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