Combining Body Pose and Movement Modalities for No-audio Speech Detection

Liandong Li, Zhuo Hao, Bo Sun
Beijing Normal University, China
bnulee@hotmail.com,hz@mail.bnu.edu.cn,tosunbo@bnu.edu.cn

ABSTRACT

Speech detection is important to automatic social behaviour analysis. In this paper, we describe our approach for no-audio speech detection. We estimate speaker pose and movement by both cameras and acceleration sensors. The multimodal features are combined and are utilized for per-second speech status prediction. The approach is tested on the MediaEval 2019 No-Audio Multimodal Speech Detection Task.

1 INTRODUCTION

Speech detection is important to automatic social behaviour analysis. There has been research focusing on this task using audio signal. However, the utilization of audio could be restricted in certain situations like noisy environment or when privacy-preserving is required [7]. Thus, some research explores this task by analysing human body behaviour [5, 7].

The 2019 No-Audio Multimodal Speech Detection Task[2] provides recorded data of speakers in social situations. Visual signal is captured through over-head cameras. Besides, tri-axial body accelerations are collected using wearable devices [3]. The modality signals are captured at 20Hz FPS, while the binary speaking status are annotated at each time-step.

In our work, we estimate human body pose and movement using the multimodal signals. The accelerators provide information about the overall movement of a person. However, it cannot describe body language which is expressed by the movement and pose of human body. Thus, we estimate body pose points in every frame to represent detailed body movement. The detail of proposed approach is described in Section 2.

2 APPROACH

In this section, we introduce our framework for no-audio speech detection. The framework consists of two components: multimodal representation and sequential classification. The first component extracts multimodal feature representations while the second component classifies the sequential data.

2.1 Multimodal Representation

Two types of modalities are utilized in our framework: tri-axial acceleration and visual. For tri-axial acceleration signal, we follow the method of [7]. Acceleration features are extracted from 3s windows with 1.5s overlap of the raw signal, absolute values of signal and the magnitude of the acceleration. Mean, variance and the power spectral density are calculated to form the final representation. The 70 dimension acceleration feature X_{acce} is standardised to have zero mean and unit variance.

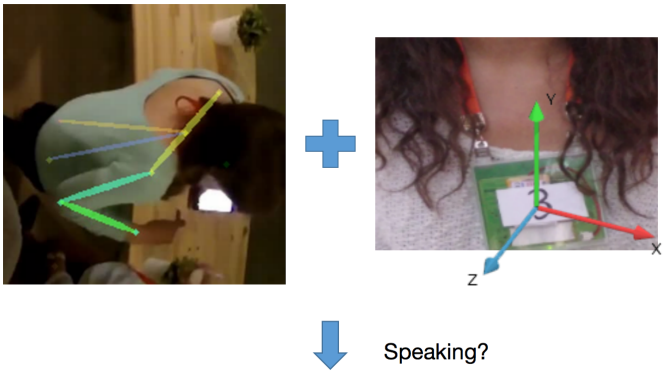

Figure 1: The scheme of our proposed method. Pose and acceleration signal are combined to predict the speech probabilities.

To extract visual representation, we utilize the Regional Multi-Person Pose Estimation Model [1]. This model consists of two components: Symmetric Spatial Transformer Network (SSTN) and Parametric Pose Non-Maximum-Suppression (p-Pose NMS). The SSTN network combines the spatial transformer network (STN), single-person pose estimator [6] (SPPE) and spatial detransformer network [4] (SDTN). The STN selects the potential area of human body and the SPPE estimate the body pose. Then, the estimated human pose is mapped back to the original image coordinate. The NMS is used to eliminate redundant pose estimations. During the training progress, Pose-Guided Proposals Generator is utilized to augment the training data. The extracted 34 dimension pose feature X_{pose} is normalized per speaker.

2.2 Sequential Logistic Classification

Given the multimodal feature representation of speech data, we follow a sequential classification scheme to predict the speech status. However, The typical sequential model may face the challenge of limited data that could be used for training. Instead, we use logistic regression model to classify speech statues at time-step. Then, a filter is used to eliminate the outlier of prediction.

Specifically, given feature X_{acce} and X_{pose}, we firstly concatenate them to be X_{con}. A logistic classifier $h_\Theta(X) = f(\Theta^T X_{con} + b)$ is utilized to train and predict the binary speech status y_t at time-step t. The sequential prediction $y_s(t)$ of a speaker is then convoluted by a N dimensional filter $g(t)$: $Y = y_s(t) * g(t)$, where $g(t) = [\frac{1}{N}, \frac{1}{N}, \ldots]$

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

Table 1: Cross validation AUC results with different value of N

N	3	5	7	9	11	13	15	17	19
AUC	0.5439	0.5486	0.5505	0.5517	0.5524	0.5528	0.5529	0.5527	0.5523

Table 2: Cross validation AUC results of different modalities

Method	Raw	Filtered
Accel	0.5281	0.5393
Video	0.5215	0.5305
Fusion	0.5341	0.5529

3 RESULTS AND ANALYSIS

We evaluate the performance of the proposed method on the MediaEval 2019 No-Audio Multimodal Speech Detection Task Dataset. The data is split into training set and testing set, with 54 and 16 videos respectively. Each video is 22 minutes long. Logistic regression model is trained to predict binary speech status at each timestep. For training the logistic regression model, 54×22×60=71280 samples are used. To evaluate the model performance, we utilize three-fold cross validation on the training set, with the result shown in Table 2. Results are measured by the Area Under the ROC Curve (AUC).

Testing is done on each video of testing set. After gaining the prediction on test videos, a N dimensional filter $[\frac{1}{N}, \frac{1}{N}, ...]$ is convoluted with the prediction probability vector to smooth the output. The dimension N is chosen through grid search on the training set. The testing results are shown in Table 3. From the result, we can see that the acceleration signal outperforms visual signal. Though the fusion prediction gets higher result on the training data, its testing performance is not as good as acceleration modality.

It is understandable that utilizing visual signal is very challenging in this task, especially with the limited video data. In spite of the large number of image frames that could be used to train image level classifier, it is actually difficult to tackle the task with image classifier. Considering that, we employed pose estimator trained on larger video dataset. However, the position of the camera from which the data was recorded is different from regular video. Thus, the pose estimator does not work well in this task, which brings large proportion of inaccurate pose features. In comparison, the acceleration signal is acquired by wearable physical sensor, which guarantees its reliability. Despite this, as we mentioned before, the acceleration signal is ambiguous about the local body movement. It is still important to explore more efficient and accurate visual feature representation.

4 DISCUSSION AND OUTLOOK

In this paper, we describe our work for no-audio speech detection. We estimate speaker pose and movement through cameras and acceleration sensors. The multimodal features are combined and utilized for per-second speech status prediction. Results show that acceleration modality outperforms visual modality. In the future, we will explore more efficient and accurate visual feature representation.

Table 3: Testing data AUC results

Subject ID	Accel	Video	Fusion
2	0.6735	0.4234	0.5742
3	0.6759	0.5127	0.6534
15	0.7465	0.4958	0.7438
17	0.6556	0.5555	0.6539
26	0.7023	0.4557	0.6225
39	0.5429	0.5131	0.5468
40	0.6123	0.4733	0.5701
43	0.6751	0.5484	0.6421
51	0.4020	0.2527	0.2588
54	0.6652	0.6763	0.7191
59	0.6067	0.6430	0.6611
65	0.5718	0.5812	0.6254
67	0.7621	0.6173	0.7590
80	0.8573	0.4529	0.7941
83	0.6234	0.5034	0.5682
85	0.5274	0.5100	0.5208
Mean	0.6438	0.5134	0.6196

ACKNOWLEDGMENTS

This work is supported by the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *ICCV*.

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

[3] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. 2013. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 207–210.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and others. 2015. Spatial transformer networks. In *Advances in neural information processing systems*. 2017–2025.

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

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*. Springer, 483–499.

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