No-Audio Multimodal Speech Detection in Crowded Social Settings task at MediaEval 2018

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

ABSTRACT

This overview paper provides a description of the automatic Human Behaviour Analysis (HBA) task for the MediaEval 2018. In its first edition, the HBA task focuses on analyzing one of the most basic elements of social behavior: the estimation of speaking status.

Task participants are provided with cropped videos of individuals while interacting freely during a crowded mingle event that was captured by an overhead camera. Each individual is also wearing a badge-like device hung around the neck recording tri-axial acceleration.

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

This task focuses on analyzing one of the most basic elements of social behavior: speaking status. This information is quite valuable since it is one of the key behavioural cues that is used for studying conversational dynamics in face to face settings [9]. Previous work has also 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 [7], or cohesion [6].

However, the automated analysis of conversational dynamics in large unstructured social gatherings such as networking or mingling events, 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 [10].

The majority of speaking status detection works focus on exploiting the audio signal but most unstructured social gatherings such as parties or cocktail events (also called mingle scenarios) tend to have inherent background noise due to the nature of these events. Because of this restriction, recording audio in such cases in an easy manner is challenging. For example, to collect good quality audio signals, participants need to wear personal headset microphones to minimise ambient noise. However, 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.

The goal of the task is to automatically estimate if a person is speaking or not using alternative modalities instead of audio. The specific modalities used in this task are video and wearable acceleration. The accelerometer is embedded inside a smart ID badge which is hung around the neck. These modalities are easy to use and replicate for these type of crowded environments.

The presence of body movements such as gesturing while speaking has been well-documented by social scientists [8]. Thus, an automatic estimation system should exploit the natural human movements that accompany speech (e.g. conversational gestures). This alternative approach for speaking status detection also enables a more privacy-preserving method of extracting socially relevant information and has the potential to scale to settings where recording audio may be impractical.

This approach is motivated by past work which estimated speaking status from a single body worn tri-axial accelerometer, hung around the neck [4, 5]. This form of sensing could be embedded into a smart ID badge that could be used in settings such as conferences, networking events, or organizational settings. In addition, other works have used video to estimate speaking status during standing conversations [3].

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

2 TASK DETAILS

This task consists of two subtasks; unimodal and multimodal estimation.

2.1 Unimodal estimation of speaking status

For this subtask participants must design and implement separate speaking status estimators for each 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) 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.

Due to the evaluation metric used in this task (see more in Section 4), all estimations must be non-binary prediction scores (e.g. posterior probabilities, distances to the separating hyperplane, likelihood, etcs).

MediaEval'18, 29-31 October 2018, France

Participants are allowed to submit up to 5 runs per modality. The output of each run should consist of n vectors (where n is the number of subjects in the test set) with the estimations every second.

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 (early, late or hybrid fusion) [1], and are allowed to submit up to 5 runs for this subtask.

The goal of this subtask is to leverage the complementary nature of the modalities to better estimate the speaking status. Thus, teams are encouraged to go beyond a normal fusion (e.g. concatenation or majority voting) and really think about the impact of each modality on the estimation. For example, if the occlusion level in the video is high, is it meaningful to give the same importance to both modalities?

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 created as a resource to analyze 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 (cocktail parties) for over 45 minutes. To eliminate the possible effects of acclimatization (e.g. people entering mingle area) only 30 minutes in the middle of the event are used. These subjects were separated using stratified sampling to create the train and test sets (see Figure 2). This 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).

Task participants are provided with videos of individuals recorded at 20FPS participating in a conversation that was captured by an overhead camera. Note that due to the crowded nature of the events, there can be strong occlusions between participants in the video. Although the interactions were simultaneously recorded by up to 3 cameras, the video for each person has been cropped from the entire frame and provided in separated videos. Note that the cameras were arranged to ensure maximum coverage of the scene and the views do not have sufficient overlap for 3D visual processing. Note that due to the crowded nature of social gatherings, the cropped scenes do not just capture the behavior 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 for by substracting the mean of each axis and normalizing with the variance of each respective axis. All the data is synchronized.

[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.

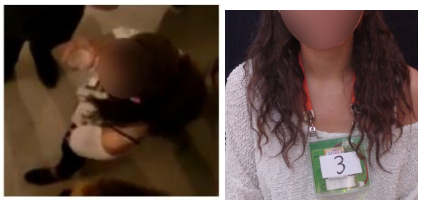

Figure 1: Alternative modalities to audio used for the HBA MediaEval task. Left: Individual video of each participant while interacting freely. Right: Wearable triaxial acceleration recorded by a device hung around the neck.

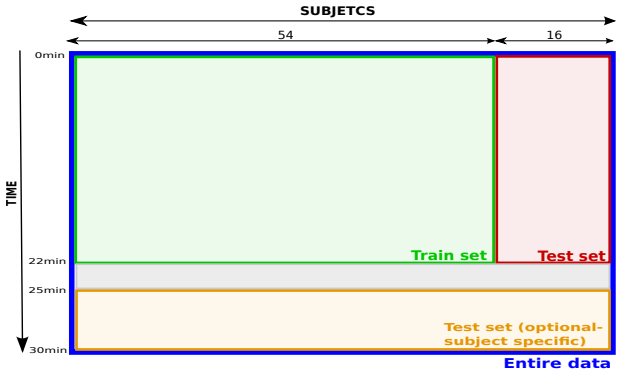

Figure 2: Separation of train and test set for HBA task.

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

Since the classes are severely imbalanced, we will be using the Area Under the ROC Curve (ROC-AUC) as the evaluation metric. Thus, 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 **independent manner**. This means that all samples are provided to the algorithm together, irrespective of the subject that the samples came from. Note that the test samples we provide will be the samples taken from people who are not in the training data.

Optional evaluation. As an optional evaluation, teams can also submit up to 5 runs (per person) using a **person specific** 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. Thus, only samples generated from the same subject are provided to the classifier, so one classifier is trained for each person with test results output per person-specific classifier.

This alternative evaluation can be a useful sanity check as the performance of the method, in theory, should perform better when trained on a specific person rather than other people.

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.

REFERENCES

[1] P.K. Atrey, M.A. Hossain, A. El Saddik, and M.S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* (2010).

[2] L. Cabrera-Quiros, A. Demetriou, E. Gedik, Meij v.d. L, and H. 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] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino. 2011. Look at who's talking: Voice activity detection by automated gesture analysis. *Workshop on Interactive Human Behavior Analysis in Open or Public Spaces, International Joint Conference on Ambient Intelligence* (2011).

[4] E. Gedik and H. Hung. 2016. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing* (2016).

[5] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. 2013. Classifying social actions with a single accelerometer. In *International Joint Conference on Pervasive and Ubiquitous Computing (UBIComp)*.

[6] 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.

[7] 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.

[8] D. McNeill. 2000. *Language and Gesture.* Cambridge University Press.

[9] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. 2012. Bridging the gap between social animal and unsocial machine: a survey of social signal processing. *IEEE Transactions on Affective Computing* (2012).

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