Group-Level Focus of Visual Attention for Improved Next Speaker Prediction

Chris Birmingham
cbirming@usc.edu
University of Southern California
Los Angeles, California, USA

Kalin Stefanov
kalin.stefanov@monash.edu
Monash University
Melbourne, Victoria, Australia

Maja J Matarić
mataric@usc.edu
University of Southern California
Los Angeles, California, USA

ABSTRACT

In this work we address the Next Speaker Prediction sub challenge of the ACM '21 MultiMediate Grand Challenge. This challenge poses the problem of turn taking prediction in physically situated multiparty interaction. Solving this problem is essential for enabling fluent real-time multiparty human-machine interaction. This problem is made more difficult by the need for a robust solution that can perform effectively across a wide variety of settings and contexts. Prior work has shown that current state-of-the-art methods rely on machine learning approaches that do not generalize well to new settings and feature distributions. To address this problem, we propose *the use of group-level focus of visual attention as additional information.* We show that a simple combination of group-level focus of visual attention features and publicly available audio-video synchronizer models is competitive with state-of-the-art methods fine-tuned for the challenge dataset.

CCS CONCEPTS

• **Computing methodologies** → Neural networks; Supervised learning; Unsupervised learning; • **Human-centered computing** → Collaborative interaction; Laboratory experiments.

KEYWORDS

turn taking prediction; focus of visual attention; neural networks

ACM Reference Format:
Chris Birmingham, Kalin Stefanov, and Maja J Matarić. 2021. Group-Level Focus of Visual Attention for Improved Next Speaker Prediction. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3474085.3479213

1 INTRODUCTION

Understanding and predicting changes in participant roles on the conversational floor (*i.e.*, speaker, addressee, bystander), known as *footing* [5, 6], is a prerequisite for most natural and effective human-machine interaction. To fluently participate in a situated, multiparty conversation, a system must understand when the speaker will take and relinquish their turn, known as *next speaker prediction* and *turn taking prediction.*

Turn taking prediction has been formulated in many ways, and requires estimating which members of the group, if any, will be speaking at a future point in time [18], in order to interact with an individual or a group without excessively long pauses and without interrupting/cutting into other group member's turns.

While most people can perform turn taking prediction in complex multiparty settings with relative ease, computers struggle to do so accurately even in simple settings. The task is inherently multimodal, requiring accurate synthesis of and reasoning about visual, auditory, and linguistic information. In physically situated interactions this challenge is amplified by perceptual limitations, *e.g.*, monocular cameras and far-field microphones. Additionally, natural conversations can be noisy, with overlaps, cut-ins, and backchannels that can blur the distinction between the active speaker and other group members.

Given these challenges, it is helpful to incorporate information beyond the candidate speaker's own visual and auditory data. Such information can include objects of interest in the environment and, in the case of multiparty interactions, it can include information from other group members, such as their focus of visual attention.

In this work, we utilize group members' focus of visual attention along with publicly available audio-video speech synchronizer models to demonstrate a competitive, dataset agnostic method for turn taking prediction on the '21 MultiMediate Challenge [9–11]. Our prior work (under review) has shown that this simple combination can achieve state-of-the-art (SOTA) results when working *across* similar datasets. Here, we evaluate our method against prior SOTA models on the challenge's validation set and against the newly developed models submitted to the challenge's hidden test set, showing competitive performance in both cases.

2 BACKGROUND

Turn taking in spoken dialogue systems is the coordination of system speech with the person or persons with whom the system is speaking. Turn taking modeling can be formulated as the process of estimating whether or not a given group member will be speaking at a future point in time, also known as next speaker prediction. Turn taking is commonly defined as having four cases: *hold*, when a person is talking and continues to do so; *yield*, when a person is talking and is about to stop; *take*, when a person is not talking but will start to talk; and *listen*, when a person is not talking and will continue to not talk. Prior turn taking decision modeling typically addresses the yielding and holding cases. These models estimate whether a person is yielding or holding when a short pause occurs during their speech. Continuous turn taking makes a prediction about future speech in order to address all four cases. Skantze [18]

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3479213

provides a comprehensive review of the history and state-of-the-art in turn taking methods.

The first model of continuous turn taking was introduced in Skantze [19], built to support an autonomous multiparty robot in an interactive museum. The approach predicted a vector of speech activity 60 frames or 3 seconds into the future using a multimodal Long Short-Term Memory network (LSTM). The model incorporated acoustic and part of speech features, but it was found that it performed almost as well with only the acoustic features. That seminal approach has been extended in Roddy *et al.* [15, 16] through the use of a multiscale Recurrent Neural Network (RNN) architecture, in which different modalities were modelled in individual sub-network LSTMs that operated at their own independent timescales, with a separate LSTM that fused the modalities to form predictions at a regular rate. Ward *et al.* [21] also extended the work in [17] through the use of an improved multilayer LSTM utilizing parametric rectified linear units and by testing the model across multiple languages and conversational genres. Masumura *et al.* [8] developed a novel model for end of turn detection utilizing a cross-modal representation trained with a punctuated text dataset.

In face-to-face interactions, eye-gaze is one of the best studied and strongest visual cues for coordinating turn taking and managing attention by dialogue partners [13]. Within the fields of Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI), eye-gaze detection has already been utilized for addressee and backchannel detection (*e.g.*, [1]). Eye-gaze control has been successfully deployed for turn signaling and control by the agent. Despite the theoretical importance of eye-gaze in group interactions and the use of eye-gaze in HRI and HCI, eye-gaze remains an underutilized feature for active speaker detection and turn taking prediction, inspiring the work in this paper.

3 METHODS

The goal of the methods described in this section is to detect the speaking state (*i.e.*, speaking or not speaking) of all visible faces in a physically situated multiparty interaction at some point in the near future (*i.e.*, turn taking prediction).

3.1 Task Definitions

Given a fixed context window of sensor data pertaining to a number of potential speakers, turn taking prediction consists of the task of determining which speakers will be active at some fixed point in the future, by using information from the current point in time. The MultiMediate Challenge formulates the problem of predicting the state of each potential speaker (*i.e.*, speaking or not speaking) as a binary classification task, with independent labels for each potential speaker.

3.2 Models

The backbone model used in this work is the SOTA audio-video *synchronizer* termed PerfectMatch [4]. Following the state-of-the-art in active speaker detection [2], the synchronizers are further turned into *predictors* by adding a Temporal Convolution Network (TCN) and a Bidirectional Long Short-Term Memory (BLSTM). Here we propose models that augment the synchronizers and predictors with information about the group member's focus of visual attention.

3.2.1 Features. Our work combines two sets of features: PerfectMatch and VisualAttention features. Next, we describe how those features are computed.

PerfectMatch – Chung and Zisserman [3] trained a Convolutional Neural Network (*i.e.*, SyncNet) to produce audio and video embeddings for the purpose of synchronizing audio and video tracks of individuals talking. The network consisted of 6 convolutional layers followed by 2 fully connected layers for separate audio and video. The model took as input 5 video frames and the corresponding audio samples. This model has been shown to be effective for active speaker detection by comparing the magnitude of the difference between the final audio and video features and smoothing with a median filter. Chung et al. [4] re-trained the original SyncNet model for the purpose of synchronizing audio and video tracks of individuals talking. The new model (*i.e.*, PerfectMatch) was shown to outperform the original SyncNet model. The output of the final convolutional layer from the PerfectMatch model that had been pre-trained on VoxCeleb [12] was directly used to produce $512D$ audio and $512D$ video features for each frame.

VisualAttention – The visual attention features are inspired by Stefanov *et al.* [20] and include *continuous* and *binary* representations of whether or not a group member is looking at the candidate speaker. The binary representation considers the group member to be looking at the candidate speaker if the group member's head pose is pointed closer to the candidate speaker than the head pose of any of the other group members. Looking at the candidate speaker is represented by a 1, otherwise a 0. For the continuous representation, the angle between the head pose of the group member and the vector between the head of the group member to the head of the candidate speaker is then used as a measure for how far away they are looking from the candidate speaker. This angle is normalized using,

$$\mathrm{f} = 1 - \theta/q \tag{1}$$

where f is the feature from that group member to the candidate speaker and θ is the angle between the group member's head pose and the candidate speaker-group member vector and q is a normalization factor based on the field of view of the group member.

In both the continuous and binary cases we created the mean VisualAttention feature for the candidate speaker by averaging the features produced by the other group members with respect to that candidate speaker. This produced a single value for the average of the continuous and binary features. The closer this value was to 1, the more likely the individual was to be the focus of attention of the group; conversely, the closer the value was to 0, the more likely it was that the individual was not the focus.

3.2.2 Model Architecture. The **synchronizer** model architecture is described in [3]. This includes a window of 5 frames of video and the corresponding audio which produces $512D$ audio and video embeddings. These embeddings are cross-correlated to find the minimum difference (the matching synchronization), then the difference between the embeddings is median filtered and normalized for each video. This produces a single value between 0 and 1 which is used as the confidence that the candidate speaker is speaking. The synchronizer utilizes the entire video as context for determining the median, and thus is not suited for real-time use.

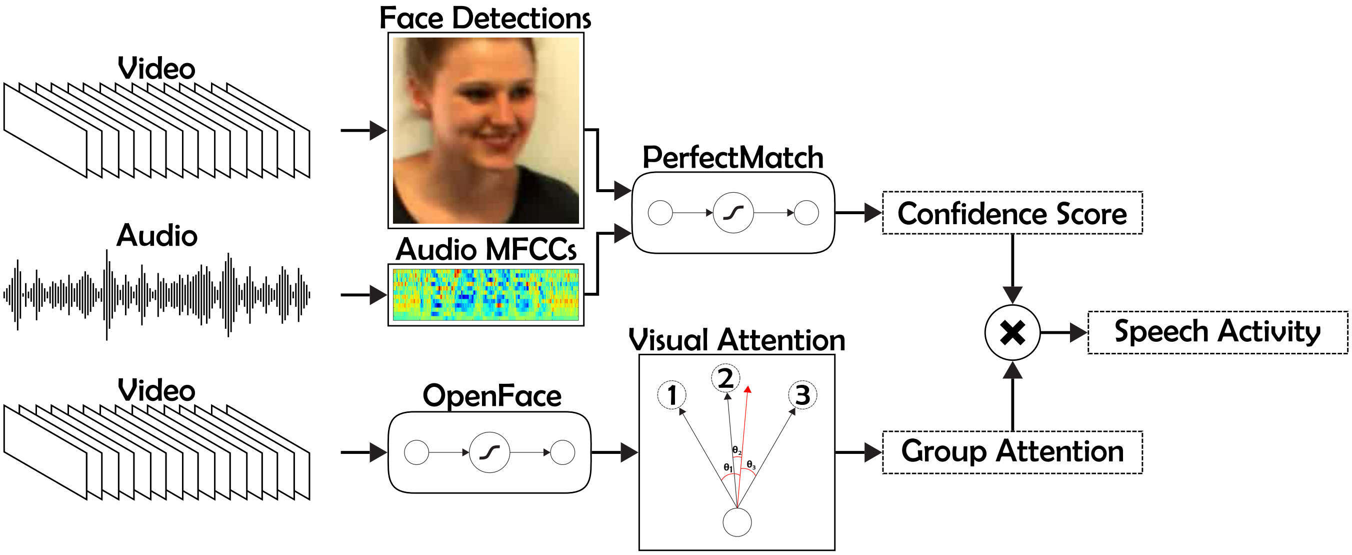

Figure 1: The augmented VisualAttention synchronizer consists of a speech activity confidence score generated by the PerfectMatch model for the candidate speaker and the group attention score for each group member except the candidate speaker from the latest available frame. The figure shows the method for generating continuous rather than binary VisualAttention features. Group attention and speech activity are combined to produce a label estimate for future speech for the candidate speaker. This is repeated for all group members in the interaction.

The **predictor** model architectures are described in [2]. A window of 5 of the $512D$ audio and video embeddings from each frame are used as input to a 2 layer time series model before being combined into a fully connected layer. The fully connected layer is then connected to a softmax layer to produce the final output probabilities. As in [2], we used both a Temporal Convolution Network (TCN) and a Bidirectional Long Short-Term Memory (BLSTM) as the time series models. These models do not require the full video context and can be used in real-time applications.

The **VisualAttention augmented models** combined the output of the synchronizer and predictors described above with the mean VisualAttention feature described in 3.2.1. To augment the predictions of the synchronizer and predictor models, which were also between 0 and 1, we multiplied the confidence value with the VisualAttention feature, which was shifted from $[0, 1]$ to $[0.5, 1.5]$ to allow the feature to adjust the model confidence. The full architecture for the VisualAttention augmented synchronizer can be seen in Figure 1.

4 EXPERIMENTS

This section describes the dataset used to train and evaluate the turn taking predictors and the general experiment setup.

4.1 Datasets

The '21 MultiMediate Challenge used the published MPIIGroupInteraction dataset [10]. The dataset consists of 22 German language conversations between three to four people, each with an approximate length of 20 minutes. Participants in each conversation were instructed to discuss a controversial topic and were recorded by 8 frame-synchronised video cameras and 4 microphones. The challenge provides the recording from all cameras (one from behind each participant) and one of the microphones for each session. Every frame of each recording is labelled with a binary representation of who is speaking.

4.2 Experimental Setup

4.2.1 Model Implementation. The **synchronizer** was implemented with a SOTA pre-trained model. The model was pre-trained on VoxCeleb [12], which may contain significantly different data distributions than the challenge datasets.

The **predictors** were implemented and trained as described in [2]. The input features from the synchronizer were held constant and the TCN and BLSTM models were trained on the MPIIGroupInteraction dataset for each experiment. The training was implemented in PyTorch [14]. The models were trained for 25 epochs, with a batch size of 64. The Adam [7] optimizer was used with the default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) and a fixed learning rate of 0.001. The loss was calculated with the cross entropy loss function.

The **VisualAttention augmented models** required additional processing to generate the features for this dataset. In the MPIIGroupInteraction dataset, the position of the cameras was not publicly available, so the positions of the group members relative to each other was estimated. The cameras were not centered exactly for each group member but were consistent across all sessions, so the estimates were created using the distributions in the training set of head poses for each group member when the other group members were speaking. For the *binary representation*, this estimation was accomplished by dividing the head yaw rotation of all participants in a given seat into three equal quantiles, where one quantile was assigned to the person to their left, one across, and one to their right. When the direction a participant was facing was within a given quantile, the participant was considered to be looking at the seat in that quantile and and not at the other two. Creating a *continuous representation* requires a more exact method for estimating the location of a given participant relative to each of the others. This was accomplished by taking the mean of the distribution of head yaw rotations when each of the other group members was speaking and using it as a proxy for the location of that group member.

4.2.2 Model Evaluation. Each of the models was evaluated on the validation set provided by the '21 MultiMediate Challenge organizers. The VisualAttention Augmented Synchronizer was evaluated on the hidden test set, as the intention of this work is to evaluate the model that was found to generalize best to new datasets in our prior work.

We follow the reporting requirements for the '21 MultiMediate Challenge in reporting the unweighted average recall scores (UAR). Here, metrics are calculated for each candidate speaker and their unweighted mean is found, where the unweighted recall score for each candidate speaker is defined as:

$$R = tp/(tp + fn) \tag{2}$$

where R is the recall, tp is the number of true positives, and fn is the number of false negatives.

5 RESULTS

This section reports the results of the experiments described in Section 4.

Challenge Validation Set		
Model	**Binary Aug.**	**Continuous Aug.**
BLSTM Predictor	0.719	**0.747**
TCN Predictor	0.572	0.721
Synchronizer	0.692	0.715

Table 1: UAR performance of models augmented by Binary and Continuous VisualAttention features on the validation set provided by the MultiMediate competition.

The results of the experiment on the '21 MultiMediate validation set are reported in Table 1. The best performing model was the BLSTM. The synchronizer augmented with continuous VisualAttention features performed only slightly worse than the SOTA BLSTM and TCN predictors augmented with VisualAttention, by −0.032 and −0.006, respectively, both of which have been explicitly fine-tuned on the '21 MultiMediate training set.

For the competition test set we submitted the synchronizer augmented with continuous and binary VisualAttention features. The binary features performed worse than the continuous, with a score of 0.628 and 0.632, respectively. The provided baseline was 0.51. Both submissions achieved SOTA performance, outperforming all other competitors.

Challenge Test Set	
Authors	**Binary**
Ours (Cont. Aug. Synchronizer)	**0.632**
Ours (Binary Aug. Synchronizer)	0.628
HNU VPAI	0.57
Jiangeng	0.53
MM Baseline	0.51

Table 2: UAR performance of competitor models and our synchronizer model augmented by Binary and Continuous VisualAttention features on the held-out test set provided by the MultiMediate competition.

6 CONCLUSION

In this work we show how simple focus of visual attention features can improve the performance of general purpose synchronizers to be competitive with SOTA methods on the '21 MultiMediate Next Speaker Prediction Challenge. While the calculation of VisualAttention features requires an understanding of the physical relationship between individuals in a scene, it does not require the retraining or fine-tuning of another model. These models do not outperform SOTA models that have been fine-tuned for a specific setting, as can be seen in the validation results, but our prior work (under review) has shown they can outperform when transferred to new scenes without an existing dataset and where fine-tuning is not possible.

It is important to note that the PerfectMatch synthesizer models used for this competition were not trained for next speaker prediction but for active speaker detection. The augmentation with visual focus of attention features may help for the turn taking and turn yielding cases, but this requires further investigation. Additionally, because the '21 MultiMediate Challenge utilized multiple cameras without providing their relative location to one another, in this work we utilized crude, approximate measures of visual attention. Future work will attempt to utilize more fine grained measures of visual attention. Additionally, the presented work utilized the simple method of multiplying for feature augmentation. In future work we will develop better ways for combining these features.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation National Robotics Initiative, IIS-1925083.

REFERENCES

[1] H. Admoni and B. Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

[2] J. S. Chung. 2019. Naver at ActivityNet Challenge 2019 – Task B Active Speaker Detection (AVA). *arXiv preprint arXiv:1906.10555* (2019).

[3] J. S. Chung and A. Zisserman. 2016. Out of Time: Automated Lip Sync in the Wild. In *Proceedings of the Workshop on Multi-view Lip-reading*.

[4] S.-W. Chung, J. S. Chung, and H.-G. Kang. 2019. PerfectMatch: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 3965–3969.

[5] E. Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience.* Harvard University Press.

[6] E. Goffman. 1981. *Forms of Talk.* University of Pennsylvania Press.

[7] D. P. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. *Computing Research Repository* abs/1412.6980 (2014).

[8] R. Masumura, M. Ihori, T. Tanaka, A. Ando, R. Ishii, T. Oba, and R. Higashinaka. 2019. Improving Speech-Based End-of-Turn Detection Via Cross-Modal Representation Learning with Punctuated Text Data. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. 1062–1069.

[9] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. *MultiMediate : Multi-modal Group Behaviour Analysis for Artificial Mediation.* Technical Report. 1–6 pages.

[10] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behavior. In *Proc. ACM International Conference on Intelligent User Interfaces (IUI)*. 153–164. https://doi.org/10.1145/3172944.3172969

[11] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour. In *Proc. ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 31:1–31:10. https://doi.org/10.1145/3204493.3204549

[12] A. Nagrani, J. S. Chung, and A. Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

[13] C. Oertel, M. Wlodarczak, J. Edlund, P. Wagner, and J. Gustafson. 2013. Gaze Patterns in Turn-Taking. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

[14] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic Differentiation in PyTorch. In *Proceedings of the NeurIPS Autodiff Workshop*.

[15] M. Roddy, G. Skantze, and N. Harte. 2018. Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. In *Proceedings of the Annual Conference of the International Speech Communication Association*. 586–590.

[16] M. Roddy, G. Skantze, and N. Harte. 2018. Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. In *Proceedings of the ACM International Conference on Multimodal Interaction*. 186–190.

[17] G. Skantze. 2017. Towards a General, Continuous Model of Turn-Taking in Spoken Dialogue Using LSTM Recurrent Neural Networks. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 220–230.

[18] G. Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 101–178.

[19] G. Skantze, M. Johansson, and J. Beskow. 2015. Exploring Turn-Taking Cues in Multi-Party Human-Robot Discussions About Objects. In *Proceedings of the ACM International Conference on Multimodal Interaction*. 67–74.

[20] K. Stefanov, G. Salvi, D. Kontogiorgos, H. Kjellström, and J. Beskow. 2019. Modeling of Human Visual Attention in Multiparty Open-World Dialogues. *ACM Transactions on Human-Robot Interaction* 8, 2 (2019), 21.

[21] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes. 2019. Turn-Taking Predictions Across Languages and Genres Using an LSTM Recurrent Neural Network. In *Proceedings of the IEEE Spoken Language Technology Workshop*. 831–837.