idoariav@gmail.com (I. Ariav), davidd@tx.technion.ac.il (D. Dov), icohen@ee.technion.ac.il (I. Cohen).

http://dx.doi.org/10.1016/j.sigpro.2017.07.006
0165-1684/

noise using the "Dropout" technique [14]. Despite the improved performance, the network in [13] classifies each time frame independently, thus ignoring temporal relations between segments of the signal. The studies presented in [15–18] propose using a recurrent neural network (RNN) to naturally exploit temporal information by incorporating previous inputs for voice detection. These methods however still struggle in frames that contain both speech and transients. Since transients are characterized by fast variations in time and high energy values, they often appear more dominant than speech. Therefore, frames containing only transients appear similar to frames containing both transients and speech, so that they are wrongly detected as speech frames.

A different school of studies suggests improving the robustness of speech detection to noise and transients by incorporating a video signal, which is invariant to the acoustic environment. Often, the video captures the mouth region of the speakers, and it is represented by specifically designed features, which model the shape and movement of the mouth in each frame. Examples of such features are the height and the width of the mouth [19,20], key-points and intensity levels extracted from the region of the mouth [21–24], and motion vectors [25,26].

Two common approaches exist in the literature concerning the fusion of audio and video signals, termed early and late fusion [27,28]. In early fusion, video and audio features are concatenated into a single feature vector and processed as single-modal data [29]. In late fusion, measures of speech presence and absence are constructed separately from each modality, and then combined using statistical models [30,31]. Dov et al. [32,33], for example, proposed to obtain separate low dimensional representations of the audio and video signals using diffusion maps. The two modalities are then fused by a combination of speech presence measures, which are based on spatial and temporal relations between samples of the signal in the low dimensional domain.

In this paper, we propose a deep neural network architecture for audio-visual voice activity detection. The architecture is based on specifically designed auto-encoders providing an underlying representation of the signal, in which simultaneous data from audio and video modalities are fused in order to reduce the effect of transients. The new representation is incorporated into an RNN, which, in turn, is trained for speech presence/absence classification by incorporating temporal relations between samples of the signal in the new representation. The classification is performed in a frame-by-frame manner without temporal delay, which makes the proposed deep architecture suitable for online applications.

The proposed deep architecture is evaluated in the presence of highly non-stationary noises and transient interferences. Experimental results show improved performance of the proposed architecture compared to single-modal approaches that exploit only the audio or video signals, thus demonstrating the advantage of audio-video data fusion. In addition, we show that the proposed architecture outperforms competing multimodal detectors.

The remainder of the paper is organized as follows. In Section 2, we formulate the problem. In Section 3, we introduce the proposed architecture. In Section 4, we demonstrate the performance of the proposed deep architecture for voice activity detection. Finally, in Section 5, we draw conclusions and offer some directions for future research.

2. Problem formulation

We consider a speech signal simultaneously recorded via a single microphone and a video camera pointed at a front-facing speaker. The video signal comprises the mouth region of the speaker. It is aligned to the audio signal by a proper selection of the frame length and the overlap of the audio signal as described in Section 4. Let $\mathbf{a}_n \in \mathbb{R}^A$ and $\mathbf{v}_n \in \mathbb{R}^V$ be feature representations of the nth frame of the *clean* audio and video signals, respectively, where A and V are the number of features. Similarly to $\mathbf{a}_n$, let $\tilde{\mathbf{a}}_n \in \mathbb{R}^A$ be a feature representation of the audio signal contaminated by background noises and transient interferences. The audio and the video features are based on the Mel Frequency Cepstral Coefficients (MFCC) and motion vectors, respectively, and their construction is described in Section 4.

We consider a dataset of N consecutive triplets of frames $(\mathbf{a}_1, \tilde{\mathbf{a}}_1, \mathbf{v}_1), (\mathbf{a}_2, \tilde{\mathbf{a}}_2, \mathbf{v}_2), \ldots, (\mathbf{a}_N, \tilde{\mathbf{a}}_N, \mathbf{v}_N)$ containing both speech and non-speech time intervals. We use the clean signal $\{\mathbf{a}_n\}_1^N$ to label each time frame n according to the presence or absence of speech. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two hypotheses denoting speech absence and presence, respectively, and let $\mathbb{I}(n)$ be a speech indicator of frame n, given by:

$$\mathbb{I}(n) = \begin{cases} 1, & n \in \mathcal{H}_1 \\ 0, & n \in \mathcal{H}_0 \end{cases}. \tag{1}$$

The goal in this study is to estimate $\mathbb{I}(n)$, i.e., to classify each frame n as a speech or non-speech frame.

Voice activity detection is especially challenging in the presence of transients, which are typically more dominant than speech due to their short duration, high amplitudes and fast variations of the spectrum [2]. Specifically, frames that contain both speech and transients, for which $\mathcal{H}_1$ holds, are often similar in the feature space to non-speech frames that contain only transients, so that they are often wrongly classified as non-speech frames. To address this challenge, we introduce a deep neural network architecture, which is designed to reduce the effect of transients by exploiting both the clean and the noisy audio signals, $\mathbf{a}_n$ and $\tilde{\mathbf{a}}_n$, respectively, and the video signal $\mathbf{v}_n$.

3. Deep architecture for audio-visual voice activity detection

3.1. Review of autoencoders

The proposed deep architecture is based on obtaining a transient reducing representation of the signal via the use of auto-encoders, which are shortly reviewed in this subsection for the sake of completeness [34]. An auto-encoder is a feed-forward neural network with an input and output layers of the same size, which we denote by $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^D$, respectively. They are connected by one hidden layer $\mathbf{h} \in \mathbb{R}^M$, such that the input layer $\mathbf{x}$ is mapped into the hidden layer $\mathbf{h}$ through an affine mapping:

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{2}$$

where $\mathbf{W}$ is a $D \times M$ weight matrix, $\mathbf{b}$ is a bias vector and σ is an element-wise activation function. Then, $\mathbf{h}$ is mapped into the output layer $\mathbf{y}$:

$$\mathbf{y} = \tilde{\sigma}\left(\tilde{\mathbf{W}}\mathbf{h} + \tilde{\mathbf{b}}\right), \tag{3}$$

where $\tilde{\mathbf{W}}, \tilde{\mathbf{b}}, \tilde{\sigma}$ are defined similarly to $\mathbf{W}$, $\mathbf{b}$ and σ.

Optimal parameters (weights) $\tilde{\mathbf{W}}, \mathbf{W}, \tilde{\mathbf{b}}, \mathbf{b}$ are those that allow reconstructing the signal $\mathbf{x}$ at the output $\mathbf{y}$ of the auto-encoder, and they are obtained via a training procedure, by optimizing a certain loss function $L(\mathbf{x}, \mathbf{y})$, e.g., a square error, which we use here. It has been shown [35,36] that minimization of the auto-encoder's loss function $L(\mathbf{x}, \mathbf{y})$ is equivalent to maximization of a lower bound on the retained information between the input and output of the auto-encoder. Thus, the hidden layer $\mathbf{h}$, obtained by (2) with optimized parameters $\mathbf{W}$ and $\mathbf{b}$, has the maximal mutual information with the input signal $\mathbf{x}$. The activation functions $\sigma, \tilde{\sigma}$ are usually chosen to be non-linear functions; here, we use a sigmoid function $\sigma(z) = \frac{1}{1+\exp -z}$, so that the hidden layer $\mathbf{h}$ incorporates non-linear relations between different parts of the input signal [34,37]. In addition, the dimension M of $\mathbf{h}$ is typically set smaller than that of

the input signal D. Therefore, the hidden layer $\mathbf{h}$ is often considered as a non-linear low dimensional representation of the input signal.

A deep architecture of auto-encoders is constructed by stacking L auto-encoders such that the lth hidden layer, denoted by $\mathbf{h}^l$, is used as an input for the $(l+1)$th layer. The training is performed one layer at a time in a bottom-up fashion. The first layer of the deep architecture is trained with $\mathbf{x}$ as input, and once trained, $\mathbf{h}^1$ is calculated by (2) using the optimized parameters $\mathbf{W}^1$ and $\mathbf{b}^1$, where $\mathbf{W}^l$ and $\mathbf{b}^l$ denote the parameters of the lth layer. Then, we fix parameters $\mathbf{W}^1$ and $\mathbf{b}^1$ and use the obtained $\mathbf{h}^1$ as input for the training procedure of the second layer and similarly for all layers up to L.

3.2. Transient-reducing audio-visual autoencoder

We adopt ideas from [38,39] of using autoencoders to fuse multimodal signals. We propose a specifically designed deep architecture, based on feeding the auto-encoder with an audio-visual signal contaminated by acoustic noises and transients, while reconstructing the clean signal. Specifically, let $\mathbf{z}_n \in \mathbb{R}^{A+V}$ and $\tilde{\mathbf{z}}_n \in \mathbb{R}^{A+V}$ be feature vectors of frame n, obtained by concatenating the video features $\mathbf{v}_n$ along with the audio features $\mathbf{a}_n$ and $\tilde{\mathbf{a}}_n$, respectively, such that $\mathbf{z}_n = [\mathbf{a}_n^T, \mathbf{v}_n^T]^T$ and $\tilde{\mathbf{z}}_n = [\tilde{\mathbf{a}}_n^T, \mathbf{v}_n^T]^T$. The auto-encoder is fed by the noisy audio-visual feature vector $\tilde{\mathbf{z}}_n$, and is trained to reconstruct the clean signal $\mathbf{z}_n$, i.e., to minimize $L(\hat{\mathbf{z}}_n, \mathbf{z}_n)$ where $\hat{\mathbf{z}}_n \in \mathbb{R}^{A+V}$ is the output of the auto-encoder.

This approach simultaneously serves two purposes; it both allows fusing of the audio and the video modalities, and reduces the effect of transients. According to (2), the hidden layer $\mathbf{h}$ is obtained by a non-linear fusion between the entries of $\tilde{\mathbf{z}}$, and specifically, by the fusion of the audio and the video modalities. In addition, the effect of transients is reduced in the hidden layer $\mathbf{h}$ since the training process is designed to reconstruct the clean signal at the output. As a result, the hidden layer only captures factors that are related to the clean signal, as we demonstrate in Section 4.

We stack L such auto-encoders to form a deep neural network as described in Section 3.1. For layers $l>1$ we can no longer use the clean and the noisy speech signals; instead, we follow the principle of a de-noising auto-encoder [35], i.e., corrupt each input $\mathbf{h}_n^l$ with random noise, and train the auto-encoder to reconstruct the uncorrupted input. Vincent et al. [35] have shown that stacking several auto-encoders yields an improved representation for the input data over an ordinary one layer auto-encoder, since the added layers allow the auto-encoder to learn more complex higher-order relations across the modalities. Assuming an architecture of L such auto-encoder layers, we consider the last layer of the network, denoted by $\mathbf{p}_n \triangleq \mathbf{h}_n^L$, as the new underlying representation of the audio-visual signal.

It is worth noting that the proposed representation significantly differs from the common early and late fusion approaches [27,28] since it is obtained via the exploration of complex relations between the audio and video signals.

3.3. Recurrent neural network for voice activity detection

Speech is an inherently dynamic process comprising rapidly alternating speech and non-speech segments, i.e., a speech segment followed by a non-speech segment (pause) and vice versa. Indeed, temporal information is widely used for improving voice activity detection by incorporating several consecutive frames in the decision process [8,9]. However, the number of previous frames that should be considered and their weight on the decision process is not straightforward, and can change over time. For example, a common assumption is that speech is present with a higher probability if it was present in previous frames rather than after a non-speech (silent) frame. Thus, predetermining the amount of past information considered in the classification process for all frames can result in suboptimal results. We address this issue by incorporating an RNN for the classification of each frame $\tilde{\mathbf{z}}_n$.

An RNN is a feed-forward multi-layered neural network in which loop connections, which are added to the hidden layers, allow to incorporate temporal information in the decision process.

Given the auto-encoder's output at time frame n, $\mathbf{p}_n$, an RNN with one hidden layer $\bar{\mathbf{h}}_n$ computes the output layer $\bar{\mathbf{y}}_n$ using a hidden layer at time frame $n-1$, according to:

$$\bar{\mathbf{h}}_n = \hat{\sigma}\left(\hat{\mathbf{W}}\mathbf{p}_n + \grave{\mathbf{W}}\bar{\mathbf{h}}_{n-1} + \hat{\mathbf{b}}_n\right) \tag{4}$$

$$\bar{\mathbf{y}}_n = \bar{\sigma}\left(\bar{\mathbf{W}}\bar{\mathbf{h}}_n + \bar{\mathbf{b}}_n\right) \tag{5}$$

where $\hat{\mathbf{W}}, \grave{\mathbf{W}}$ and $\bar{\mathbf{W}}$ are weight matrices, $\hat{\mathbf{b}}$ and $\bar{\mathbf{b}}$ are the bias parameters, and $\hat{\sigma}, \bar{\sigma}$ are the corresponding activation functions. The case of an RNN with one hidden layer is extended to the case of an RNN with $\bar{L}>1$ layers by iteratively calculating the hidden layers for $l=1$ to $\bar{L}$:

$$\bar{\mathbf{h}}_n^l = \hat{\sigma}\left(\hat{\mathbf{W}}^l\bar{\mathbf{h}}_n^{l-1} + \grave{\mathbf{W}}^l\bar{\mathbf{h}}_{n-1}^l + \hat{\mathbf{b}}_n^l\right) \tag{6}$$

where $\bar{\mathbf{h}}_n^l$ is the lth hidden layer at time n, and $\hat{\mathbf{W}}, \grave{\mathbf{W}}$ and $\hat{\mathbf{b}}$ are defined as in (4). The first layer is the input layer, i.e., $\bar{\mathbf{h}}_n^0 \triangleq \bar{\mathbf{x}}_n$, and the output layer $\bar{\mathbf{y}}_n$ is calculated from (5) using the last hidden layer $\bar{\mathbf{h}}_n^{\bar{L}}$.

We incorporate the proposed transient-reducing representation $\{\mathbf{p}_n\}_1^N$ into the deep RNN in order to exploit the temporal information inherent in speech for voice activity detection. Specifically, for each frame n, we feed the new representation $\mathbf{p}_n$ to the RNN and iteratively compute the hidden layers $\bar{\mathbf{h}}_n^l$ according to (6). Then, we use the output layer $\bar{\mathbf{y}}_n$, and apply a sigmoid function to constrain its values to the range of $0-1$. Thus, we consider the output as a probability measure for the presence of speech in frame n, and propose to estimate the speech indicator $\mathbb{I}(n)$ in (1) by comparing the output to a threshold t:

$$\mathbb{I}(n) = \begin{cases} 1, & \bar{\mathbf{y}}_n \geq t \\ 0, & \bar{\mathbf{y}}_n < t \end{cases}. \tag{7}$$

The RNN has two beneficial properties for voice activity detection. First, the length of the temporal window used for speech detection is implicitly incorporated in the weights $\{\grave{\mathbf{W}}^l\}_1^{\bar{L}}$, and is automatically learned during the training process rather than being arbitrarily predefined. Second, the speech indicator in (7) is obtained via a supervised procedure, which exploits the true labels of the presence of speech, and allows for an accurate detection of speech as we show in Section 4.

4. Experimental results

4.1. Experimental setting

4.1.1. Dataset

We evaluate the proposed deep architecture for voice activity detection using the dataset presented in [32]. The dataset includes audio-visual sequences of 11 speakers reading aloud an article chosen from the web, while making natural pauses every few sentences. Thus, the intervals of speech and non-speech range from several hundred ms to several seconds in length. The video signal uses a bounding box around the mouth region of the speaker, cropped from the original recording, and it is of 90×110 pixels. The audio signal is recorded at 8 kHz with an estimated SNR of $\sim$25 dB. It is processed using short time frames of length 634 samples with 50% overlap such that it is aligned to the video frames which are processed at 25 frames/s. Each of the 11 sequences is

120 s long, and it is divided into two parts such that the first 60 s are used to train the algorithm and the rest of the sequence is used for evaluation.

The clean audio signal is contaminated with various background noises such as white Gaussian noise, musical instruments noise and babble noise, and with transients, such as a metronome, keyboard typing and hammering, taken from [40]. The training data extracted from each speaker contains all possible combinations of background noises and transients.

4.1.2. Feature selection

For the representation of the audio signal, we use MFCC[41], which represent the spectrum of speech in a compact form using the perceptually meaningful Mel-frequency scale. The MFCCs were found to perform well for voice activity detection under challenging conditions such as low SNR and non-stationary noise [32,42]. Each MFCC feature vector is composed of 12 cepstral coefficients, and their first and second derivatives, Δ and $\Delta\Delta$, respectively. Accordingly, the dimensions of the clean and contaminated audio feature vectors, $\mathbf{a}_n$ and $\tilde{\mathbf{a}}_n$, are $A = 36$. We note that by using Δ and $\Delta\Delta$ MFCCs, we incorporate temporal information into the process of learning the transient reducing representation. This allows for a better distinction between transients and speech, where the former typically vary faster over time. Even though the temporal information is also incorporated in the RNN, we found in our experiments that the use of Δ and $\Delta\Delta$ MFCCs further improves the detection results.

For the representation of the video signal, we use motion vectors, calculated using the Lucas–Kanade method [43,44]. Motion vectors are suitable for speech-related tasks since they capture both spatial and temporal information, i.e., the movement of the mouth, and they were previously exploited for voice activity detection in [25]. The feature representation, $\mathbf{v}_n$, is obtained by concatenating the absolute values of the velocities of each pixel from 3 consecutive frames $n-1, n, n+1$, so that its dimension is $V = 297$. We refer the reader to [32] for more details on the construction of the dataset and the audio-visual features.

4.1.3. Training process

The concatenated feature vector $\tilde{\mathbf{z}}_n$, of size $A+V = 333$, is fed as input to the transient-reducing audio-visual auto-encoder. The entries of $\tilde{\mathbf{z}}_n$ are normalized over the training set such that they have zero mean and unit variance in order to prevent saturation of the auto-encoder's neurons. We use an auto-encoder architecture with $L = 2$ hidden layers containing 200 neurons each, and with a logistic sigmoid activation function. During the training of the second hidden layer, in which we can no longer use the clean and contaminated signals for training, we contaminate the input for that layer with Gaussian noise with zero mean and variance 0.05 as described in Section 3.2.

The input layer of the RNN has 200 neurons, matching the output of the transient-reducing audio-visual auto-encoder, $\mathbf{p}_n$. The RNN comprises $\bar{L} = 3$ hidden layers with 50,50, and 30 neurons, activated with a logistic sigmoid function, so that the full system architecture is of the form 333($\tilde{\mathbf{z}}_n$)-200($\mathbf{h}_n^1$)-200($\mathbf{p}_n$)-50($\bar{\mathbf{h}}_n^1$)-50($\bar{\mathbf{h}}_n^2$)-30($\bar{\mathbf{h}}_n^3$)-1($\bar{\mathbf{y}}_n$).We used a sigmoid activation function in order to constrain the output of the entire network to be in the range [0, 1] so that it can be used as a probability measure for speech presence. For consistency, we also use the sigmoid for the activation of the hidden layers, and note that we found in our experiments that it performs similarly to the widely used ReLU [45]. We train the RNN layers in a supervised end-to-end manner using back propagation through time [36], and the whole system is optimized with gradient descent with a learning rate of 10^{-5} and momentum 0.9. All of the weights are initialized with values from a random normal distribution with zero mean and variance 0.01.

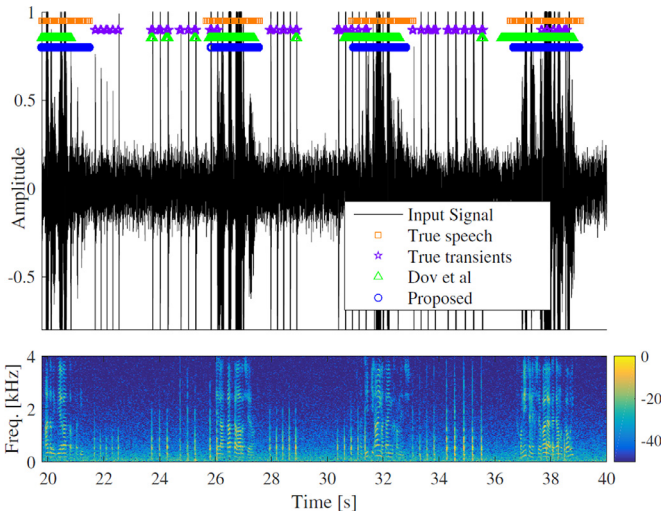

Fig. 1. Example of voice activity detection. Acoustic environment: colored Gaussian noise with 10 dB SNR and hammering transient interferences. (Top) Time domain, input signal – black solid line, true speech- orange squares, true transients – purple stars, competing method [32] with a threshold set for 90% correct detection rate – green triangles, proposed deep architecture with a threshold set for 90% correct detection rate – blue circles. (Bottom) Spectrogram of the input signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To prevent over-fitting, we use the early stopping procedure [46]; specifically, we use 30% of the training data as a validation set, on which we evaluate the network once every 5 epochs. The training procedure is stopped when the loss function of the network stops improving, and specifically, when 5 consecutive increases in validation error are obtained. To further increase robustness against convergence into suboptimal local minima, we train three realizations of the same network with different random weight initializations and average the predictions of the network over all realizations. The training time of all three realizations of the network took about 8 hours on an ordinary desktop computer.

4.2. Evaluation

To evaluate the performance of the proposed deep architecture, we compare it to the audio-visual voice activity detectors presented in [32] and [28], which are denoted in the plots by "Dov AV" and "Tamura", respectively. In Fig. 1 we present an example of speech detection in the presence of hammering transient. The performance of the proposed deep architecture is compared to the algorithm presented in [32] by setting the threshold value t in (7) to provide 90% correct detection rate, and comparing their false alarm rates. Fig. 1 shows that the proposed architecture yields significantly fewer false alarms compared to the competing detector, where the latter wrongly detects transients as speech, e.g., in seconds 33 – 36.

In Figs. 2–4 we compare the different algorithms in the form of receiver operating characteristic (ROC) curves, which present the probability of detection versus the probability of false alarm. The ROC curves are generated by spanning the threshold in (7) over all values between zero and one. Moreover, the maximal performance of each method for different acoustic environments is presented in Table 1. They are obtained using a threshold value that provides the best results in terms of true positive (TP) rate plus true negative (TN) rate.

In order to further demonstrate the benefit of the fusion of the audio and video signals for voice activity detection, we evaluated single modal versions of the proposed architecture based only on the audio or video modalities. The single modal versions are de-

Table 1
Comparison of different VADs in terms of TP + TN. The best result in each column is highlighted in bold fonts.

	Babble 10 dB SNR Keyboard	Musical 10 dB SNR Hammering	Colored 5 dB SNR Hammering	Musical 0 dB SNR Keyboard	Babble 15 dB SNR Scissors
Tamura	73.6	83.8	83.9	73.8	81.2
Dov - Audio	87.7	89.9	87.8	86.5	90.2
Dov - Video	89.6	89.6	89.6	89.6	89.6
Dov - AV	92.9	94.5	92.8	92.9	94.6
Proposed - AV	**95.8**	**95.4**	**95.9**	**95.1**	**97.2**

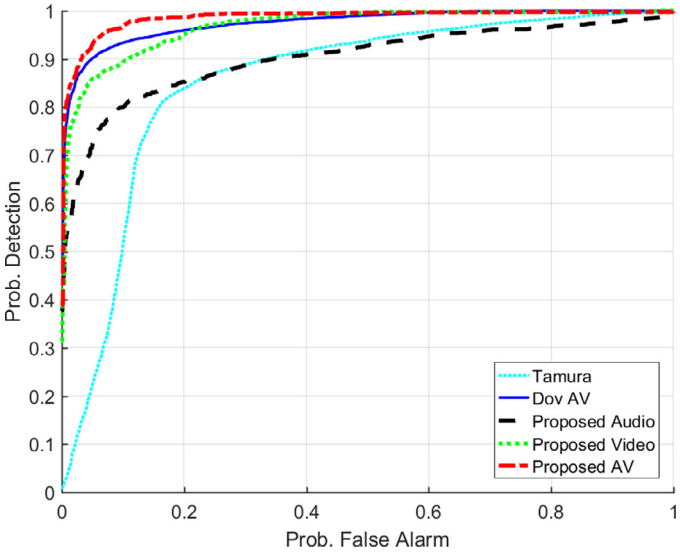

Fig. 2. Probability of detection versus probability of false alarm. Acoustic environment: Musical noise with 10 dB SNR and hammering transient interferences (best viewed in color).

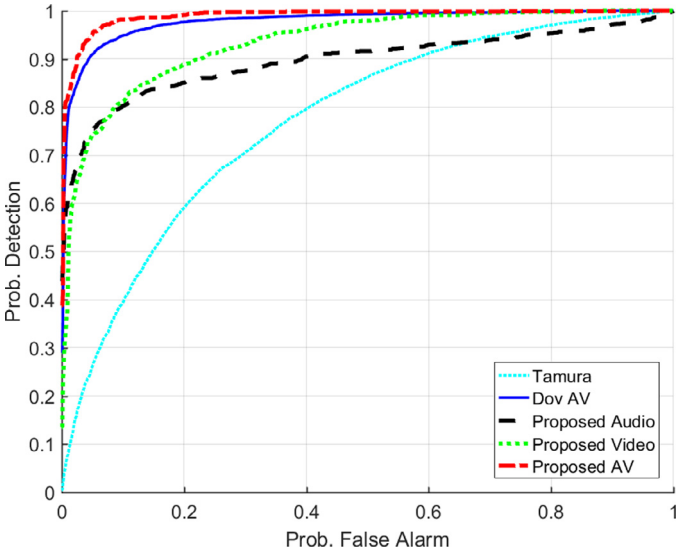

Fig. 3. Probability of detection versus probability of false alarm. Acoustic environment: Babble noise with 10 dB SNR and keyboard transient interferences (best viewed in color).

noted in the plots by "Proposed Audio" and "Proposed Video", respectively. When tested in a single modal version, the proposed deep architecture is fed only with features from one modality, after making proper changes to the input layer size. Then, the entire network is trained as described in Section 3. The benefit of fusing the audio and the video modalities is clearly shown in Figs. 2–4, where the proposed audio-visual architecture significantly outperforms the single modal versions. Also, the proposed deep architecture outperforms the audio-visual methods presented in [28] and [32].

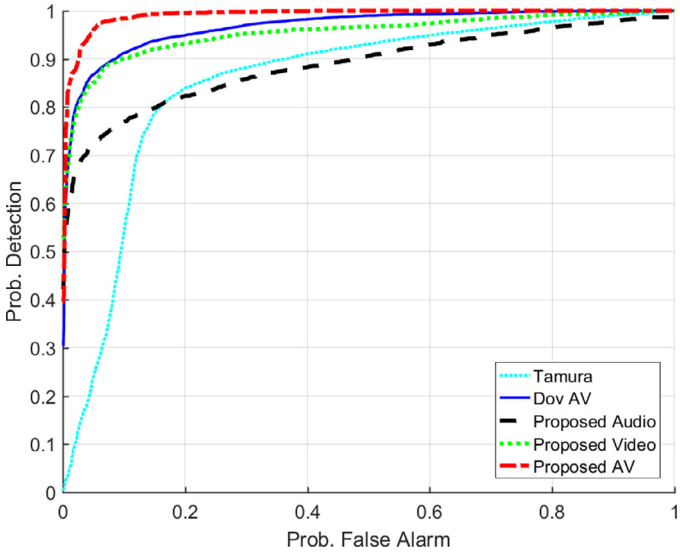

Fig. 4. Probability of detection versus probability of false alarm. Acoustic environment: Colored noise with 5 dB SNR and hammering transient interferences (best viewed in color).

In contrast to [32], where the modalities are merged only at the decision level, the proposed architecture exploits complex relations between the modalities learned by the transient-reducing auto-encoder. Moreover, in [32] the temporal context is only considered by concatenating features from a predefined number of consecutive frames, while in the proposed architecture the weights associated with previous frames are automatically learned by the supervised training process of the RNN, allowing for varying durations of temporal context to be exploited for voice activity detection.

5. Conclusions and future work

We have proposed a deep architecture for speech detection, based on specifically designed auto-encoders providing a new representation of the audio-visual signal, in which the effect of transients is reduced. The new representation is fed into a deep RNN, trained in a supervised manner to generate voice activity detection while exploiting the differences in the dynamics between speech and the transients. Experimental results have demonstrated that the proposed architecture outperforms competing state-of-the-art detectors providing accurate detections even under low SNR conditions and in the presence of challenging types of transients.

Future research directions include considering more complex variations of recurrent neural networks for the classification process. For example, bidirectional RNNs may be used to exploit the temporal context from future frames, and long short-term memory (LSTM) networks may facilitate learning even longer-term de-

pendencies between the inputs. Another next step is to perform a fine-tuning of the entire network from end to end in a supervised manner, while simultaneously updating the weights of the auto-encoder and the RNN via back propagation.

References

[1] D. Dov, I. Cohen, Voice activity detection in presence of transients using the scattering transform, in: Proc. 28th Convention of the Electrical & Electronics Engineers in Israel (IEEEI), IEEE, 2014, pp. 1–5.

[2] D. Dov, R. Talmon, I. Cohen, Kernel method for voice activity detection in the presence of transients, IEEE/ACM Trans. Audio, Speech Lang. Process. 24 (12) (2016) 2313–2326.

[3] D.A. Krubsack, R.J. Niederjohn, An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech, IEEE Trans. Signal Process. 39 (2) (1991) 319–329.

[4] J.-C. Junqua, B. Mak, B. Reaves, A robust algorithm for word boundary detection in the presence of noise, IEEE Trans. Speech Audio Process. 2 (3) (1994) 406–412.

[5] S. Van Gerven, F. Xie, A comparative study of speech detection methods, in: Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH), 1997, pp. 1095–1098.

[6] N. Cho, E.-K. Kim, Enhanced voice activity detection using acoustic event detection and classification, IEEE Trans. Consumer Electron. 57 (1) (2011) 196–202.

[7] J.-H. Chang, N.S. Kim, S.K. Mitra, Voice activity detection based on multiple statistical models, IEEE Trans. Signal Process. 54 (6) (2006) 1965–1976.

[8] J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6 (1) (1999) 1–3.

[9] J. Ramírez, J.C. Segura, C. Benítez, A. De La Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information, Speech Commun. 42 (3) (2004) 271–287.

[10] J.W. Shin, J.-H. Chang, N.S. Kim, Voice activity detection based on statistical models and machine learning approaches, Comput. Speech Lang. 24 (3) (2010) 515–530.

[11] J. Wu, X.-L. Zhang, Maximum margin clustering based statistical VAD with multiple observation compound feature, IEEE Signal Process. Lett. 18 (5) (2011) 283–286.

[12] X.-L. Zhang, J. Wu, Deep belief networks based voice activity detection, IEEE Trans. Audio, Speech Lang. Process. 21 (4) (2013) 697–710.

[13] V.S. Mendelev, T.N. Prisyach, A.A. Prudnikov, Robust voice activity detection with deep maxout neural networks, Mod. Appl. Sci. 9 (8) (2015) 153.

[14] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[15] S. Leglaive, R. Hennequin, R. Badeau, Singing voice detection with deep recurrent neural networks, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 121–125.

[16] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.

[17] T. Hughes, K. Mierle, Recurrent neural networks for voice activity detection, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7378–7382.

[18] W.-T. Hong, C.-C. Lee, Voice activity detection based on noise-immunity recurrent neural networks, Int. J. Adv. Comput. Technol. (IJACT) 5 (5) (2013) 338–345.

[19] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, C. Jutten, An analysis of visual speech information applied to voice activity detection, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1, 2006, pp. 601–604.

[20] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, C. Jutten, A study of lip movements during spontaneous dialog and its application to voice activity detection, J. Acoustical Soc. Am. 125 (2) (2009) 1184–1196.

[21] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, C. Jutten, Two novel visual voice activity detectors based on appearance models and retinal filtering, in: Proc. 15th European Signal Processing Conference (EUSIPCO), 2007, pp. 2409–2413.

[22] E.-J. Ong, R. Bowden, Robust lip-tracking using rigid flocks of selected linear predictors, in: Proc. 8th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2008.

[23] Q. Liu, W. Wang, P. Jackson, A visual voice activity detection method with adaboosting, in: in Proc. Sensor Signal Processing for Defence (SSPD), 2011, pp. 1–5.

[24] S. Siatras, N. Nikolaidis, M. Krinidis, I. Pitas, Visual lip activity detection and speaker detection using mouth region intensities, IEEE Trans. Circuits Syst. Vid. Technol. 19 (1) (2009) 133–137.

[25] A. Aubrey, Y. Hicks, J. Chambers, Visual voice activity detection with optical flow, IET Image Process. 4 (6) (2010) 463–472.

[26] P. Tiawongsombat, M.-H. Jeong, J.-S. Yun, B.-J. You, S.-R. Oh, Robust visual speakingness detection using bi-level HMM, Pattern Recognit. 45 (2) (2012) 783–793.

[27] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Syst. 16 (6) (2010) 345–379.

[28] S. Tamura, M. Ishikawa, T. Hashiba, S. Takeuchi, S. Hayamizu, A robust audio-visual speech recognition using audio-visual voice activity detection, in: Proc. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 2694–2697.

[29] I. Almajai, B. Milner, Using audio-visual features for robust voice activity detection in clean and noisy speech, in: Proc. 16th European Signal Processing Conference (EUSIPCO), 2008.

[30] T. Yoshida, K. Nakadai, H.G. Okuno, An improvement in audio-visual voice activity detection for automatic speech recognition, in: Proc. 23rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2010, pp. 51–61.

[31] V.P. Minotto, C.B. Lopes, J. Scharcanski, C.R. Jung, B. Lee, Audiovisual voice activity detection based on microphone arrays and color information, IEEE J. Selected Topics Signal Process. 7 (1) (2013) 147–156.

[32] D. Dov, R. Talmon, I. Cohen, Audio-visual voice activity detection using diffusion maps, IEEE/ACM Trans. Audio, Speech Lang. Process. 23 (4) (2015) 732–745.

[33] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmonic Anal. 21 (1) (2006) 5–30.

[34] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.

[36] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back--propagating errors, Nature 323 (1986) 533–536.

[37] Y. Bengio, Y. LeCun, Scaling learning algorithms towards AI, Large-Scale Kernel Mach. 34 (5) (2007) 1–41.

[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proc. 28th International Conference on Machine Learning (ICML-11), 2011, pp. 689–696.

[39] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Advances in Neural Information Processing Systems, 2012, pp. 2222–2230.

[40] [Online]. Available: http://www.freesound.org.

[41] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoustics, Speech Signal Process. 28 (4) (1980) 357–366.

[42] S. Mousazadeh, I. Cohen, Voice activity detection in presence of transient noise using spectral clustering, IEEE Trans. Audio, Speech Lang. Process. 21 (6) (2013) 1261–1271.

[43] J.L. Barron, D.J. Fleet, S.S. Beauchemin, Performance of optical flow techniques, Int. J. Comput. Vis. 12 (1) (1994) 43–77.

[44] A. Bruhn, J. Weickert, C. Schnörr, Lucas/kanade meets horn/schunck: combining local and global optic flow methods, Int. J. Comput. Vis. 61 (3) (2005) 211–231.

[45] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proc. 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

[46] L. Prechelt, Early stopping-but when? in: G. Montavon, G. B. Orr and K.-R. Müller (Eds.), Neural Networks: Tricks of the Trade, Springer, 2012, pp. 53–67.