No-Audio Multimodal Speech Detection task at MediaEval 2019

Ekin Gedik[1], Laura Cabrera-Quiros[3,1,2], Hayley Hung[1]
[1]Delft University of Technology, Netherlands
[2]Instituto Tecnológico de Costa Rica, Costa Rica.
[3]Eindhoven University of Technology, Netherlands
{l.c.cabreraquiros,e.gedik,h.hung}@tudelft.nl

ABSTRACT

This overview paper provides a description of the No-Audio multimodal speech detection task for the MediaEval 2019. Same as the first edition that was held in 2018, the task again focuses on the estimation of speaking status from multimodal data. Task participants are provided with cropped videos of individuals interacting freely during a crowded mingle event, captured by an overhead camera. Each individuals tri-axial acceleration throughout the event, captured with a single badge-like device hung around the neck, is also provided. The goal of this task is to automatically estimate if a person is speaking or not using these two alternative modalities. In contrast to conventional speech detection approaches, no audio is used for this task. Instead, the automatic estimation system must exploit the natural human movements that accompany speech. The task seeks to achieve competitive estimation performance compared to audio-based systems by exploiting the multi-modal aspects of the problem.

1 INTRODUCTION

Speaking status is one of the most essential elements of social behaviour since it is one of the key behavioural cues that is used for studying conversational dynamics in face to face settings [10]. This task focuses on the automatic detection of speaking status. Previous work has shown the benefit of deriving features from speaking turns (which can be obtained from the speaking status of different people) for estimating many different social constructs such as dominance [8], or cohesion [7].

However, automated analysis of conversational dynamics in large unstructured social gatherings is an under-explored problem despite the fact that attendance of these type of events have shown to be contributing factors for career and personal success [11]. The majority of speaking status detection work focuses on utilising the audio signal mainly captured through microphones. However, most unstructured social gatherings such as parties or cocktail events tend to have inherent background noise due to the nature of these events. Because of this restriction, recording audio in such cases is challenging. For example, to collect good quality audio signals, participants need to wear personal headset microphones. This requires uncomfortable and intrusive equipment to be worn. Recording audio can also have certain negative connotations as it can be perceived as an invasion of privacy to have the precise verbal contents of a conversation to be recorded.

Estimating a person's speaking status using the provided *video* and *wearable acceleration* data instead of audio is the main goal of this task. The accelerometer is embedded inside a smart ID badge which is hung around the neck. These modalities are easy to use and replicate for crowded environments such as conferences, networking events, or organisational settings. This approach also enables a more privacy-preserving method of extracting socially relevant information.

The presence of body movements such as gesturing while speaking has been well-documented by social scientists [9]. Thus, an automatic estimation system should exploit the natural human movements that accompany speech. Past work which estimated speaking status from a single body worn tri-axial accelerometer [5, 6] and other work that used video to estimate speaking status during standing conversations [4] motivated this task.

Despite these efforts, one of the major challenges of these alternative approaches has been achieving competitive estimation performance against audio-based systems. As of 2019, exploiting the multi-modal aspects of the problem is still under-explored and this is the main focus of this challenge.

2 TASK DETAILS

2.1 Unimodal estimation of speaking status

For this subtask, participants are expected to design and implement separate speaking status estimators for each modality. If participants prefer to focus on developing an estimator for only one of the modalities, they can use the provided baseline approach for the other modality. For the video modality, the algorithm will have a video of a person interacting freely in a social gathering (see Figure 1) as input and should provide a estimation of that persons' speaking status (speaking/non-speaking) estimation every second. Similarly, for the wearable modality, the method will have the wearable tri-axial acceleration signal of a person as input and must return a speaking status estimation every second. We provide baseline codes for each modality. The baseline using acceleration implements the logistic regression approach in [5] and the video baseline employs dense trajectories and multiple instance learning, as explained in [3].

2.2 Multimodal estimation of speaking status

For this subtask teams must provide an estimation of speaking status every second by exploiting both modalities together. Teams can use any type of fusion method they see fit [1]. The goal is to leverage the complementary nature of the modalities to better estimate the speaking status. Thus, teams are encouraged to go beyond basic fusion and really think about the impact of each modality on the estimation.

3 DATA

The data for this task is a subset of the MatchNMingle dataset [2], which is open to the research community. This dataset was

MediaEval'19, 27-30 October 2019, Sophia Antipolis, France

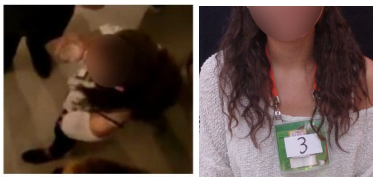

Figure 1: Alternative modalities to audio used for the task. Left: Individual video of each participant while interacting freely. Right: Wearable triaxial acceleration recorded by a device hung around the neck.

created as a resource to analyse unstructured mingle scenarios and seated speed dates.[1] The subset for this task contains data for 70 people who attended one of three separate mingle events for over 45 minutes. To eliminate the effects of acclimatisation, only 30 minutes in the middle of the event are used. Subjects were separated using stratified sampling to create the train and test sets (see Figure 2). Stratification was done with various criteria to ensure balanced distributions in both sets for speaking status, gender, event day, and level of occlusion in the video.[2] An additional segment of the data (orange in Figure 2) is left for the optional subject specific evaluation (see more in Section 4). Train and test sets provided to the participants this year are **entirely same** with the one used in last year's iteration, making comparisons possible between solutions of different years.

Task participants are provided with videos of individuals recorded at 20FPS, captured by an overhead camera. Due to the crowded nature of the events, there can be strong occlusions between participants in the video. The video for each person has been cropped from the entire frame and provided in separated videos. Note that due to the crowded nature of social gatherings, the cropped scenes do not just capture the behaviour of the person of interest, as cross contamination between bounding boxes does occur.

Each individual is also wearing a badge-like device, recording tri-axial acceleration at 20Hz. Task participants have access to the raw tri-axial acceleration, for which only the effect of gravity was compensated by subtracting the mean of each axis and normalising with the variance of each respective axis. All the data is synchronised.

Finally, binary speaking status (speaking/non-speaking) was annotated every frame by 3 different annotators. Inter-annotator agreement for a 2 minute segment of the data reported a *Fleiss' kappa* coefficient of 0.55.

4 EVALUATION

Due to class imbalance, we use the Area Under the ROC Curve (ROC-AUC) as the evaluation metric. Participants need to submit non-binary prediction scores (posterior probabilities, distances to the separating hyperplane, etc.).

The task will be evaluated using a subset of the data left as a test set (as shown by the red section of Figure 2). All the samples of this test set will be for subjects who are not present in the training set, as can be seen in Figure 2.

Required evaluation. For each subtask, each team must provide up to 5 runs with their non-binary estimations for a persons' speaking status. The evaluation will be made in a **person independent manner**. This means that the training set **will not** include any data from the participants in the test set.

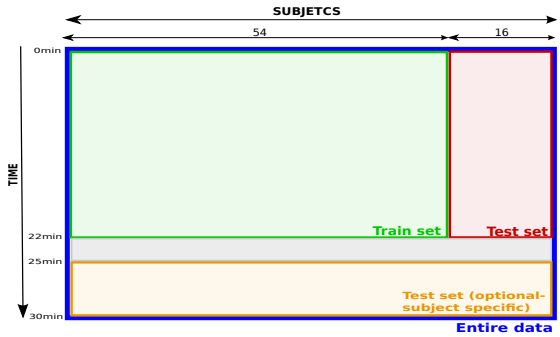

Figure 2: Separation of train and test set for the task.

Optional evaluation. As an optional task, teams can also submit up to 5 runs (per person) using a **person dependent** training scheme. To do so, a separate 5 minutes interval for all people in the training set is provided, as shown by the orange section in Figure 2. In this setting, samples originating from the same subject (which are temporally non-adjacent) can be used in the training in addition to data from the other subjects. This evaluation can be a sanity check as the performance of the method, in theory, should perform better when trained on a specific person rather than other people.

5 DISCUSSION AND OUTLOOK

This task aims to investigate the use of alternative modalities for the detection of speaking status. With the information gained from this task, we aim to learn more about the nature of the connection between speaking and body movements, providing valuable insights for both social science and the multimedia communities. Moreover, we expect these insights will pave the way for solutions that are privacy-preserving and scalable.

In its first edition in 2018, we saw that the task was received as untypical and challenging. The participation for the task was limited and no participant managed to provide a better performance or explanation than the baseline method provided by the organisers. Various properties make the task challenging. The chosen modalities are not for directly sensing the physical manifestation of the task (audio). Acceleration and video provides an indirect way of sensing speaking and requires carefully designed approaches that can exploit the connection between body movements and speech. Secondly, the connection between speech and body movements has been found to be person-specific [5], further complicating the task. In its current edition, we aimed to increase participation by providing the baseline codes for each modality.

In addition, we are reaching out to different communities (affective computing, multimedia, computer vision, and speech). We believe each of these communities can bring their own expertise to the task. In the following years as well as augmenting the data, we aim to focus on the person dependent task and its fundamental differences from a person independent training setting.

ACKNOWLEDGMENTS

This task is partially supported by the Instituto Tecnológico de Costa Rica and the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

[1] MatchNMingle is openly available for research purposes under an EULA at http://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/

[2] Occlusion levels can be requested if needed for training set.

REFERENCES

[1] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16, 6 (2010), 345–379.

[2] 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. *IEEE Transactions on Affective Computing* (2018).

[3] Laura Cabrera-Quiros, David MJ Tax, and Hayley Hung. 2019. Gestures in-the-wild: detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. *IEEE Transactions on Multimedia* (2019).

[4] Marco Cristani, Anna Pesarin, Alessandro Vinciarelli, Marco Crocco, and Vittorio Murino. 2011. Look at whoâĂŹs talking: Voice activity detection by automated gesture analysis. In *International Joint Conference on Ambient Intelligence*. Springer, 72–80.

[5] Ekin Gedik and Hayley Hung. 2017. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing* 21, 4 (2017), 723–737.

[6] 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.

[7] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[8] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling Dominance in Group Conversations Using Nonverbal Activity Cues. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 3 (2009), 501–513.

[9] David McNeill. 2000. *Language and gesture*. Vol. 2. Cambridge University Press.

[10] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87.

[11] Hans-Georg Wolff and Klaus Moser. 2009. Effects of networking on career success: a longitudinal study. *Journal of Applied Psychology* 94, 1 (2009), 196.