AUDIO-VISUAL TARGET SPEAKER EXTRACTION ON MULTI-TALKER ENVIRONMENT USING EVENT-DRIVEN CAMERAS

Ander Arriandiaga[1] *Giovanni Morrone*[2] *Luca Pasa*[3] *Leonardo Badino*[3] *Chiara Bartolozzi*[1]

[1] iCub Facility, Istituto Italiano di Tecnologia, Genova, Italy
[2] Department of Engineering Enzo Ferrari, University of Modena and Reggio Emilia, Modena, Italy
[3] CTNSC, Istituto Italiano di Tecnologia, Ferrara, Italy

ABSTRACT

In this work, we propose a new method to address audio-visual target speaker extraction in multi-talker environments using event-driven cameras. All audio-visual speech separation approaches use a frame-based video to extract visual features. However, these frame-based cameras usually work at 30 frames per second. This limitation makes it difficult to process an audio-visual signal with low latency. In order to overcome this limitation, we propose using event-driven cameras due to their high temporal resolution and low latency. Recent work showed that the use of landmark motion features is very important in order to get good results on audio-visual speech separation. Thus, we use event-driven vision sensors from which the extraction of motion is available at lower latency computational cost. A stacked Bidirectional LSTM is trained to predict an Ideal Amplitude Mask before post-processing to get a clean audio signal. The performance of our model is close to those yielded in frame-based fashion.

Index Terms— audio-visual target speaker extraction, event-driven camera, optical-flow, LSTM, deep learning

1. INTRODUCTION

The ability to disentangle and correctly recognise the speech of a single speaker among other speakers (the well known cocktail party effect [1]) is paramount for effective speech interaction in unconstrained environments. As such, it is an extremely useful feature for artificial agents, such as speech assistants and robots. Humans solve this problem using complementary and redundant strategies such as physical sound source separation (thanks to stereo sound acquisition [2]) but also using cues from observing the motion of lips [3].

In artificial systems, good results were achieved using Long-Short Memory Networks (LSTM) [4, 5, 6, 7] or dilated convolutional layers [8] to extract speaker-independent clean audio from multi-talker environment using single-channel audio signals. However, these approaches need to determine the correspondence between the target speaker and the output clean speech signal, and the number of speakers has to be known in advance. A solution is to give as input to the model some target speaker dependant features [9, 10], using an LSTM-based speaker encoder to produce speaker-discriminative embeddings. However, this solution needs a reference utterance of the speaker and an additional trainable Deep Neural Network (DNN), making the speech separation performance conditioned on the performance of the speaker encoder network.

Inspired on the findings that viewing the target speaker's face improves the listener ability to track the speech in a Cocktail Party setting [3], methods that combine visual cues and speech processing achieved remarkably good results. They were based on residual networks (ResNet [11]) pre-trained on a word-level lip-reading task [12] and [13], or on a pre-trained face recognition model, in combination with 15 dilated convolutional layers [14]. Such architectures require heterogeneous and large audio-visual datasets to train the models. A possible approach that allows to use smaller datasets (such as the GRID dataset [15]) is to rely on pre-trained models. Good, but speaker-dependent, results were obtained with the use of images and corresponding optical flow as inputs to a pre-trained dual tower ResNet extracting visual features [16]. Another alternative is to extract the video features directly from the image without using trainable methods. In such a case, the neural networks are smaller and can be trained with smaller datasets without overfitting. Following this idea, in [17] they used face landmark movements as input visual features to a bidirectional LSTM that achieved good speaker-independent results on the GRID dataset. In this work, the use of landmark motion features rather than positional features turned out to be a key factor. Inspired by this finding, we propose to substitute the visual pipeline implemented with traditional frame-based sensors, face tracking and extraction of motion landmarks, with an equivalent pipeline, based on the use of a novel type of vision sensors – the event-driven cameras (EDC) – from which the extraction of motion is available at lower computational cost and latency.

EDCs asynchronously measure the brightness change for

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

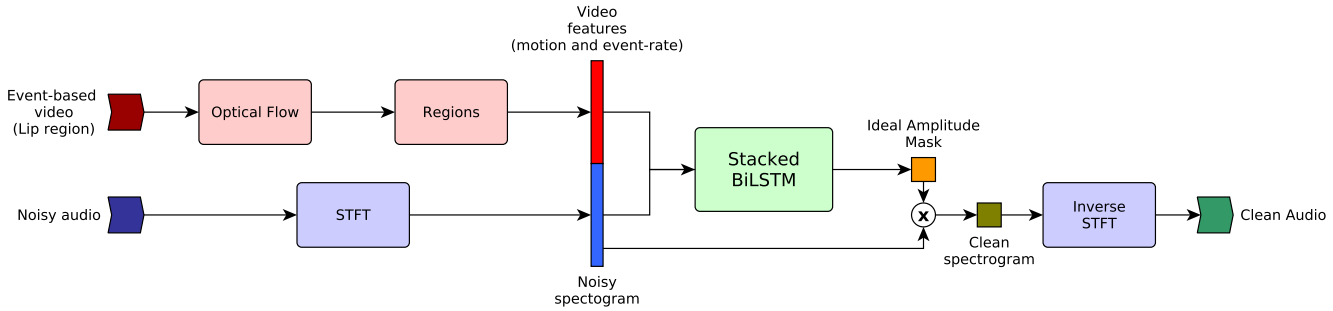

Fig. 1. Audio-Visual Speech Separation pipeline.

each pixel, featuring a temporal resolution as high as 1 μs, extremely low latency and data compression (as only active pixels communicate data). With such an input, the audio-visual system can use the same temporal discretization of the auditory pipeline (10 ms), rather than the one of the visual pipeline (30 ms is the standard frame-rate of traditional sensors).

Event-driven vision sensors have been widely used with good results for object tracking [18, 19], detection [20] and segmentation [21], and for gesture recognition [22]. Recently, they have been applied in the context of speech processing: vision-only speech recognition (i.e., lip-reading) on GRID exploited EDCs as input to a Deep Neural Network architecture [23]; lip movements detected by an EDC were used to detect speech activity and enable an auditory-based voice activity detection [24], for embedded applications that require low computational cost.

This work presents an audio-visual target speaker extraction system on multi-talker environment using event-driven vision sensors that compute motion at lower latency and computational cost. Following [17], we propose a non trainable method to extract visual features combined with deep learning techniques to extract the target talker speech in multi-talker environment. We use the GRID corpus in order to compare this approach with frame-based methods. To the best of our knowledge, this is the first work that presents an audio-visual target speaker extraction system that uses event-driven cameras.

2. METHODS

This work is based on [17]. The main differences are the way we capture the video, using event-driven camera, and how we extract video futures in a non-trainable fashion, using the optical flow. We get visual video from event-driven camera and then compute the optical flow from event-driven video to get visual features. Besides, we extract audio features and concatenate both, audio and video features, to train a Recurrent Neural Network (RNN) in order to get a time-frequency mask that we can multiply by noisy spectrogram to obtain the clean spectrogram (see Figure 1).

2.1. Event-driven camera

Event-driven cameras output asynchronous events whenever each pixel detects changes in log intensity larger than a threshold. These events have an associated timestamp, t, pixel position, $< x, y >$, and polarity (log intensity increase or decrease), p. Outputting these events asynchronously, event-driven cameras reduce latency and increase the dynamic range compared with frame-based cameras [25]. Likewise, the event-driven cameras only emit events for moving objects, they remove any data redundancy from not moving objects on the field of view. Thus, the camera outputs large amount of events when something is moving fast and it outputs few events when something is moving slowly in the scene. These properties make event-driven cameras suitable for extracting motion features like those happening when a subject is talking. Besides, event-driven cameras do not have any time restriction like happens with frame-based cameras that record at 30 fps, where there is only visual information available every 33 ms.

2.2. Optical-flow

Although event-driven cameras are suitable to extract motion features, this is no easy task because traditional motion estimation methods can not be used with events due to high temporal precision of the events. In this work, we use a method presented by Benosman *et al.* [26]. This method uses the time of the events from the camera to compute the direction and amplitude of the motion of each event. One interesting thing of this method is that it relies on the precise timing of each event and performs the optical flow of each event without using time intervals and thus, there is no need to generate frames to compute the optical flow. Besides, this method is computationally and temporally very efficient.

3. EXPERIMENTAL SETUP

The complete pipeline of the system we use in this work can be seen in Figure 1.

3.1. Dataset

The GRID dataset is used in this work [15]. The dataset consists of audio and videos of 3 seconds of 34 speakers speaking 1000 sentences in front of the camera. The dataset was generated using 200 sentences from 33 speakers (one was discarded because the videos from the speaker were not available). Then, with each sentence from each speaker other 3 different audio-mixed samples were generated using other speaker sentences. Finally, for each speaker 600 mixed-audio were available. From the total amount of samples, samples from 25 speakers were for training, from 4 speakers for validation and from the last 4 speakers for testing the model. The video was upscaled to 100 fps using video processing software to have more temporal information. The event-based video was recorded pointing ATIS event-driven camera [25] with 8mm lens to a high definition LED monitor. Due to the low quality of the original videos (360×288 pixels resolution) and in order to preserve the details in lip movements, the mouth area with 100×50 pixels resolution from each event-based video is extracted.

3.2. Model training

In order to train the model we need to pre-process the original audio waveform, pre-process the original video to extract the visual features and define the RNN architecture and configuration.

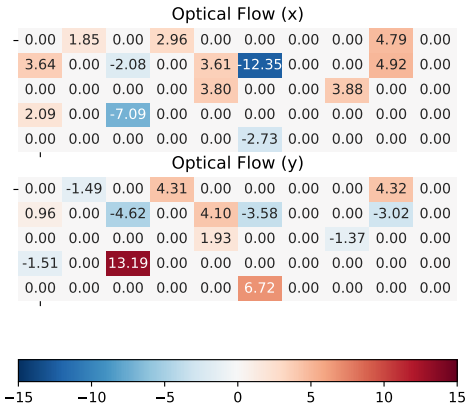

Fig. 2. Optical flow representation with regions of 10×10 pixels

3.2.1. *Audio pre- and post-processing*

Following the most of the works that address speech separation and speech enhancement task, the audio original waveforms were pre-processed as follows: First, the audio files were resampled to 16 kHz. Then, over the resampled audio waveforms Short Time Fourier Transform (STFT) was applied using Fast Fourier Transform (FFT) size of 512, Hann window of length 25 ms, and hop length of 10 ms. The spectrogram $|x|^p$ of each input audio sample was obtained performing power-law compression of the STFT magnitude with p = 0.3. Finally, the data was normalized per-speaker with 0 mean and 1 standard deviation. To reconstruct the clean audio, on the post-processing stage, the inverse STFT to the estimated clean spectrogram was applied using the phase of the noisy input signal.

3.2.2. *Video pre-processing*

First we need to generate frames every 10 ms to align the visual and audio features. Over each frame we compute the optical flow with the method explained in 2.2. However, due to the nature of event-driven cameras, the number pixels that generate events in each frame is different and therefore, the number of video features in each frame is different. To avoid this problem, we generate regions of same size across the 100×50 pixels.

For each location we compute the mean of the x component and y component of optical flow and the event-rate, the number of events on each location at each frame. For example, with regions of 10×10 pixels we have a total of 50 regions and if we compute the event-rate and the mean of the x and y components, we have 150 video features. In Figure 2 an example of the x component and y component of optical flow for a specific frame can be seen.

3.2.3. *Deep Neural Network*

The RNN model consists of 5 stacked Bidirectional Long Short-Term Memory (BiLSTM) layers, with 250 neurons in each layer. The inputs of the model are the audio and visual features concatenated. The output of the network is an Ideal Amplitude Mask (IAM) and the loss function J_{mr} [17].

The models are trained using the Adam optimizer and 20% of dropout to avoid overfitting. Each model is trained up to 500 epochs and early stopping is applied on the validation set to stop the training process.

4. RESULTS

To measure the performance of each model, we use the well known source-to-distortion ratio (SDR) to measure the separation of target speech from the concurrent speech, and PESQ [27] to measure the quality of cleaned speech (i.e. the speech enhancement measure).

	SDR	PESQ
Noisy signal	0.21	1.94
Frame-based approach [17]	**7.37**	**2.65**
Event-based approach (150 features)	7.03	**2.65**
Event-based approach (400 features)	6.58	2.59
Event-based approach (LSTM)	3.79	2.22

Table 1. GRID dataset results.

Table 1 shows the results yielded with different models. To train the first model we use 150 video features as input (concatenated with the audio features). These 150 features correspond to the x and y components of the optical flow and the event-rate (the number of events) for each of the 50 regions (10×10 pixels each region). It can be seen that the results are quiet good, with higher than 7.0 SDR and on pare with the frame-based approach on PESQ performance. This shows that, although the original GRID dataset is frame based, the event-based approach can work almost as good as frame-based approach. It is noteworthy that in the frame-based approach there is need to perform a linear interpolation to align video features with the audio features, something that is similar to what we did when we upscale the original video to 100 fps. However, from our dataset we have to record the video with event-based cameras from a high definition screen, something that generates noise on the event-based videos. Even with these drawbacks, it can be said that event-based approach is almost on pare with the frame based approach.

In the next experiment we decrease the size of each region to 5×5 pixels in order to have more localized video features. However, the number of input features increases enormously. That is why we only used x and y components of the optical flow (400 input features in total). Although the results are good (6.58 SDR and 2.59 PESQ), they are not close to those achieved with 150 input features. One of the drawbacks of BiLSTMs used in the previous experiments is that they need to pass all the features forward and backward before giving a prediction. That means that BiLSTM is not a suitable RNN architecture to work on real-time. That is why we carried out one final experiment using LSTM instead of BiLSTM. However, the results show that the performance of deep LSTM is far from that yielded by the models with BiLSTM. That means that neither BiLSTM, neither deep LSTM are the best methods to address the Cocktail Party problem on real time.

5. CONCLUSIONS AND FUTURE WORK

This work presents a RNN for speaker target extraction on multi-talker environment using event-driven camera for the GRID dataset. We show that although this approach does not outperform the frame based approach, the performance of the presented method is almost on pair to the frame-based approach. This work also show that the x and y components of the optical flow from the lip region can be useful video features on speech separation. To the best of our knowledge, this is the first work that uses event-driven cameras to address the speech separation task.

However, due to the frame-based nature of the original dataset, the processed ATIS video signal is not as sharp as the original one, actually, it is a more noisy. Therefore, one can think that with an event-based original video the results achieved will improve. On the other hand, the BiLSTM is not a suitable RNN architecture for speech separation in real-time but the performance of LSTM is far to be comparable o that of BiLSTM.

One of the main drawback of event-driven cameras is the lack of audio-visual benchmark datasets. Therefore, future work will extend to generate an audio-visual dataset for speaker target extraction in event-driven fashion. With this dataset it will be easier to study the potential of using event-driven cameras for speech separation and enhancement. It is interesting to appoint that with event-driven cameras, unlike with frame-based cameras, the temporal limitation to perform speech separation is the temporal window to calculate the STFT. Thus, in order to do speech separation in real-time we have to move towards to new methods to extract audio features and other architectures different to RNNs. One interesting approach might be to work with audio features in time domain instead of in time-frequency domain [8] and using the methodology presented in this work to extract video features. Besides, another interesting research line might be working on asynchronous fashion using event-driven audio and video features with Spiking Neural Networks (SNN).

6. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, , no. 25, pp. 975–979, 1953.

[2] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Atten Percept Psychophys*, vol. 77, no. 5, pp. 1465–1487, 2015.

[3] Elana Zion Golumbic, Gregory B. Cogan, Charles E. Schroeder, and David Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party"," *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[4] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech*, 2016.

[5] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[6] C. Han, Y. Luo, and N. Mesgarani, "Online deep attractor network for real-time single-channel speech separation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 361–365.

[7] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 696–700.

[8] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal timefrequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.

[9] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures.," in *Interspeech*, 2017, pp. 2655–2659.

[10] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[12] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[13] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu, "Time domain audio visual speech separation," 2019.

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112:1–112:11, July 2018.

[15] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[16] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," 2017, pp. 3051–3055.

[17] Giovanni Morrone, Sonia Bergamaschi, Luca Pasa, Luciano Fadiga, Vadim Tikhanoff, and Leonardo Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6900–6904.

[18] A. Glover and C. Bartolozzi, "Robust visual tracking with a freely-moving event camera," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 3769–3776.

[19] L. A. Camuas-Mesa, T. Serrano-Gotarredona, S. Ieng, R. Benosman, and B. Linares-Barranco, "Event-driven stereo visual tracking algorithm to solve object occlusion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4223–4237, Sep. 2018.

[20] M. Iacono, S. Weber, A. Glover, and C. Bartolozzi, "Towards event-driven object detection with off-the-shelf deep learning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 1–9.

[21] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza, "Event-based motion segmentation by motion compensation," 2019.

[22] J. Maro, G. Lenz, C. Reeves, and R. Benosman, "Event-based visual gesture recognition with background suppression running on a smartphone," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, May 2019, pp. 1–1.

[23] X. Li, D. Neil, T. Delbruck, and S. Liu, "Lip reading deep network exploiting multi-modal spiking visual and auditory sensors," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2019, pp. 1–5.

[24] A. Savran, R. Tavarone, B. Higy, L. Badino, and C. Bartolozzi, "Energy and computation efficient audio-visual voice activity detection driven by event-cameras," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 333–340.

[25] C. Posh, 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," in *IEEE Journal of Solid-State Circuits*, Jan. 2011, vol. 46, pp. 259–275.

[26] R. Benosman, C. Clercq, X. Lagorce, S. Ieng, and C. Bartolozzi, "Event-based visual flow," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 407–417, Feb 2014.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, May 2001, vol. 2, pp. 749–752 vol.2.