No-Audio Multimodal Speech Detection Task at MediaEval 2020

Laura Cabrera-Quiros[1], Jose Vargas[2], Hayley Hung[2]
lcabrera@itcr.ac.cr,{j.d.vargasquiros,h.hung}@tudelft.nl
[1]Instituto Tecnológico de Costa Rica, Costa Rica
[2]Delft University of Technology, Netherlands

ABSTRACT

This overview paper provides a description of the No-Audio multimodal speech detection task for MediaEval 2020. Similar to the previous two editions, the participants of this task are encouraged to estimate the speaking status (i.e. person speaking or not) of individuals interacting freely during a crowded mingle event, from multimodal data. In contrast to conventional speech detection approaches, no audio is used for this task. Instead, the automatic estimation system proposed must exploit the natural human movements that accompany speech, captured by cameras and wearable sensors. Task participants are provided with cropped videos of individuals while interacting, captured by an overhead camera, and the tri-axial acceleration of each individual throughout the event, captured with a single badge-like device hung around the neck. This year's edition of the task also focuses on investigating posible reasons for interpersonal differences in the performances obtained.

1 INTRODUCTION

Speaking status is one of the key signals that is used for studying conversational dynamics in face to face settings [10]. From the speaking status of multiple people one can also derive speaking turns, and other features that have shown beneficial for estimating many different social constructs such as dominance [8], or cohesion [7]. Overall, automated analysis of conversational dynamics in large unstructured social gatherings is an under-explored problem despite the relevance of such events [11], and automated speaking detection one of its key components.

The majority of works regarding speaking status detection focuses on utilizing the audio signal captured by microphones. However, most unstructured social gatherings such as parties or cocktail events tend to have inherent background noise and to collect good quality audio signals, participants need to wear uncomfortable and intrusive equipment. Recording audio also risks to be perceived as an invasion of privacy due to the access to the precise verbal contents of the conversation, further limiting the natural behavior of the individuals involved. Because of these restrictions, recording audio in such cases is challenging.

As a suitable alternative, the main goal of this task is to estimate a person's speaking status using *video* and *wearable acceleration* data from a smart ID badge which is hung around the neck, instead of audio. Such alternative modalities are more privacy-preserving, and easy to use and replicate for crowded environments such as conferences, networking events, or organizational settings.

Body movements such as gesturing tend to co-occur with speaking, as it has been well-documented by social scientists [9]. Thus, an automatic estimation system should exploit the natural human movements that accompany speech. This task is motivated by such insights, and past work which estimated speaking status from a single body worn tri-axial accelerometer [5, 6] and video [4].

Despite many efforts, one of the major challenges of these alternative approaches has been achieving competitive estimation performance against audio-based systems. Moreover, results from past editions of this task have shown a significant difference in the performance of different individuals, and lower performances for a particular subset of them (failure cases) not fully understood yet.

2 TASK DETAILS

2.1 Unimodal estimation of speaking status

Participants are encouraged to design and implement separate speaking status estimators for each modality. However, baseline approaches for each modality are provided, in case they prefer to focus on improving an estimator for only one of the modalities, or the fusion technique. The baseline using acceleration implements the logistic regression in [5] and the video baseline employs dense trajectories and multiple instance learning, as explained in [3].

For the video modality, the input will be a video of a person interacting freely in a social gathering (see Figure 1), and a estimation of that persons' speaking status (speaking/non-speaking) should be provided every second. For the wearable modality, the method will have the wearable tri-axial acceleration signal of a person as input and must also return a speaking status estimation 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 [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.

2.3 Analysis of failure test cases

As a new addition for this year's edition, teams must analyze the differences in the performance results for the test set, focusing on the three subjects with the lower performances, and hypothesize about the reasons the method underperforms for these persons. Participants are encouraged to think about the circumstances for the subjects (e.g. occlusion) or interpersonal differences that could explain such dissimilarities.

3 DATA

A subset of the MatchNMingle dataset[1] [2] is used for this task. It contains data for 70 people who attended one of three separate

MediaEval'20, 14-15 December 2020, Online

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

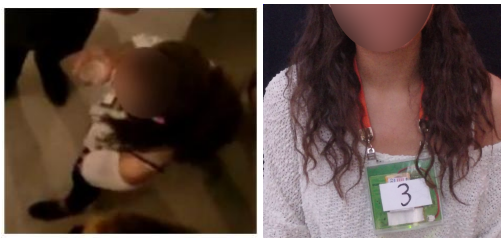

Figure 1: Alternative modalities to audio used for the task. Left: Individual video of each participant while interacting freely. Right: Wearable triaxial acceleration recorded by a device hung around the neck.

mingle events for over 45 minutes. To eliminate the effects of acclimatization, only 30 minutes in the middle of the event are used. Subjects were separated using stratified sampling to create the train (54 subjects) and test sets (16 subjects). 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 was created for the optional subject specific evaluation of the task (see more in Section 4). While the dataset used this year is the same as the one used in previous versions of the challenge, making comparisons possible between solutions of different years, focus is given to the differences shown by the 16 subjects in the test set.

Videos were captured from an overhead view at 20FPS. The rectangular (bounding box) area around each subject has been cropped, in such a way that a video is provided per person. Important challenges in the automatic analysis of this data include the significant amount of cross-contamination and occlusion, both in self-occlusion and occlusion by other subjects, due to the crowded nature of the event (cocktail party).

Subjects also wore a badge-like body-worn accelerometer (see Figure 1), recording tri-axial acceleration at 20Hz. These acceleration readings were processed via whitening applied per axis. All video and wearable the data is synchronized.

Finally, binary speaking status (speaking/non-speaking) was annotated by 3 different annotators. Inter-annotator agreement was calculated on a 2 minute segment of the data, which resulted in a *Fleiss' kappa* coefficient of 0.55.

4 EVALUATION

The Area Under the ROC Curve (ROC-AUC) is used as evaluation metric, since it is robust against class imbalance which exists in our scenario. Therefore, participants need to submit continuous prediction scores (posterior probabilities, distances to the separating hyperplane, etc.) obtained by running their method on the evaluation set. These scores will be compared against the test labels, which are not available to participants.

Required evaluation. For unimodal and multimodal estimations, each team must provide up to 5 runs with their scores for a persons' speaking status. As mentioned, the evaluation set does not contain any data from participants in the test set to achieve person independent results.

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

Optional evaluation. Teams may optionally submit up to 5 runs (per person) using person dependent training. To do so, a separate 5 minutes interval for all people in the training set is provided. Thus, samples and labels from the same subject can be used to train or fine-tune and then test for a specific test subject's data, which is temporally to adjacent to the training samples. This method would be expected to perform better when trained or fine-tuned on the target person rather than other people.

5 DISCUSSION AND OUTLOOK

With this task, we aim to support the study of speaking status detection in the wild using alternative modalities to audio. We aim to learn more about the connection between speaking and body movements, expecting that in the future this will bring on valuable insights for both the social science and multimedia communities.

Participation in previous editions of the task has been limited, with only small improvements over the baseline. We believe this is due to the variety of ways in which this task is atypical. For example, the connection between speech and body movements has been found to be person-specific [5]. Additionally, the interaction between the two modalities of interest (chest acceleration and video) is not traditionally explored, i.e. the combination of these two modalities is not common. This leaves open opportunities to explore their complimentarity, to better understand in which situations one modality is more reliable over the other, and develop or apply appropriate fusion strategies.

Moreover, differences in the performances between test subjects was consistently found in previous editions, further supporting past research [5]. Thus, this year participants are encouraged to focus on such failure cases and hypothesize about the reasons of such dissimilarities.

We are reaching out to different communities (affective computing, multimedia, computer vision, and speech), as 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 include and explore the implications of the spatial social component of the mingle (e.g. F-Formations) on the speaking status detection.

ACKNOWLEDGMENTS

This task is partially supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

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 who's 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.