Multimodal Fusion of Body Movement Signals for No-audio Speech Detection

Xinsheng Wang[1,2], Jihua Zhu[1], Odette Scharenborg[2]
[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
[2]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
wangxinsheng@stu.xjtu.edu.cn,zhujh@xjtu.edu.cn,o.e.scharenborg@tudelft.nl

ABSTRACT

No-audio Multimodal Speech Detection is one of the tasks in MediaEval 2020, with the goal to automatically detect whether someone is speaking in social interaction on the basis of body movement signals. In this paper, a multimodal fusion method, combining signals obtained by an overhead camera and a wearable accelerometer, was proposed to determine whether someone was speaking. The proposed system directly takes the accelerometer signals as input, while using a pre-trained 3D convolutional network to extract the video features that work as input. Experiments on the No-audio Multimodal Speech Detection task show that our method outperforms all submissions of previous years.

1 INTRODUCTION

There is a close relationship between body movements, e.g., gesturing, and speaking status, i.e., whether someone is speaking or not. This might make it possible to determine whether a person is speaking by analyzing the person's body movements. This No-Audio Multimodal Speech Detection task of MediaEval 2020 focuses on analyzing the problem of determining the speaking status of standing subjects in crowded mingling scenarios with the information recorded by an overhead camera and a single body-worn triaxial accelerometer, hung around the neck of the subjects [1]. In this paper, we fuse the signals from these two modalities to perform the No-audio Speech Detection task. The details of the proposed approach will be described in the following section[1].

2 APPROACH

The architecture of the proposed method is shown in Fig. 1. The proposed model consists of three parts, i.e., AccelNet, VideoNet, and the fusion part for the accelerometer data input, the video input, and the multi-modality fusion respectively. According to the requirements of this task, the AccelNet and VideoNet are also designed to be able to predict the speaking status individually.

2.1 Data processing

In the provided database, video and accelerometer data were recorded with a duration of 22 minutes at 20Hz. For training, we segmented the video and accelerometer data into 11 segments, each of which has a duration of 2 minutes, resulting in a video segment with 2400 frames and an accelerometer data segment with a size of 3×2400.

2.2 AccelNet

As shown in Fig. 1, the AccelNet consists of 3 1-D convolution layers and a bi-directional GRU layer. Between every two adjacent convolutional layers, a batch normalization layer is adopted. The 3 convolution layers take kernel sizes of 5, 3, and 3 respectively, and take stride sizes of 5, 2, and 2 respectively, resulting in a feature with a receptive field of 23 frames, which is similar to the sampling rate of 20Hz. Therefore, we can assume that each frame out of the total of 120 frames from the last convolutional layer, with a dimension of 256, represents the movement status within a second. Intuitively, the speaking status in one moment would have a relationship with the previous and following several time steps, the bi-directional GRU, with 256 units, is adopted after the last 1-D convolutional layer to capture this relationship.

Concatenating the features of two directions at each time step, the bi-directional GRU results in a 512-d feature with a sequence length of 120. Then this feature will be concatenated with the video feature to perform the multimodal speech detection task. In order for the AccelNet to detect speaking status on the basis of the accelerometer data only, a linear transformation followed by a sigmoid layer can be added after the bi-directional GRU.

2.3 VideoNet

The C3D [7] pre-trained on Sports-1M [4] is adopted to extract the video features. The video was recorded with a frequency of 20Hz, while the C3D model only uses 16 consecutive frames as context to obtain the 3D convolutional features. In practice, we dropped the last 4 frames within each second in the video, so that we can use the C3D to extract video features of each second, resulting in 120 feature vectors with a dimension of 512 for each video segment (2 minutes). The C3D features go through a bi-directional GRU, with 256 units, before being fused with the accelerometer features.

Similar to the AccelNet, the output of VideoNet can also be used for unimodal speech detection.

2.4 Fusion and objective function

The early fusion strategy is adopted in this paper. Specifically, the accelerometer feature from the AccelNet and the visual feature from the VideoNet are concatenated, resulting in a feature with 1024 dimensions and 120 frames. Two linear transformation layers are used to transform the feature dimension from 1024 to 1, and then a sigmoid layer is utilized after the last linear transformation layer to obtain the final prediction probability.

[1]The code of the proposed method can be found at: https://github.com/xinshengwang/No-audio-speech-detection

MediaEval'20, December 14-15 2020, Online

To train the model, the binary cross-entropy loss is adopted on the frame level. First, the AccelNet and VideoNet are trained for the unimodal prediction task individually. Next, the pre-trained models are used in the multimodal task. During multimodal task training, we only updated the fusion network, i.e., two linear transformation layers, while keeping the parameters of the pre-trained AccelNet and VideoNet fixed.

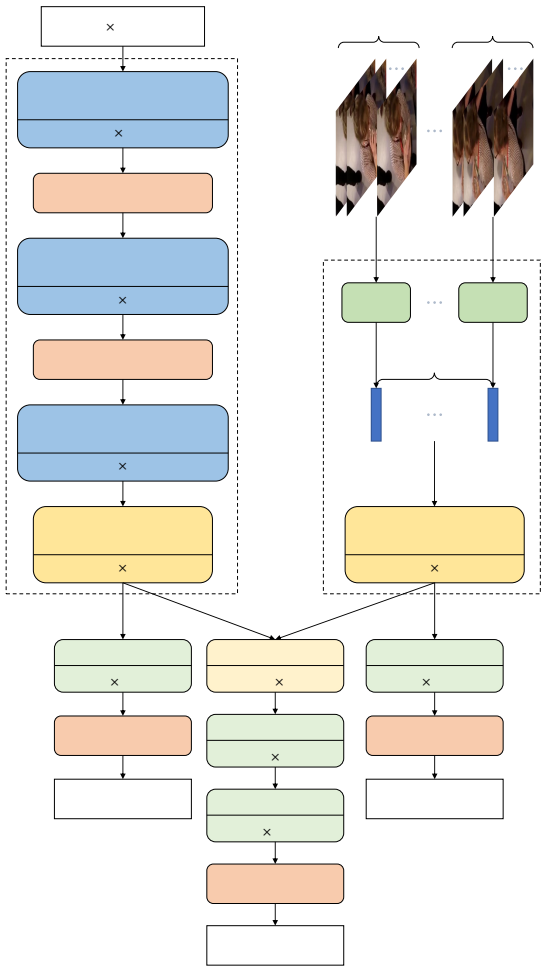

Figure 1: The proposed multimodal speech detection network.

3 RESULTS

In order to evaluate our speech detection approach, we followed the given split method of the No-audio Speech Detection task. The model was trained on data from 54 subjects and tested on data from 16 unseen subjects that non-overlap with the subjects in the training set. We report the Area Under Curve (AUC) metric for each test subject and each modality. The mean AUC scores computed over all test subjects are shown in Table 1, while the AUC scores for each test subject separately are shown in Fig. 2.

Table 1: Performance of each of the previously submitted results and our proposed method for the unimodal and multimodal speech detection tasks. Bold indicates best result.

Method	Accel	Video	Fusion
Cabrera-Quiros et al. [2]	0.656±0.074	0.549±0.079	0.658±0.073
Liu et al. [6]	0.533±0.020	0.512± 0.021	0.535±0.019
Giannakeris et al. [3]	0.649±0.066	0.614±0.067	0.672±0.051
Li et al. [5]	0.644	0.513	0.620
Vargas et al. [8]	0.692	0.552	0.693
The proposed model	0.689±0.094	0.656±0.076	**0.712±0.081**

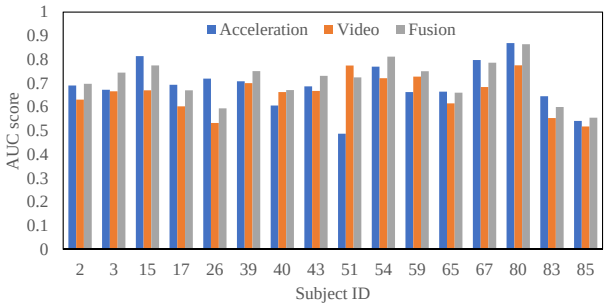

Figure 2: AUC scores for each test subject.

In Table 1, our method is compared with the submission results of pervious years. Our method achieves the better performance on the multimodal speech detection task. On the unimodal tasks, our AccelNet outperforms our VideoNet. Moreover, our accelerometer data-based method is only slightly lower than that of [8], while our video-based method achieves a much higher performance than the second best approach [3], indicating the good performance of C3D on extracting video features and also the good design of the VideoNet. The best performance of our multimodal result benefits from the good performance of the VideoNet.

From Fig. 2 we can see that the accelerometer modality-based method does not always outperform the video-based method, indicating that the signals from the accelerometer and video could be complementary, which could explain the higher performance of the fusion of the two modalities compared to the unimodal methods. However, fusion did not lead to an improved performance for all individual test subjects (see subjects 17 and 83), and a better fusion method should be considered in the future.

4 CONCLUSION

In this paper, we proposed a multimodal speech detection model, with video and accelerometer data as input. Our model showed competitive results on the unimodal speech detection tasks with either video or accelerometer data as input, and it outperformed previous methods on the multi-modal task which uses both types of input.

REFERENCES

[1] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 2018. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates.