Multimodal Fusion of Appearance Features, Optical Flow and Accelerometer Data for Speech Detection.

Panagiotis Giannakeris[1], Stefanos Vrochidis[1], Ioannis Kompatsiaris[1]
[1]Centre for Research & Technology Hellas - Information Technologies Institute, Greece
{giannakeris,stefanos,ikom}@iti.gr

ABSTRACT

In this paper we examine the task of automatic detection of speech without microphones, using an overhead camera and wearable accelerometers. For this purpose, we propose the extraction of hand-crafted appearance and optical flow features from the video modality, and time-domain features from the accelerometer data. We evaluate the performance of the separate modalities in a large dataset of over 25 hours of standing conversation between multiple individuals. Finally, we show that applying a multimodal late fusion technique can lead to a performance boost in most cases.

1 INTRODUCTION

An increasing interest exists for applications that require automatic voice activity detection. It is significantly insightful to recognize the speech status of people gathered at crowded environments, such as meetings or conferences, as speech is one of the primary elements of social interaction.

This paper presents the algorithms and results from CERTH-ITI's participation to the No-Audio Multimodal Speech Detection task at MediaEval 2019 [2]. The task focuses on automatic speech detection using an overhead camera and wearable accelerometers. The camera records a meeting event where several individuals participate in standing conversations. Each subject wears a tri-axial accelerometer that captures body movement. The use of microphones is not suitable in many cases since they may introduce background noise from the environment, or be uncomfortable to wear, or even raise privacy concerns. In contrast, an overhead camera is not as invasive, and the accelerometers are isolated instruments free of environment noise.

2 APPROACH

2.1 Detecting Speech from Video

We aim to process short, non-overlapping, video segments in order to classify them into speech or not-speech status. For this purpose we chose to extract low-level descriptors for each frame that represent body pose movements and speech gestures and then aggregate the information along the short temporal windows.

The videos are all taken from a single overhead camera which captures the full meeting space. Each video clip is a cropped version of the full resolution video that shows the subject and the immediate surrounding space. The subjects move freely inside the room, changing conversation partners and as such the videos follow the subjects at all times. There are several challenges posed as a result of this particular setting:

- Facial characteristics are severely occluded. A subject's body may be partially occluded as well, as a result of his movements and interactions with others.
- Multiple other subjects may appear inside a subject's immediate area cross contaminating the video data.
- When the cropped region is moving to follow a subject global camera motion is introduced.
- The orientation of the video is not aligned with head pose orientation making it difficult to obtain structured information consistent with pose or gaze.

In order to deal with occlusions and the changing orientation of the human body we select to extract appearance features and specifically the Histogram of Oriented Gradients (HOG) descriptor in a spatial 3×3 grid. Therefore, 9 different HOG descriptors are obtained and concatenated to form the HOG vector of a frame. We hypothesize that using HOG features in this manner we introduce some structure to the final representation regarding: (a) the primary subject's pose orientation and (b) the surrounding area elements which may consist of other people as well as background space.

To capture gestures and body movements from the speaker we compute dense optical flow for each frame. Then, we extract Histogram of Optical Flow (HOF) features in a spatial grid as described above. The grid partitioning here should make our representations capable of describing movement in different areas of the frame. The surrounding environment may contain other people talking and moving which can indicate that the primary subject in the center is currently not speaking. It is expected in these cases that HOF descriptors in peripheral grid cells have higher values. To compensate for camera motion we also extract Motion Boundary Histogram (MBH) features for each cell of the spatial grid. HOF and MBH are generally known to have complementary benefits for activity recognition tasks.

All the low-level frame descriptors of the same type are L2 normalized and averaged across temporal windows of 20 frames and then concatenated together to form a single representation for each second. Since the annotations are provided for each frame, we assign the label that the majority of the frames hold in order to annotate each 1 second segment. We remove any black screen instances from the training set and since the classes are severely imbalanced we remove random negative samples as well to balance the training set. We chose under-sampling instead of over-sampling in order to avoid having duplicates in the training set. Finally a Linear SVM classifier is trained using cross-validation on a random split, leaving the 30% of the subjects out, to obtain the optimal value of the regularization parameter C.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

2.2 Detecting Speech from Accelerometers

We deal with the task of speech detection from accelerometers in a similar fashion. We slide non-overlapping windows of 20 steps to segment the continuous x, y, z signal values, computing the magnitude values in each window:

$$M = [m_1, m_2, \dots m_{20}], \quad m_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$$

Then the following time-domain features are extracted from the magnitude values:

(1) Kurtosis
(2) Interquartile range
(3) Mean value
(4) Standard Deviation
(5) Min and Max values
(6) Number of zero crossings

Again, due to the fact that we try to solve the task by classifying each temporal window, we remove random negative instances in order to balance the training set. A Linear SVM classifier is trained here as well, cross-validating on a random split, leaving the 30% of the subjects out, to obtain the optimal C.

2.3 Late Fusion

We deploy a late fusion mechanism in order to explore the multimodal nature of the task. We feed the visual and accelerometer SVMs with all the test samples, in order to obtain for each one a pair of distances from the two separating hyper planes respectively. Then, we assign the label that corresponds to the farthest absolute distance of the two. This simple late fusion mechanism can guarantee that the most confident classifier for a particular sample is trusted.

3 RESULTS AND ANALYSIS

In order to evaluate our speech detection algorithms we train our classifiers on videos taken from 54 subjects and test on videos from 16 unseen subjects. We report the Area Under Curve (AUC) metric for each test subject and each modality (Fig. 1). Also the mean AUC scores for all subjects is presented in Table 1 and the performance is compared with last year's participation on this task. Our video estimator has the lowest mean score with 61% mean AUC and the accelerometer estimator performs higher by nearly 5%. The late fusion scheme achieves the best result gaining another 2%, which looks promising given that our fusion scheme is a fairly simple one.

We hypothesize that the shortcomings of our video estimator lie on the ineffectiveness of our approach with respect to the frequent head pose orientation changes of the subjects. Nevertheless, it performs better by a good margin from the dense trajectories of [1] and the colorhist+LBP of [3], which enhances our belief that the spatial grid structure is a good first step towards making the video estimators achieve more competitive results in this task. Another step for improvement would be to detect the head pose of the primary subject and align the spatial grid accordingly to ensure that each cell encapsulates visual information from a similar position relative to the speaker across all subjects.

The accelerometer estimator yields a satisfying performance compared with other methods presented at a previous version of this task despite the fact that no frequency domain signal processing was performed. The under-sampling strategy during the training phase may be a factor of improvement in this case as well as for the video estimator.

Table 1: Comparison of mean AUC±std scores between speech detection algorithms.

Method	Accel	Video	Fusion
[1]	0.656±0.074	0.549±0.079	0.658±0.073
[3]	0.533±0.020	0.512± 0.021	0.535±0.019
Ours	0.649±0.066	0.614±0.067	**0.672±0.051**

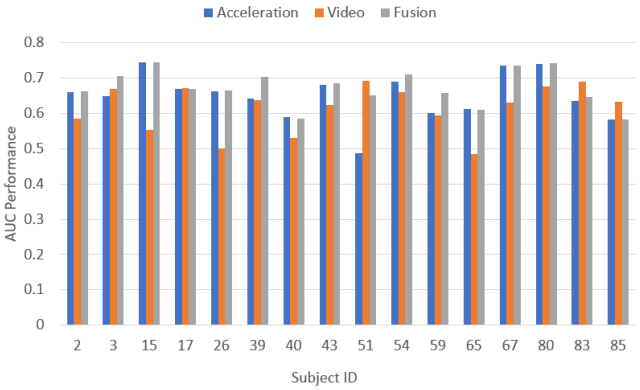

Figure 1: AUC scores for each test subject.

The fusion scores are better than the video and accelerometer scores for the majority of the test subjects. This shows that the confidence of the individual classifiers is actually a trustworthy measure for producing fused predictions in this task.

In this paper we tackle this task by classification of temporal segments. A promising alternative would be to deploy statistical modeling to the sequences of the extracted features, like Hidden Markov Models. Additionally, in neither technique did we adopt any speech behavioral modeling for the subjects which is a topic yet to be explored.

4 DISCUSSION AND OUTLOOK

In this work we have managed to achieve competitive results for the video modality regarding the task of no-audio speech detection and as a result we have made the late fusion estimator more effective using only the confidence of the individual classifiers. However, there is still a lot of experimentation to be done with early fusion techniques as well. Finally, we have proposed some key areas for improvement that should be examined thoroughly in order to achieve better performance from the separate modalities.

ACKNOWLEDGMENTS

This work was supported by SUITCEYES project funded by the European Commission under grant agreement No 780814.

REFERENCES

[1] Laura Cabrera-Quiros, Ekin Gedik, and Hayley Hung. Transductive Parameter Transfer, Bags of Dense Trajectories and MILES for No-Audio Multimodal Speech Detection. In *Proc. of the MediaEval 2018 Workshop*. 2018.

[2] Ekin Gedik, Laura Cabrera-Quiros, and Hayley Hung. No-Audio Multimodal Speech Detection task at MediaEval 2019. In *Proc. of the MediaEval 2019 Workshop*. Sophia Antipolis, France, Oct. 27-29, 2019.

[3] Yang Liu, Zhonglei Gu, and Tobey H Ko. Analyzing Human Behavior in Subspace: Dimensionality Reduction + Classification. In *Proc. of the MediaEval 2018 Workshop*. 2018.