Energy and Computation Efficient Audio-visual Voice Activity Detection Driven by Event-cameras

Arman Savran*, Raffaele Tavarone†, Bertrand Higy*‡, Leonardo Badino† and Chiara Bartolozzi*
*iCub Facility, Istituto Italiano di Tecnologia, Italy
†CTNSC, Istituto Italiano di Tecnologia, Italy
‡Università di Genova, Italy

***Abstract*—We propose a novel method for computationally efficient audio-visual voice activity detection (VAD) where visual temporal information is provided by an energy efficient event-camera (EC). Unlike conventional cameras, ECs perform on-chip low-power pixel-level change detection, adapting the sampling frequency to the dynamics of the activity in the visual scene and removing redundancy, hence enabling energy and computational efficiency. In our VAD pipeline, first, lip activity is located and detected jointly by a probabilistic estimation after spatio-temporal filtering. Then, over the lips, a feather-weight speech-related lip motion detection is performed with minimum false negative rate to activate a highly accurate but expensive acoustic deep neural networks-based VAD. Our experiments show that ECs are accurate at detecting and locating lip activity; and EC-driven VAD can result in considerable savings in computations as well as can substantially reduce false positive rates in low acoustic signal-to-noise ratio conditions.**

I. INTRODUCTION

Voice activity detection (VAD) is the first step in many speech processing systems. However, in certain applications like mobile hands free scenarios, where VAD has to continuously run in background, it can be computationally expensive and drain the battery. Computation becomes even more demanding when robustness is increased by detecting the type of the noise [15]. As such, many consumer applications by default rely on the user activation of speech processing, like button press.

As for the majority of audio-visual applications, during the use of smartphones or tablets, with specific apps or video calling, the user typically faces the device and the facial movements gathered by the frontal camera represents an additional source of information about speech production. Prior work exploited this fact to obtain acoustic-noise-robust VAD using video cameras [3], [7], [11], [13], [17], [21]. Optical flow [7], [17], [21] and Discrete Cosine Transform (DCT) [3], [13] have commonly been employed for feature extraction. Authors have performed various statistical learning methods of significant complexity on the visual as well as the audio features. Recently deep auto-encoder and recurrent networks have been applied on audio-visual data as well [4]. Although there are no reports of the computational complexity of the proposed techniques, vision processing implies a considerable computation burden due to the dense image data and the fixed video-frame rate. Moreover, we see that complexity has significantly increased with more recent methods in order to improve the detection [4], [11], [17].

To achieve energy and computation efficiency for acoustic-noise-robust VAD, we propose a novel approach based on event-driven vision sensors (or event cameras, ECs). ECs implement pixel-level temporal change detection [19] and only respond to changes in their field of view, asynchronously generating pixel-events at a rate proportional to the change of contrast. Their temporal resolution can be as high as 1μs and the sampling rate adapts to the dynamics of the stimulus: for stationary inputs they do not produce a significant amount of events, while the sampling rate can be very high for fast changing stimuli. This corresponds to power consumption at 50mW for a static scene and at 175mW maximum for high activity scenes. In comparison, power consumption of conventional cameras mounted in mobile phones are around 1W and can go up to 1.5W depending on the model, and microphones consume slightly more above 100mW [9], [22]. Thus ECs clearly provide substantial energy efficiency incomparable to conventional vision cameras, and intrinsic data compression that in turn enables design of computationally efficient processing, where the cost of computation dynamically changes with the amount of activity in the visual scene.

The VAD method we propose is based on an extremely efficient EC-driven processing which activates a very robust but costly audio-only deep neural network (ADNN) detector only when it detects lip activity. This EC-based gating mechanism provides dramatical energy and computation savings, especially at idle time. The temporally and spatially sparse but highly informative nature of the EC signals has a great potential for computationally efficient VAD, which is impossible to achieve with conventional cameras where dense video-frame images have to be processed at a constant frame rate (independently of whether there is movement or not in front of the camera). Moreover, the high sampling rate

This work is supported by the European Union's Horizon2020 project ECOMODE (grant No 644096).

978-1-5386-2335-0/18/$31.00

of ECs prevents motion blur problems which have degrading effects on visual VAD [7]. Another crucial advantage is the extremely high dynamic range which make ECs work similarly well under varying and inhomogeneous lighting conditions where conventional vision methods fail or require even more complex processing to combat these difficulties. Prior work with EC-based detection involves [5], [16]. To our knowledge, the present work is the first study on event-camera based speech activity detection. The characterization of this method shows that the visual gate runs very efficiently (Sec. IV-C), and overall VAD is also efficient by calling the complex ADNN only when there is potential lip-activity, yielding remarkable low error rates (Sec. IV-D).

II. Event-Camera Gated VAD

Fig. 1 shows a block diagram of the proposed EC-gated VAD. An event camera is used to monitor the visual scene, generating ON/OFF pixel-events for increasing/decreasing contrast. In the resulting event-stream there's no notion of the traditional "frame", as events are continuously and asynchronously produced. Therefore each event is represented as a tuple $e_i = (\mathbf{x}_i, t_i, p_i)$ where $\mathbf{x}_i$ is the 2D pixel position, t_i is the time stamp and p_i is the polarity (ON/OFF). An event at time t is generated as soon as a change in the log intensity, $\log I$, exceeds a threshold θ, $|\log I(x,t) - \log I(x,t^p)| \geq \theta$, where t^p is the timestamp of the previous event at the same pixel. However, as a single event is not informative enough, we use a spatio-temporal volume of fixed size and duration of events to perform computation. Given the current volume of events, we perform event-rate detection to distinguish activity elicited by global camera motion or a potential user talking in front of the sensor. This first thresholding activates an event-based filter with salient activity selection that enables lip-activity localization. Then a lip activity detection-based gating operation activates (or not) an ADNN-VAD. In the "Collect" stage, a decision probability is assigned to each acoustic frame, as either the ADNN posterior or as "0" if lips have not been detected from the visual stream and the ADNN-VAD is inactive. The final decision is made by uniform temporal averaging over a sliding window of 61 frames centered at the current acoustic frame (corresponding to a 600ms temporal window). This averaging performs smoothing as well as fusion since negative decisions of the visual detector are combined with ADNN-VAD posteriors.

A. Event-rate Detection

Given the characteristic activation of EC, the number of events in an output event-stream is highly variable, corresponding to different computational loads. Still scenes fire pixel-events very sparsely (only noise), while large camera motion or sudden and fast movements in front of the camera produce dense pixel-event clouds in space-time volume. A person talking in front of the camera, instead, elicits an intermediate activation level of event-pixels. To limit the computational demand, we simply avoid processing if the event-rate (total number of events in a time window) is high than expected. We determine the threshold as the maximum event-rate observed in the training set of talking face clips, sampling with the same rate as of the spatio-temporal filter.

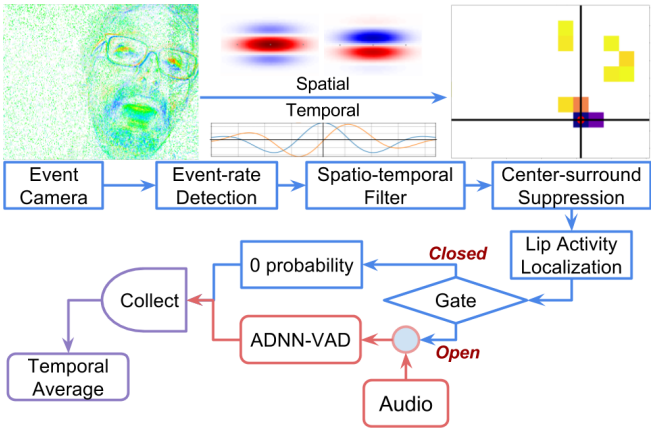

Figure 1. Block diagram of the EC-gated VAD. The acoustic signal is sent to ADNN-VAD via the gate only when lip activity is detected. At the top are events accumulated over 200ms (left), spatio-temporal components of the 3D Gabor filter (middle) and the sparse activation map from the center-surround suppression (right).

B. Spatio-temporal Filtering

Small head movements in front of EC elicit events corresponding to face, eyes, nose and lip contours. Speech related lip movements appear as horizontally elongated objects, with dominant motion component in the vertical axis, modulated in the 2-7 Hz frequency band [8]. Based on these characteristics, we apply a spatio-temporal Gabor filter, i.e., temporal extension of commonly used spatial Gabor filters in image analysis [18]. The filter (centered at (x_0, y_0, t_0)) is a 3D-Gaussian modulated by a complex sinusoid on the temporal axis (t) as well as on the vertical axis (y) to focus only on horizontally oriented lip shape:

$$g(x,y,t) = e^{-\left(\frac{(x-x_0)^2}{\sigma_x^2} + \frac{(y-y_0)^2}{\sigma_y^2} + \frac{(t-t_0)^2}{\sigma_t^2}\right)} \cdot e^{\{-2\pi j[f_y(y-y_0) + f_t(t-t_0)]\}}. \quad (1)$$

We set the spatial modulation as $f_y = 24$ pixels/cycle and temporal modulation as $f_t = 10$Hz. High temporal fidelity of ECs allow exact calculation of the temporal response. To determine the Gabor envelope, we use the half-response frequency bandwidth, using bandwidth of 1.0 octaves for the temporal component, and 2.7 octaves for the spatial domain, to better perform spatial localization. The spatial aspect ratio of the filter is 0.5, hence $\sigma_y = \sigma_x/2$. Real and imaginary profiles of the filter are shown in Fig. 1. The spatio-temporal filter size is set according to its standard deviation (43 pixels, 200ms). At each location we have separate filters for ON and OFF events, thus we obtain two separate output maps.

This filtering (convolution) stage is the computationally dominant part of the whole visual processing. For efficient computation, we create a 3D look-up table of the spatio-temporal volume. Depending on the space-time overlap of the filter kernels, more than one kernel can occupy the same index in the look-up table. Therefore, for each table index, the list of occupying kernels are stored together with the corresponding kernel weight (complex scalar). Thus, unlike a standard convolution implementation which strides the kernel in the space-time domain requiring a loop over the 3D look-up table (multiple-passes over an event), we do event-driven convolution which requires only one-pass over an event without looping over the space-time table, and only performing floating point additions per event to accumulate the Gabor filter responses. The complexity of this technique is $O(N_E/r^3)$, where N_E is the number of events and r is the filters' overlap ratio; hence it does not depend on the kernel size unlike in image convolution. We set $r = 0.5$, resulting in 21 pixels and 100 ms step sizes and complexity of $O(8 \times N_E)$. Outputs are calculated at each time step, by evaluating the Gabor filter magnitude responses.

C. Center Surround Suppression

Center-Surround Suppression (CSS) – applied to the output of Gabor filters – selects local maxima, considerably reducing redundant information, improving detection. As in [18], we implement CSS as difference of Gaussians and half-wave rectification: $\mathbf{D} = |\mathbf{G}_2 - \mathbf{G}_1|^+$, where the 2D filter kernels are $\mathbf{D}, \mathbf{G}_2, \mathbf{G}_1 \in R^{5x5}$. $|.|+$ is the half-wave rectifier applied at each kernel element by setting the negative values to 0. $\mathbf{G}_1$ is the matching envelope of the Gabor filter and $\mathbf{G}_2$ is the scaled Gaussian with $\sigma_2 = k_2.\sigma_1$, where k_2 corresponds to the surround which inhibits the center. For a given inhibition strength factor α, CSS activations are calculated by a convolution and subtraction $\mathbf{A} = \mathbf{M} - \alpha\mathbf{M} * \mathbf{D}/||\mathbf{D}||_1$, where $\mathbf{M}$ is the 2D input magnitude response map. We can simply perform the same operation with a convolution using the suppression filter kernel $\mathbf{CSS} = \mathbf{I} - \alpha.\mathbf{D}/||\mathbf{D}||_1$, where $||.||_1$ is the L_1 norm and $\mathbf{I}$ is the identity matrix. For $\alpha = 2$ and $k_2 = 4$ (see [18]), CSS provides high sparsity, typically activating 5% to 10% of cells in our system, suppressing noise and reducing false positives.

Fig. 2 shows some example inputs (the first column) and outputs (the second column) of CSS, however, by combining the maximum of ON and OFF cells at each cell location as a single map for visualizing compactly. The third column shows the input event-data with accumulated pixel-events where the filter locations are depicted by the overlaid 9×12 cell grid. We see in the first row that in the absence of activity there are only weak activations due to noise and slight movements. In the second row talking lips yield strong activations. The face has a moderate movement in the third row, however, the spatio-temporal filter does not respond to the moving vertical edges. Although there are still other strong responses, the filter-matched CSS disregards them. In the fourth row we see moderate activations due to the eyebrow movements, which are not as strong as the lip activity. However, there can also be cases like in the last row, albeit rare, where there is no lip activity but other strong activations like eye blinks. If those activations are detected as lip activity, the audio detection is activated unnecessarily. Nevertheless, our method can also handle those cases by means of an estimation with a location prior as explained in Section II-D.

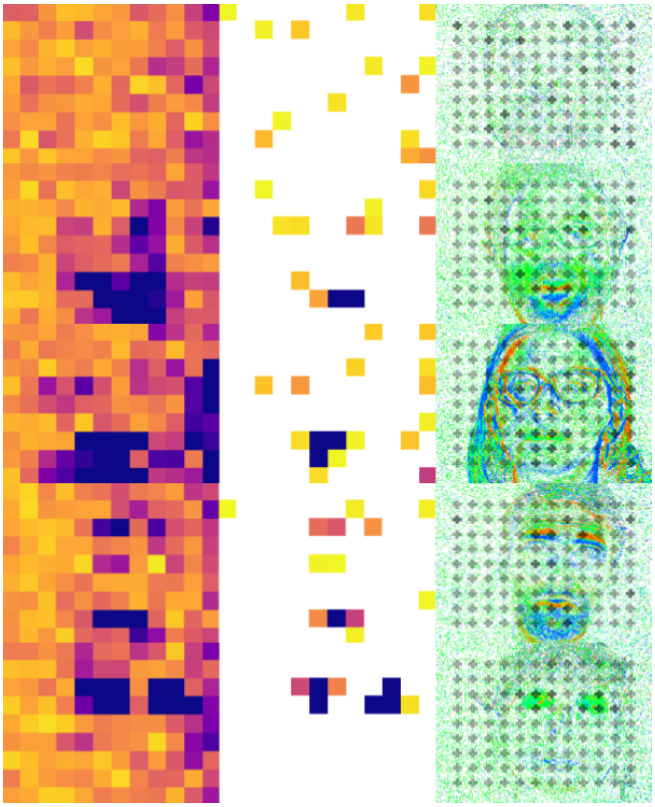

Figure 2. Gabor magnitude (the first column) and the center surround suppression activation ($\mathbf{A}$, in the second column) maps for several moments in different speech clips, by selecting the maximum values from the filtered ON and OFF polarity maps. Event-data (in the last column) is displayed by accumulating pixel-events over 200 ms window (blue: ON events, red: OFF events, green: combination), where the activation grids are overlaid.

D. Detection and Localization of the Lip Activity

With event cameras, lips are typically observed only if a subject is moving or talking in front of the camera. In this scenario, we can formulate lip activity detection and localization problems jointly. Given a sparse activation map $\mathbf{A}$, detection and location probability distributions can be evaluated by marginalizing out from the joint distribution, $p(y, \mathbf{x}|\mathbf{A})$, where the binary random variable $y \in \{0, 1\}$ denotes the presence of the activity when $y = 1$, and $\mathbf{x}$ is the 2D coordinate of the activity center. The location distribution

is approximated with a probability mass function over the activation grid cells. Thus the joint distribution is evaluated via the Bayes rule

$$p(y, \mathbf{x}_{ij}|\mathbf{A}) = \frac{p(\mathbf{A}|y, \mathbf{x}_{ij})p(y, \mathbf{x}_{ij})}{p(\mathbf{A})} \quad (2)$$

where $\mathbf{x}_{ij}$ is the location vector at the grid cell (i, j). Since the location and activity of the lips are independent, $p(y, \mathbf{x}_{ij}) = p(y)p(\mathbf{x}_{ij})$. We use equal prior probabilities for the values of y, and a 2D Gaussian density is employed as the location prior

$$p(\mathbf{x}_{ij}) \propto N(\mathbf{x}_{ij}; \mu_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}}). \quad (3)$$

The likelihood function of $\mathbf{A}$ given there is a lip activity like activation $(y = 1)$ is modeled by linear logistic regression. However, since the activation values are exponentially distributed, we first do a logarithmic transformation. Thus,

$$p(\mathbf{A}|y = 1, \mathbf{x}_{ij}) = \frac{1}{1 + exp(-w. \log A_{ij} + b)} \quad (4)$$

where A_{ij} is the activation value at the grid cell (i, j), w is the gain and b is the bias. On the other hand, the likelihood given $y = 0$ has a uniform shape since when there is no activity lips are not seen and can be anywhere in the scene. In that case the posterior (2) equals to the prior $p(\mathbf{x}_{ij})$ due to the Bayes rule.

After calculating the posterior we obtain the location distribution by marginalizing out y,

$$p(\mathbf{x}_{ij}|\mathbf{A}) = \sum_{y} p(y, \mathbf{x}_{ij}|\mathbf{A}). \quad (5)$$

Then the detection location is found on the activation grid by Maximum a-posterior (MAP) estimate (exhaustive search)

$$i^*, j^* = \arg\max_{i,j} p(\mathbf{x}_{ij}|\mathbf{A}). \quad (6)$$

Thus the estimated location is $\mathbf{x}_{i^*j^*}$. However, we only trust and use the location if the detection probability is high

$$p(y|\mathbf{x}_{i^*j^*}, \mathbf{A}) = \frac{p(y, \mathbf{x}_{i^*j^*}|\mathbf{A})}{p(\mathbf{x}_{i^*j^*}, \mathbf{A})}. \quad (7)$$

In the localization experiments we compare use of two weak priors: i) talking face scene prior, and ii) face center prior. While the former models the distribution of the mouth center in the scene by a Normal density, the latter models relative to the face bounding box center so that the detector can benefit from a face detector if available (in that case, $p(\mathbf{x}_{ij}) \propto N(\mathbf{x}_{ij}; \mu_{\mathbf{x}} + \mathbf{x}_{\mathbf{face}}, \mathbf{\Sigma}_{\mathbf{x}})$).

E. Visual Gating by Lip Activity Detection

After the localization, we perform local analysis over the lips at every 100 ms for voice activity detection. We observed that activation value of the cell which is nearest to the estimated location (within 2 std. of spatial Gabor kernel) provides accurate voice activity detection (when there is visible lip activity). Knowing the location, we apply a low threshold on the nearest activation cell. The threshold is determined to attain low false negative rate (see Sec. IV-C). When visual activity is detected at the lips, audio processing is enabled, however, for more than 100 ms. This visual gate takes into account possible delay between visible articulation and voice production [8], and the likely longer duration of speech segments in the absence of visible articulation. The minimum duration of the gate-open state is set to 200 ms, and extended up to 500 ms as discussed and assessed in Sec. IV-C. The visual detector runs continuously – also during the activation of the ADNN-VAD – and the gate is kept open for as long as lip activity is detected. Although this visual gate can possibly detect short non-verbal activity, in practice ADNN-VAD easily handles them with negligible cost, since they are rare events of short duration.

F. Acoustic Deep Neural Network based VAD

Recent audio-VAD studies have shown that DNN-based algorithms achieve state-of-the-art [12], [15]. Therefore we use a DNN-based audio VAD, as a feed-forward network with 4 hidden layers, 2000 rectified linear units per layer, and a 2-node softmax output layer. The acoustic signal is converted into 40 log mel-filtered spectral coefficients plus deltas and delta-deltas, computed every 10 ms over a 25 ms window. Then 11 contiguous frames are concatenated to be used as ADNN input. Thus computation-complexity is roughly $O(frames \times layers \times nodes^2)$, i.e., 1600 MFlops.

III. DATASET

We collected an audio-visual speech dataset using the "ATIS" (304×240 pixels resolution) Event Camera [19] for the visual signal, and a high quality directional microphone at 44.1 kHz for the auditory signal. The dataset was collected in a lab environment with standard room illumination and using 8 mm lens to video-record subjects standing 70 cm away from microphone and EC. The dataset consists of 18 subjects (9 males and 9 females) with different variations among eye-glasses, beard, mustaches and head poses (e.g. with the head slightly rotated to one-side). Subjects were free to slightly move their head while speaking. A unique set of 20 utterances were selected from the TIMIT [23] speech corpus for each subject. There are 5 sentences in common between only one pair of subjects. In total there are 360 audio-visual clips for a total duration of 28 minutes. We also collected non-speech facial actions from all the subjects, including free head rotations, lip-biting, lip-snap, breathing with open mouth, smiling and occlusions of face and lips due to hand gestures. Finally, we collected event streams corresponding to arbitrary motions of the EC – without subjects – to evaluate visual gating false positives in presence of typical hand-held devices handling.

For accurate synchronization of visual and auditory signals, we manually triggered an audio-visual signal – a LED

light and a buzzer – for each session, and manually marked the audio-visual signals to re-align possible time-shifts. For the ground-truth end points, speech segments were manually marked on the audio waveforms. There are silence sections at the beginning and ending parts of each clip. The overall voice percentage on the speech corpus is about 64%.

In order to test performance in different background noise conditions, the audio clips were mixed with recordings of subway, cafeteria, square backgrounds [1], at seven levels of SNR ratio from 15dB to -15dB with 5dB decrements. For each clean audio clip, we thus obtained 21 noisy clips.

IV. Experimental Results

In the experiments we first evaluate the lip activity localization performance (Section IV-A). We explain our DNN training procedure to achieve a robust acoustic VAD in Section IV-B. Then we evaluate our lip activity detection and the visual gate for VAD (Section IV-C). Finally, audio-visual EC-Gated VAD evaluation is given in Section IV-D. In all the experiments, the first 6 speakers are used for training and the remaining 12 subjects are used for testing.

A. Lip Activity Localization

To evaluate the lip activity localization performance and to perform training, we annotated mouth and face bounding boxes on the visual event-data, by creating frame videos with 40 ms window and step sizes, as shown in Fig. 3. However, since on the event-data face and lips appear only when there is motion, we annotated the bounding boxes by manually finding the key motion frames, sometimes separately for face and lips depending on the motion. If the face was moving, key frames were selected at the onset and offset of the movement as well as at several in-between frames. Using linear key-frame interpolation, bounding boxes are re-sampled at desired time points. We annotated for all the 18 subjects in the dataset, for over 4 speech clips per subject.

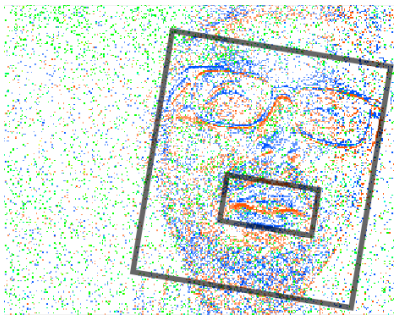

Figure 3. Annotated lip region and face bounding boxes. Event-pixels are accumulated over 40 ms window.

The Gaussian priors that we evaluate (see Sec. II-D), i.e., talking face scene and face center priors, are estimated over the re-sampled bounding box center coordinates (at 40 ms frame step). For the face center prior, estimation is done over the difference vectors $(\mathbf{x_{lips}} - \mathbf{x_{face}})$.

To train the linear logistic regression (for the activation probability), we need positive and negative activation values of the lip activity. In order to select the apex activation values, we apply a threshold. This threshold is determined for each clip independently by taking the average activation value over a space-time region where the boundaries are the spatial bounding box of the lip activity and the voiced segment. Positive activation values are the ones that are above the threshold in that region. On the other hand, values outside this region and below the threshold are assigned to the negative class. This sampling from the activation maps is performed at 100 ms steps, and results in very unbalanced sample sizes with much more negatives than the positives. Therefore we use class weighting in training (also applying L_2 regularization with the hyperparameter value set to 1).

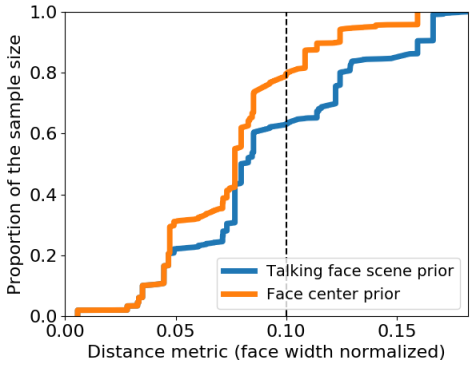

Figure 4. Localization performance over 1520 voiced samples across 12 unseen test subjects (talking face scene prior vs. face center prior).

Fig. 4 shows the lip localization performance curves for detection, comparing the talking face scene prior and the face center prior. The curves are cumulative distribution functions which show the proportion of samples that achieve less (or equal) than a given error metric. The error metric is

$$e = \|\mathbf{x} - \mathbf{x}_{gth}\|/w_{face} \tag{8}$$

where $\mathbf{x}$ is the estimated lips center coordinate vector, $\mathbf{x}_{gth}$ denotes the ground-truth location, and w_{face} is the ground-truth face width. These results are obtained over 1520 voiced samples across 12 unseen test subjects. We see that both of the detectors result in accurate localization since in all the test samples lips are located, by at most within 16% of the face width using the face center prior and with 18% of the face width using the talking face scene prior. This level of precision is sufficient since the localization precision is not as important as the accuracy for the VAD task. We also see that use of a face center prior clearly increases the precision, as 80% of the samples are already localized within 10% of the face width, while with the talking face scene prior

64% of the samples are localized with the same precision. Recall that we employ only a coarse spatial-sampling for the activation maps since the final goal is detection in time to activate audio processing. If higher precision is required for a local spatial analysis task, for instance, for visual speech recognition, a fine spatial sampling may be beneficial.

B. Training for Robust Audio VAD

We trained the ADNN-VAD on a balanced subset of the QUT-NOISE-TIMIT corpus [10], which retains all the noise types and the noise levels used for training. When tested on a different subset of the same dataset, our ADNN-VAD significantly outperforms other well known algorithms, e.g., it produces a Half Total Error Rate (HTER, average of false negative and false positives) of $\approx 1\%$ at low-level noise, vs. the 6% HTER of a Gaussian Mixture Model (Fig. 2 in [10]). If directly applied to our dataset, it performs poorly as the training dataset; because the two dataset are significantly different, both in terms of noise types and recording conditions. Thus, we performed domain adaptation [6] by fine tuning the DNN on both clean and corrupted acoustic data of the 6 speakers in the training set. These adaptation data were previously down-sampled to 16 kHz in order to match the sampling rate of QUT-NOISE-TIMIT.

C. Evaluation of the EC-Gate

We first evaluate the EC visual gate regarding its computation efficiency as well as its detection performance by varying gate threshold and open-duration, while comparing against a well known computationally efficient acoustic gating proposed by Sohn [20]. For comparison, we applied default parameters for the auditory gating, as implemented in the voicebox toolkit [2]. Speech-likelihoods are computed every 10ms with 20ms, and a Markov-model based hangover is applied for the continuity (smoothing) of the voice activity. As the proposed visual gate runs with 100 ms shift size and 200ms time-window, we down-sampled the reference acoustic method at 10 Hz by averaging over 200ms window.

We characterized the complexity of our method in terms of computational load by estimating the peak of the EC visual processing, which occurs only during the localization of the lip region since all the filter cells have to be evaluated. After localization EC processing is negligible, as computation is then performed only at the lip region that is cheaply tracked over time. Although in visual VAD studies, e.g., [7], [11], [13], the localization aspect is not evaluated (being out-of-scope), the assessment of its complexity is important as localization is the dominant part of the visual processing.

While it is not possible to make direct comparison of our method to visual non-EC based gating on our dataset – since we don't have simultaneous acquisition of traditional frame-based videos – we can make a rough comparison of the acquisition and processing costs. Mobile-device cameras consume about one order of magnitude more power (see Section I). Video-frame processing is computationally very demanding, since processing is performed on every pixel for each frame. On the contrary, event-driven processing is proportional to the dynamics of the scene content and is performed only on active pixels. After lip-region localization, while standard methods perform methods like DCT [13], optical flow [7] and diffusion mapping [11], our method performs floating point additions proportional to the number of events (with negligible fixed-rate calculations).

	Static Scene	Global Activity	Lip & Face	Visual Speech
EC-Gate (max.) *(Event-rate [Meps])*	**0.24** *(0.02)*	0.33 *(2.68)*	9.1 *(1.18)*	**10.45** *(1.02)*
Sohn-Gate	7.7 (*)			
ADNN-VAD	$>$ 1000 (*)			

Table I
COMPUTATIONAL LOAD ESTIMATES [MFLOPS]: EC-GATE (FOR DIFFERENT SCENARIOS) ARE ESTIMATED AS MFLOPS ADDITIONS AT THE PEAK OF THE VISUAL PROCESSING TO LOCATE LIPS (EC-GATE MFLOPS < 0.1, IF LOCALIZED), WHILE SOHN-GATE [20] AND ADNN-VAD ARE ESTIMATED AS MFLOPS MULTIPLICATIONS (*).

Table I shows MFlops (mega-floating-point-operations-per-second) estimates of lip search and EC-gate for VAD over different types of visual activity, including (i) static scene where only noise event-pixels exist, (ii) maximum activation when the device is moved (Global Activity), and (iv) visual speech and (iii) other lip and facial motion described in Sec. III. EC event-rate shows high variability, that depends on the amount of activity in the scene and that changes the computational load of the EC lip-search, however, our method sets an upper bound to the number of events that can be processed. The number of operations peaks if the lip search is required in the whole scene. Once the lips are localized, complexity becomes negligible due to local computations, yielding less than 0.1 MFlops additions. These peak EC MFlops estimates are much less than the computations required by ADNN, which are more than 1000 MFlops (Sec.II-F). More computationally efficient ADNN could be achieved through compression techniques, at the cost of some performance decreases (see, e.g., [14]). The complexity estimate of the default voicebox implementation of Sohn [20] was 7.7 MFlops. Sohn's complexity was roughly estimated based on the complexity of the used FFT ($O(nlog(n))$ with n = number of samples), and the ratio between FFT execution time and the execution time of all the remaining operations within Sohn's VAD.

Fig. 5 shows the receiver operating characteristic (ROC) curves for the EC and Sohn's methods at 100ms shift size with a 200ms time-window, combining clean and noisy audio examples. With the EC method there is a limit on the maximum true positive rate (TPR), which is about 80% (or 20% false negative rate) as seen in Fig. 5. This is due to the fact that visible articulation only corresponds to a

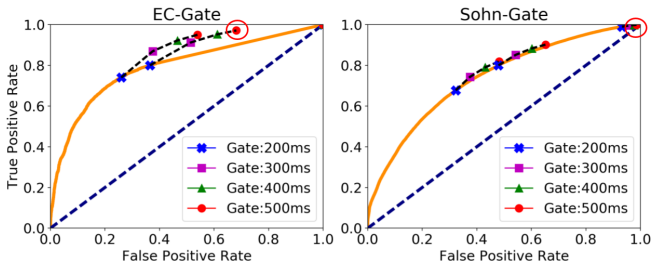

Figure 5. Receiver operating characteristic curves of the gates together with different operation points depending on the gate durations (each of 4 durations are shown for 2 thresholds over 2 dashed lines). Chosen operation points are marked with circles. All the clean and noisy audio samples are combined.

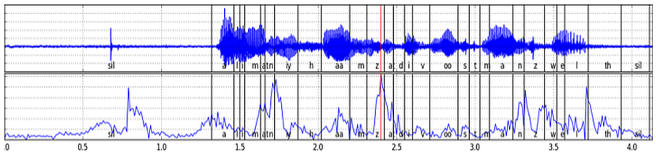

Figure 6. Audio waveform and lip-region even-rate profile of a four seconds long utterance "Alimony harms a divorced man's wealth." are shown with phoneme boundaries.

subset of phonemes, as shown in Fig. 6. Fig. 5 shows that extending the duration for EC-Gate counteracts this effect and produces a significant improvement: e.g., starting from two different thresholds on the ROC curve, we can boost the TPR much faster than the ROC curve. Thus we select the threshold value at 80% TPR and the gate duration 500ms to achieve 97% TPR, which is way higher than the natural limit of the EC. On the other hand, for Sohn-Gate, there is no significant advantage from extending the duration as the corresponding operation points are always almost on the ROC curve. This is as expected because the speech acoustic signal is informative for VAD during the voiced periods as opposed to visual VAD, and hence changing the threshold and extending the duration have similar effect. In order to perform the comparisons under equivalent conditions, for the Sohn-Gate, we select the same 500 ms gate duration, and adjust the threshold value for the same TPR which nevertheless causes very high FPR.

D. Evaluation of the EC-Gated VAD

As the complexity of the gating is much lower than the ADNN-VAD (see Sec. IV-C), the overall computational and energy efficiency of the gated-VAD is mainly related to how often the ADNN is activated (call rate). Table II compares call and FP rates of different VAD methods on speech clips, on unvoiced lip- and face-activity clips, and on unvoiced static scenes. The results are calculated over all acoustic noise types and levels, by setting a threshold on the outputs of each VAD to achieve 1% FN over the speech clips, including both clean and noisy audio. In addition to gated-VADs, we also considered the vision-only EC-VAD and the audio-only ADNN-VAD. For the EC method, we set the probability score to 1.0 if the gate is open. EC-VAD achieves 1% FN only after the temporal averaging of the outputs at the final stage of detection (Fig. 1).

Activity:	Visual Speech		Lip & Face		Static Scene	
	Voice: 64.5%		Voice: 0.0%		Voice: 0.0%	
	FN: 1%		FN: 0%		FN: 0%	
Method	**Call**	**FP**	**Call**	**FP**	**Call**	**FP**
EC-VAD	0.0	74	0.0	65	0.0	0
EC-ADNN	**86.7**	27	**58.6**	6	**0.0**	0
Sohn-ADNN	99.6	28	93.8	9	93.8	9
ADNN-VAD	100	26	100	10	100	10

Table II
ADNN CALL PERCENTAGES (CALL) AND FALSE POSITIVE PERCENTAGES (FP) OVER ALL NOISE TYPES AND LEVELS. OPERATION POINT IS SET SO THAT THE FALSE NEGATIVE PERCENTAGE (FN) IS 1% OVER THE SPEECH CLIPS COMBINING ALL NOISE CASES.

For the unvoiced static scenes, which can be assumed to be either rare or frequent, depending on the application, EC-driven VAD methods naturally lead to 0% call and FP rates. Due to the acoustic noise, Sohn-Gate yields 93.8% call rate and the very high FP rate of the gate is reduced to 9%, thanks to ADNN which alone has 10% FP rate. Negative cases, such as unvoiced lip- and facial-activity scenes are difficult to handle with vision-based methods, mostly because lip open-close actions open the gate, that correspond to 65% FP rate. In this case, EC-ADNN has 58.6% call rate, and the use of ADNN drops FP rate to 6%. Finally, in visual speech scenes (64.5% voice over non-voice activity), the call rate is 86.7% for EC-ADNN and 99.6% with Sohn-ADNN. Also in this case, the high FP rates of the gates drop by the help of ADNN, however, to rather moderate values (around 27%). Observing a slightly higher FP rate compared to ADNN (26%) looks contradictory, however, it is because at the 1% FN rate EC-ADNN attains a lower threshold value due to the temporal averaging with 0 probability values (Fig. 1).

Fig. 7 compares call, FP and FN rates of EC-ADNN and Sohn-ADNN under varying SNRs over all acoustic noise types, with the same operating points. Notice that since operating points for 1% FN is fixed by combining clips of all SNR levels including the clean audio, FN rates vary depending on the SNR (the third row in Fig. 7). Sohn-ADNN is inefficient as it calls ADNN most of the time. Efficiency is comparable or less than EC-ADNN only when the audio is clean and the scene has difficulties in handling lip- and face-activity, as 58.6% call rate with EC and 48.3% with Sohn's method. However, the FN plot shows that Sohn-ADNN achieves a poor (2.0%) FN rate with clean audio at this operation point, while EC-ADNN achieves 0.4%. This higher error rate is due to the fact that Sohn's method sometimes misses true positives, while the noise added to the voice always makes the Sohn's method open the gate which in turn activates the robust ADNN. In this context, acoustic gating proves to be inefficient. Additionally, EC-ADNN also

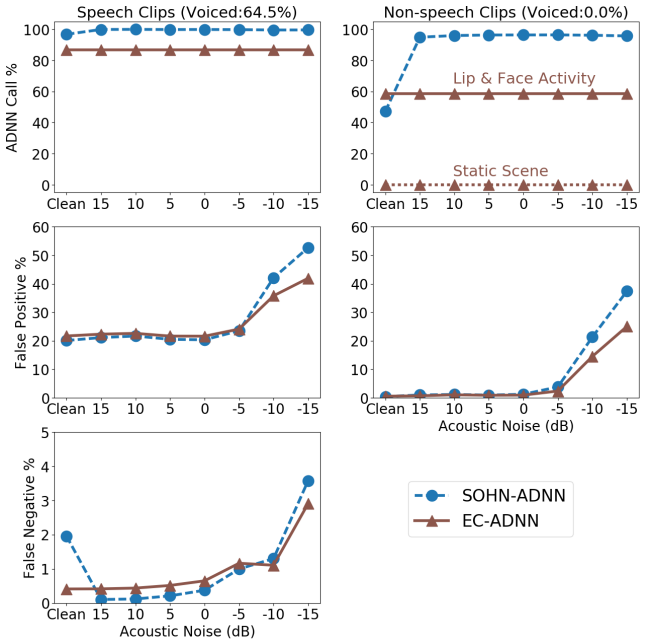

Figure 7. ADNN call, FP and FN rates varying with SNR over all noise types. Operation point is set so that the FN is 1% over the speech clips combining all noise cases.

considerably reduces the FP rate at higher acoustic noise levels, even in the presence of unvoiced lip- and face-activity, showing higher gating efficiency.

V. Conclusions

We have developed a novel audio-visual approach to VAD where efficient visual processing derived from the use of non-conventional low-power vision sensors enables the sparse activation of deep neural network based acoustic voice activity detection. The proposed event-driven vision processing tailors a spatio-temporal filter to the features of speech-related lip motion and coarsely removes activity that is not related to speech, hence provides an effective gating mechanism. We demonstrate that our method drastically reduces the overall computation and potentially the power consumption thanks to low-power compressed vision sensing, the very low computation-time complexity of the visual gate, and to a gating mechanisms that prevents complex processing when there is no potential lip activity. It also decreases the detection of false positives in high levels of acoustic noise, further reducing the activation of speech processing. The proposed method is hence extremely valuable, where VAD must continuously run in background while a user is facing towards the device, without occupying most of the resources and wasting battery power.

References

[1] Aurora. http://aurora.hsnr.de/.

[2] Voicebox. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html/.

[3] I. Almajai and B. P. Milner. Using audio-visual features for robust voice activity detection in clean and noisy speech. In *European Signal Processing Conference*, August 2008.

[4] I. Ariav, D. Dov, and I. Cohen. A deep architecture for audio-visual voice activity detection in the presence of transients. *Signal Processing*, 142(Supplement C):69 – 74, 2018.

[5] S. Barua, Y. Miyatani, and A. Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *IEEE Winter Conf. on App. of Computer Vision*, March 2016.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[7] M. Buchbinder, Y. Buchris, and I. Cohen. Adaptive weighting parameter in audio-visual voice activity detection. In *IEEE Int. Conf. on the Science of Electrical Eng.*, Nov 2016.

[8] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar. The natural statistics of audiovisual speech. *PLOS Computational Biology*, 5(7):1–18, 07 2009.

[9] X. Chen, Y. Chen, Z. Ma, and F. C. A. Fernandes. How is energy consumed in smartphone display applications? In *Workshop on Mobile Computing Sys. and App.*, 2013.

[10] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason. The qut-noise-timit corpus for the evaluation of voice activity detection algorithms. In *Interspeech*, Japan, September 2010.

[11] D. Dov, R. Talmon, and I. Cohen. Audio-Visual Voice Activity Detection Using Diffusion Maps. *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, 23(4):732–745, 2015.

[12] Y. Fujita and K.-i. Iso. Robust dnn-based vad augmented with phone entropy based rejection of background speech. In *Interspeech*, 2016.

[13] H. Ghaemmaghami, D. Dean, S. Kalantari, S. Sridharan, and C. Fookes. Complete-linkage clustering for voice activity detection in audio and visual speech. In *Interspeech*, Germany, September 2015.

[14] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[15] I. Hwang, H.-M. Park, and J.-H. Chang. Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection. *Comput. Speech Lang.*, 38:1–12, 2016.

[16] J. Li, F. Shi, W. Liu, D. Zou, QiangWang, W. Liu, D. Zou, Q. Wang, H. Lee, P.-K. Park, and H. E. Ryu. Adaptive Temporal Pooling for Object Detection using Dynamic Vision Sensor. In *British Machine Vision Conference*, 2017.

[17] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas. Visual voice activity detection in the wild. *IEEE Transactions on Multimedia*, 18(6):967–977, June 2016.

[18] N. Petkov and E. Subramanian. Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition. *Biol. Cybern.*, 97(5-6):423–439, 2007.

[19] C. Posch, D. Matolin, and R. Wohlgenannt. A QVGA 143 dB Dynamic Range Frame-Free PWM Image Sensor With Lossless Pixel-Level Video Compression and Time-Domain CDS. *IEEE J. Solid-State Circuits*, pages 1–16, 2011.

[20] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Proc. Letters*, 6(1):1–3, Jan. 1999.

[21] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu. Voice Activity Detection based on Fusion of Audio and Visual Information. In *Auditory-Visual Speech Proc. (AVSP)*, 2009.

[22] S. Tarkoma, M. Siekkinen, E. Lagerspetz, and Y. Xiao. *Smartphone energy consumption: modeling and optim.*. 2014.

[23] A. Wrench. Mocha-timit. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html, November 2006.