Who Speaks Next? Turn Change and Next Speaker Prediction in Multimodal Multiparty Interaction

Usman Malik
INSA Rouen, LITIS
Normandie University
Rouen, France
usman.malik@insa-rouen.fr

Julien Saunier
INSA Rouen, LITIS
Normandie University
Rouen, France
julien.saunier@insa-rouen.fr

Kotaro Funakoshi
FIRST, IIR
Tokyo Institute of Technology
Yokohama, Japan
funakoshi@lr.pi.titech.ac.jp

Alexandre Pauchet
INSA Rouen, LITIS
Normandie University
Rouen, France
alexandre.pauchet@insa-rouen.fr

***Abstract*—Turn change prediction and next speaker prediction are two important tasks in multimodal, multiparty human-agent interaction. Predicting a change of dialogue turn and the most probable next speaker can help an agent to decide whether he should contribute to the discussion or wait for someone else to speak. In this research, we propose a machine learning-based approach for both turn change and next speaker prediction. Individual as well as combined models are explored to tackle these tasks. Results show that the proposed models outperform baselines. An ablation study is also performed to measure the importance of different features.**

***Index Terms*—Human-Agent Interaction, Multimodal Interaction, Multiparty Interaction, Machine Learning.**

I. Introduction

In a dialogue, an utterance may induce a response from one of the listeners, depending on the context. For instance, a speaker can ask questions to one or more people that cause a particular participant to reply. This change of speaker is referred to as *turn change* and the process of turn distribution among the meeting participants is called *turn management* [1].

In human-human interactions, turn management is mostly implicit. Interaction participants are expected to know when to speak, rather than explicitly being told to talk. Furthermore, at each time the speaker can continue or can be interrupted by any of the listeners, hence there is no 'correct' next speaker and the 'real' next speaker is only known when anyone takes the turn. Speaker and listeners exploit language as well as various co-verbal and non-verbal signals such as pitch, gaze, head and hand gestures, to implicitly negotiate turn change [2]–[4]. Turn management is therefore a multimodal process.

User experience in multimodal human-agent interaction can be improved if turn changes and next speakers are predicted efficiently by the agents. Turn change and next speaker prediction not only enable to understand when to contribute to an interaction, but also help to detect who should speak next, and thus generate an appropriate behaviour.

In this article, we propose a machine learning based approach for turn change and next speaker prediction in multimodal, multiparty interaction. The proposed models are compared against baselines on two different datasets, resulting in improvement of predictions. In addition, an ablation study is performed to investigate the importance of some of the features for turn change and next speaker prediction. The remainder of this article is divided into 7 sections. Section II presents some existing works. The process of feature selection is explained in section III. Section IV formalizes the problem and describes the experimental methodology. Section V details experiments and results, and section VI concludes the article.

This work was supported by the INCA project, cofunded by the European Union with the European Regional Development Fund and by the Regional Council of Normandie. The work was partially funded by Kyoto University, Japan. The MPR 2012 dataset has been developed in collaboration between Honda Research Institute, Japan Co., Ltd and Kyoto University, Japan.

II. Related Work

This section reviews existing works, firstly with turn change prediction processed independently, and secondly when combined with next speaker prediction.

A. Models for Turn Change Prediction

One of the earliest turn management model is proposed by Sacks *et al.* [5]. The authors report that conversations proceed smoothly with only one people speaking at a time, and that sub-dialogues with multiple participants speaking simultaneously are short. Transition Relevance Places (TRPs) mark the end of a turn and initiate a new turn. Relying on TRPs, Sacks *et al.* propose rules for turn management [5]: 1) If the current speaker (S) selects the next speaker (N) in the turn, S is expected to stop speaking, and N to speak next; 2) If S does not select the next speaker, then any other participant may self-select and whoever speaks first gets the turn; 3) If no speaker self-selects, S may continue. Though these rules are sufficiently generic to apply in various situations, they are not specific enough to explain which interaction features signal turn change and which ones indicate the next speaker in case of turn change.

Guntakandla and Nielsen employ a J48 decision tree for turn prediction using n-grams of current and previous Dialogue Acts[1] (DA) and current and previous speakers as features [2]. The model is trained on Switchboard dataset [8]. They report an overall accuracy of 62.70% for turn change prediction.

Meshorer and Heeman propose a model that exploits random forests to predict turn changes [3]. The model is also

[1]A dialogue act is the meaning of an utterance at the level of illocutionary force [6] or as the function of a user's utterance [7].

trained on Switchboard and uses current and previous DA, the *relative turn length* [2] and the *relative turn floor control* [3] to predict turn change. They report an accuracy of 76.05% and a F1 score of 0.74 for turn change prediction.

Aldeneh *et al.* consider turn change as a sequence problem [9]. They propose a turn change prediction model based on LSTM applied to speech features such as loudness, intensity or zero-crossing rate. The model is trained on Switchboard and they report an F1 score of 0.65 for turn change prediction. De Kok and Heylen propose a machine learning based model for end-of-turn prediction [4], trained on AMI [10]. The proposed model exploits DA, focus of attention, head gesture and prosody as features to train CRF and HMM models. They report a F1 score of 0.61 for end-of-turn prediction.

B. Models Combining Turn Change and Next Speaker

Several researchers have proposed combined models for turn change and next speaker prediction. To this end, Petukhova and Bunt studied the importance of various multimodal signals such as gaze directions, verbal signals, lip movements and posture shift for next speaker prediction [11]. The proposed model looks for correlations between various multimodal features and turn types such as turn taking, turn grabbing and turn accepting in AMI dataset.

Kawahara *et al.* also propose a machine learning based turn change and next speaker prediction model that relies on a combination of gaze, prosody and head movements [12]. The dataset used to train their models consists of 3 participants. They report an accuracy of 70.60% for turn change and 69.06% for next speaker prediction using SVM.

Ishii *et al.* propose a probabilistic turn change and next speaker prediction model funded on participants gaze transition patterns near the end of utterance in multiparty interaction [13]. A total of 12 gaze transition patterns are used for predictions. They achieve a F1 score 0.76 for turn change prediction and an accuracy of 59.20% for next speaker prediction. Ishii *et al.* then exploit human gaze and respiratory behaviour for turn change and prediction of next speaker in multiparty interaction [14]. The tests are performed on a custom dataset of 4 participants. Sequential minimal optimization, which is a variation of SVM, is used to train the models. Results show that the model based on late fusion of eye gaze and respiratory behaviour yields a F1 score of 0.75 for turn change prediction and a F1 score of 0.52 for next speaker prediction. Ishii *et al.* also propose a two-step machine learning model that predicts initially whether or not a turn change occurs and then who is the next speaker [15]. The model exploits head movements of the speaker and listeners (amplitude and frequency) near the end of the utterance. The model is trained via SVM on a custom dataset of 4 participants. They achieve an accuracy of 75.00% for turn change prediction and an accuracy of 55.20% for next speaker prediction. Ishii *et al.* also investigate the role of mouth-opening transition pattern to predict the time interval between the current utterance and the next utterance, and the next speaker in multiparty interaction [16]. A SVM is trained for turn change and next speaker prediction on a custom dataset. They report a F1 score of 0.80 for turn change prediction and a F1 score of 0.47 for next speaker prediction.

[2]Two versions: (i) duration of a turn divided by the average speakers turn duration, and (ii) number of words of a turn divided by the average number of words of the speaker's turns.

[3]Two versions: (i) total duration of the speaker's turns divided by the duration of the whole conversation, and (ii) total number of words the speaker's turns divided by the total number of words in the whole conversation.

C. Summary and Discussion

A summary of the existing works is presented in Table I. Most of the works consider turn change as a binary classification problem at the end of each utterance. Both rule based and machine learning based approaches are used for turn change prediction. The most commonly features exploited for turn change prediction are DA, prosody, gaze information, head movements and speaker information.

Concerning next speaker prediction, existing works rely on machine learning. The accuracies reached for next speaker prediction are significantly lower than those achieved for turn change prediction. The reason can be that next speaker prediction is a multiclass classification problem which increases the chance of miss-classification. Another reason is the uncertainty of the task since there can be multiple potential next speakers after a certain utterance since any of a meeting participants can speak at any time.

III. Feature Selection

In this article, features are selected according to their relevance for turn change and next speaker, based on the literature review and an analysis of two datasets: AMI [10] and MPR [17].

A. Features

Speaker Role: Speaker role refers to the identity of the current speaker. In AMI, participants are identified by their role *i.e.* Project Manager, Marketing Expert, User Interface Designer, and Industrial designer. In MPR, participants are identified by IDs: A, B, C (human participants) and NAO (a robot). Work from [2] highlights the importance of including current speaker for turn change prediction.

Dialogue Act: DA is an important feature for turn change and next speaker prediction. For instance, a DA involving a question often prompts a turn change and leads (one of) the addressee(s) to answer. The importance of DA for the two tasks is highlighted in several existing works [2]–[4], [11].

Pause Duration: Psycho-linguistic evidence shows a strong correlation between pause duration and turn change [18], [19]. A quantitative analysis [18] of speaker changes reports that 91% of the speaker changes occur with a pause between two utterances whereas only 8% of the turn changes occur with no pause, and 1% of the utterances overlap another one.

Start and End Time of Utterance: The start and end time of an utterance enable to calculate utterance duration which can play an important role in human-agent interaction [3]. The analysis of the AMI and MPR datasets show that the utterances

TABLE I
SUMMARY OF RELATED WORKS FOR TURN CHANGE AND NEXT SPEAKER PREDICTION

Reference	Approach	Dataset	Salient Features	Turn Change	Next Speaker
Guntakandla and Nielsen [2]	Decision Tree (J48)	Switchboard	Current and previous DA, current and previous speakers	62.00% Accuracy	NA
Meshorer and Heeman [3]	Random Forest	Switchboard	Current and previous DA, relative turn length, relative floor control	76.05% Accuracy	NA
Aldeneh *et al.* [9]	LSTM	Switchboard	Speech features e.g loudness, intensity	0.65 F1 Score	NA
Petukhova and Bunt [11]	Correlation	AMI	Gaze, head movements, hand gestures	feature correlations	feature correlations
De kok and Heylen [4]	CRF & HMM	AMI	DA, focus, head movements, prosody	0.65 F1 Score	NA
Kawarah *et al.* [12]	SVM	Custom Dataset	Gaze, prosody and head movements	70.00% Accuracy	69.06% Accuracy
Ishii *et al.* [13]	Probabilistic	Custom Dataset	Gaze transition patterns	0.76 F1 Score	59.2% Accuracy
Ishii *et al.* [14]	SVM	Custom Dataset	Gaze and respiratory behaviour	0.75 F1 Score	0.52 F1 Score
Ishii *et al.* [15]	SVM	Custom Dataset	Head movements	75.00% Accuracy	55.20 % Accuracy
Ishii *et al.* [16]	SVM	Custom Dataset	Mouth opening transition patterns	0.80 F1 Score	0.47 F1 Score

at the beginning of a conversation are less likely to have turn changes compared to utterances at later stages.

Focus of Attention: The importance of gaze and focus of attention as a marker for turn change and next speaker prediction has been largely investigated [4], [11]–[14]. The results from these research works show that focus of attention is fundamental for turn change and next speaker prediction.

Addressee Role: Though there is no evidence from existing works that addressee role is important for turn change and next speaker prediction, we propose to add it in the feature set. The rationale is that if a speaker asks a question to someone, that addressee is more likely to take the turn and respond to the speaker. The analysis of AMI also reveals that utterances addressed to individuals rather than groups are more likely to cause turn change.

B. Discussion

Despite psycholinguistic evidences, to the best of our knowledge none of the existing works has exploited pause duration as a feature for turn change and next speaker prediction in multiparty interaction. A reason can be that pause duration is a dynamic attribute: pause duration between two utterances cannot be calculated before the start of the next utterance. An implementation solution is to regularly evaluate pause duration and trigger the turn change prediction.

Addressee role is another feature that, to the best of our knowledge, has not yet been exploited for turn change and next speaker prediction. This feature is selected based on the analysis of several datasets that reveal that the utterances addressed to individuals are more likely to evoke a response causing turn change than those addressed to a group. Furthermore, the addressee of the current utterance is often the next speaker.

IV. Problem Formalization and Methodology

Turn management is divided into two sub-tasks: turn change and next speaker prediction. Two approaches are proposed to solve these tasks: (i) two independent models; (ii) a connected model where turn change is predicted in the first step and then next speaker is predicted by including turn change as a feature.

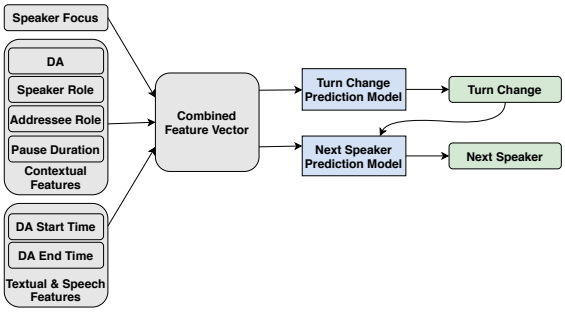

Fig. 1. ML based independent and combined turn change and next speaker prediction models. Turn change is used as input feature for next speaker prediction in combined model.

A. Turn Change Prediction and Next Speaker Prediction Considered as Independent Problems

Given a set of features, the first task consists in predicting whether the speaker of an utterance differs from the speaker of the next utterance. Turn change prediction is a binary classification problem since there are only two possible outputs.

The second task performed is to predict who is the speaker of the next utterance. Next speaker prediction is a multiclass classification problem since there are more than two possible outputs. For instance in AMI, the next speaker can be any of the four participants.

For the sake of simplification, in this article it is assumed that an utterance consists of a single DA and thus both turn change and next speaker are predicted after every DA.

Figure 1 shows the experimental methodologies considering turn change and next speaker prediction as independent as well as combined models. Input features, divided into speaker focus, contextual features, and textual and speech features, are fused together to form a combined feature vector that is used to train both models. The turn change prediction model outputs a binary value (a turn change occurs / no turn change). The next speaker prediction model predicts the speaker for the next utterance among the meeting participants.

B. Combining Turn Change and Next Speaker Prediction

Turn change and next speaker prediction are two related tasks as next speaker prediction depends on turn change: a turn

change signals that the speaker of the next utterance cannot be the current speaker and someone different from the current speaker has to speak next. Thus, predicted turn change can also be used as an additional feature.

In this case, the next speaker is predicted in two steps: 1) the input features are used to predict turn change, and 2) the predicted turn change value is added to the feature vector (the arrow from turn change to next speaker prediction model in Figure 1). The real turn change values are used to train the model whereas at run time the predicted turn change value is exploited as additional feature.

To evaluate our turn change prediction models, two baselines are selected: (i) Meshorer and Heeman [3], and (ii) majority class for turn change. The first baseline is selected because (a) it returns the highest performance on Switchboard and (b) its features are available in most of the existing datasets and therefore results can be reproduced. To evaluate our next speaker prediction model, the majority class for next speaker prediction is chosen as baseline. Indeed, the selected features are not available in the datasets exploited by existing research works, nor is it possible to reproduce their results on the commonly used datasets (i.e. Switchboard and AMI) as they are based on unavailable features. Hence, unfortunately, it is not possible to compare our proposed model with any of the existing models as baseline for next speaker prediction.

V. Experiments and Results

This section describes the datasets, the procedure followed to perform the experiments and the results obtained.

A. Datasets

The datasets used to train and test the turn change and next speaker prediction models are AMI [20] and MPR [17]: they are the only corpora that contain annotated data for all the features selected in Section III.

AMI is a multimodal interaction corpus of 100 hours of meeting recordings. Each meeting involves 4 participants, and is either task-oriented or open discussion. These participants are Project Manager (PM), Industrial Designer (ID), Marketing Executive (ME), and User Interface Expert (UI). The AMI dataset uses custom taxonomy for DA annotation.

MPR dataset contains 30 trios of Japanese individuals that participate in two 25-minute interactive sessions with a robot in which they repeatedly engage in a conversational game. The meeting participants are labelled as A, B and C, while the robot is labelled as NAO. The MPR dataset uses DIT++ [21] taxonomy for DA annotation.

B. Procedure

Three sets of experiments are performed : (i) experiments performed to evaluate performance of turn change and next speaker prediction models individually, as mentioned in Section IV-A, (ii) experiments performed for next speaker prediction using predicted turn change explained in Section IV-B, and (iii) experiments performed for ablation study in order to evaluate the importance of some of the features.

1) Experiments for Individual Turn Change and Next Speaker Prediction Models: Two separate sets of experiments are performed: one for turn change prediction model, and the other for next speaker prediction model. A conventional machine learning pipeline is followed to carry out the experiments. The categorical features in the feature set are one-hot encoded to convert them into a numerical form. The feature set is normalized using standard scaling. The data set is divided randomly into 80-20% training and test sets, respectively.

Six of the most classic machine learning classifiers have been trained and tested: XGboost (XGB) [22], Multilayer Perceptron (MLP) [23], Random Forest (RF) [24], Logistic Regression (LR) [25], Support Vector Machines (SVM) [26] and K-Nearest Neighbours (KNN) [27]. For all the classifiers, default parameters as specified in Python's Sklearn library [28] are used. Finally, accuracy and F1 measure have been employed to evaluate performances. Accuracy is used to compare the results with baselines and F1 measure is considered since the class distribution is irregular in both datasets [28].

2) Experiments for Model Combining Turn Change and Next Speaker Prediction: Concerning the experiments performed to evaluate the model combining turn change and next speaker prediction, the predicted turn change is included in the feature set in order to predict the next speaker. The values for turn change prediction are obtained with the model that yields best result for turn change prediction. The machine learning pipeline, the algorithms and the evaluation metrics are the same as those of Section V-B1.

3) Ablation study: In the proposed models, two new features (*i.e.* pause duration and addressee role) are added. Experiments are performed with and without these features to evaluate if these features significantly improve turn change and next speaker predictions.

Since the importance of different features is studied individually on turn change and next speaker prediction models, two sets of experiments are performed: (i) Turn change prediction, (ii) and Next speaker prediction.

The naming convention for the experiments follows the pattern *dataset-task-experiment*, where: dataset refers to ami or mpr; task refers to tc (for turn change prediction) or ns (for next speaker prediction); experiment refers to one of the 4 different experiment types *i.e.* -ar-pd, +ar-pd, -ar+pd, and +ar+pd. In experiments -ar-pd, the models are trained using the feature set presented in Section III without addressee role (-ar) and pause duration (-pd). In experiments +ar-pd and -ar+pd, each feature is added independently, while in +ar+pd, the whole feature set is used. To find statistical significance, t-tests are performed between pairs of experiments (-ar-pd, +ar-pd), (-ar-pd, -ar+pd), and (-ar-pd, +ar+pd).

The machine learning pipeline, the algorithms and the evaluation metrics for the ablation experiments are the same as the ones used for the other experiments.

C. Results

1) Results for Individual Turn Change and Next Speaker Prediction Models: Table II show that for both the MPR and

TABLE II
RESULTS FOR TURN CHANGE PREDICTION FOR MPR AND AMI DATASETS (ACCURACIES IN %, F1 VALUES IN BRACKETS).

Algorithm	MPR	AMI
XGB	**83.02 (0.82)**	**87.59 (0.88)**
RF	82.32 (0.82)	86.42 (0.86)
MLP	80.43 (0.80)	64.36 (0.64)
SVM	76.46 (0.71)	66.55 (0.66)
KNN	76.72 (0.75)	62.95 (0.63)
LR	75.85 (0.71)	65.83 (0.66)
Baseline 1 [3]	**77.72 (0.75)**	**60.05 (0.59)**
Baseline 2 (Majority Class)	**75.29**	**56.99**

TABLE III
RESULTS FOR NEXT SPEAKER PREDICTION FOR MPR AND AMI DATASETS (ACCURACIES IN %, WEIGHTED F1 VALUES IN BRACKETS)

Algorithm	MPR	AMI
XGB	**65.64 (0.65)**	**64.04 (0.64)**
RF	64.02 (0.64)	65.54 (0.66)
MLP	61.95 (0.61)	45.76 (0.49)
SVM	57.43 (0.57)	51.41 (0.51)
KNN	56.33 (0.56)	48.36 (0.48)
LR	56.96 (0.58)	50.88 (0.51)
Baseline (Majority Class)	**36.62**	**32.81**

AMI datasets, the proposed turn change model outperforms both of the baselines. A maximum accuracy of 87.59% is reached on AMI using XGB algorithm (baseline 1: 60.05%; baseline 2: 56.99%). On MPR, the XGB algorithm achieves a maximum accuracy of 83.02% which also outperforms baselines 1 (77.72%) and 2 (75.29%).

Table III contains the results to evaluate the performance of the next speaker prediction model. On AMI, the results show that the proposed model achieves a best case accuracy of 64.04% via the XGB algorithm, which is better than the baseline accuracy of 32.81%. Similarly on MPR, a maximum accuracy of 65.64% is achieved via the XGB algorithm, which is better than the baseline accuracy of 36.62%.

2) Results for Model Combining Turn Change and Next Speaker Prediction: Table IV depicts the results for the combined turn change and next speaker prediction model. The results show that for AMI, a maximum accuracy of 65.12%, and a F1 value of 0.65 is obtained via the XGB and RF algorithms. The value using XGB is greater than the value obtained (64.04% and F1=0.64) when next speaker prediction model is considered as an individual model. However for RF, the value achieved via the individual next speaker prediction model (65.54%, and F1 = 0.66) is greater than the combined model. Concerning MPR, a maximum accuracy of 65.39% and a F1 value of 0.65 are obtained using the XGB algorithm, which is similar to the maximum values (65.64% and F1=0.65) obtained via individual next speaker prediction model.

For both AMI and MPR datasets, the combined turn change and next speaker prediction models outperform the baseline. However, the comparison between individual and combined next speaker prediction models show that the performance difference between the two models is not significant at $p < 0.05$

TABLE IV
RESULTS FOR MODEL COMBINING TURN CHANGE AND NEXT SPEAKER PREDICTION (ACCURACIES IN %, WEIGHTED F1 VALUES IN BRACKETS).

Algorithm	MPR	AMI
XGB	**65.39 (0.65)**	**65.12 (0.65)**
RF	63.71 (0.63)	**65.12 (0.65)**
MLP	61.05 (0.61)	43.92 (0.44)
SVM	56.52 (0.56)	49.53 (0.49)
KNN	56.46 (0.56)	46.18 (0.46)
LR	56.67 (0.56)	51.51 (0.51)
Baseline (Majority Class)	**36.62**	**32.81**

(p=0.14 on AMI and p=0.66 on MPR).

3) Ablation Study: The Table V shows the results obtained during the ablation study performed for turn change prediction on AMI and MPR. The results show that for both AMI and MPR, in the best case (using the XGB algorithm) the models trained using both addressee role and pause duration (ami-tc+ar+pd, mpr-tc+ar+pd), outperform the models trained without these features (ami-tc-ar-pd, mpr-tc-ar-pd) and the models trained including only one of these features in the feature set (ami-tc+ar-pd, ami-tc-ar+pd, mpr-tc+ar-pd, and mpr-tc-ar+pd). These results are significant at $p < 0.05$ (p=0.04 on AMI and p=0.01 on MPR).

The Table VI depicts the results of the ablation study performed for next speaker prediction. The results show that for both AMI and MPR, the models trained using addressee role and pause duration outperform both those trained without these features and the models trained including only one of these features. For both AMI and MPR, the best results are obtained via the XGB algorithm. These results are significant at $p < 0.05$ (p=0.04 on AMI and p=0.01 on MPR).

VI. DISCUSSION, CONCLUSION & PERSPECTIVES

Results from Tables II, III and IV show that both individual and combined turn change and next speaker prediction models perform better than baselines on AMI and MPR. Furthermore, combining turn change and next speaker prediction (*i.e.* exploiting predicted turn change as additional feature for next speaker prediction) does not yield any significant performance improvement. One of the reason could be that the error from turn change prediction model propagates to next speaker prediction model.

Moreover, the ablation study shows a significant performance improvement for turn change and next speaker prediction when addressee role and pause duration are added to the feature set.

One implementation difficulty concerns dynamic feature values such as pause duration and speaker focus. One of the possible solutions to estimate the dynamic values of pause duration and speaker focus is to start a thread as soon as the end of an utterance is detected. Then, the thread monitors the time elapsed since the last utterance and tracks the speaker gaze. The combined feature vector that includes the updated values of pause duration and speaker focus can be transmitted to the turn change and next speaker modules regularly. When

TABLE V
ABLATION STUDY FOR TURN CHANGE PREDICTION USING AMI AND MPR DATASETS (ACCURACIES IN %, F1 VALUES IN BRACKETS).

Alg.	ami-tc-ar-pd	ami-tc+ar-pd	ami-tc-ar+pd	ami-tc+ar+pd	mpr-tc-ar-pd	mpr-tc+ar-pd	mpr-tc-ar+pd	mpr-tc+ar+pd
XGB	65.96 (0.66)	65.96 (0.66)	87.09 (0.87)	**87.59 (0.88)**	75.94 (0.70)	75.94 (0.71)	82.70 **(0.82)**	**83.02 (0.82)**
RF	63.62 (0.64)	65.46 (0.65)	**86.42 (0.86)**	**86.42 (0.86)**	68.63 (0.68)	69.27 (0.68)	82.07 (0.81)	**82.32 (0.82)**
MLP	62.19 (0.62)	63.28 **(0.63)**	62.19 (0.62)	**64.36 (0.64)**	73.90 (0.70)	74.49 (0.72)	**80.99 (0.81)**	80.43 (0.80)
SVM	66.13 **(0.66)**	66.47 **(0.66)**	66.13 **(0.66)**	**66.55 (0.66)**	75.84 (0.70)	**76.46 (0.71)**	75.85 (0.70)	**76.46 (0.71)**
KNN	62.95 **(0.63)**	62.95 **(0.63)**	63.36 **(0.63)**	**62.95 (0.63)**	73.47 (0.71)	73.79 (0.71)	**78.48 (0.77)**	76.72 (0.75)
LR	63.28 (0.63)	65.80 **(0.66)**	63.28 (0.63)	**65.83 (0.66)**	75.60 (0.70)	75.87 **(0.71)**	75.62 (0.70)	**75.85 (0.71)**

TABLE VI
ABLATION STUDY FOR NEXT SPEAKER PREDICTION USING AMI AND MPR DATASETS (ACCURACIES IN %, WEIGHTED F1 VALUES IN BRACKETS).

Alg.	ami-ns-ar-pd	ami-ns+ar-pd	ami-ns-ar+pd	ami-ns+ar+pd	mpr-ns-ar-pd	mpr-ns+ar-pd	mpr-ns-ar+pd	mpr-ns+ar+pd
XGB	50.12 (0.50)	51.38 (0.51)	61.86 (0.62)	**64.04 (0.64)**	49.26 (0.47)	57.19 (0.57)	58.39 (0.57)	**65.64 (0.65)**
RF	46.77 (0.47)	49.79 (0.50)	61.77 (0.62)	**65.54 (0.66)**	41.22 (0.41)	47.93 (0.48)	56.31 (0.56)	**64.02 (0.64)**
MLP	43.92 (0.44)	44.25 (0.44)	43.92 (0.44)	**45.76 (0.46)**	47.26 (0.46)	53.81 (0.54)	55.79 (0.55)	**61.77 (0.61)**
SVM	48.95 (0.49)	51.38 (0.51)	48.95 (0.49)	**51.41 (0.51)**	46.74 (0.42)	57.29 **(0.57)**	47.73(0.43)	**57.43 (0.57)**
KNN	45.51 (0.45)	48.28 **(0.48)**	45.59 (0.46)	**48.36 (0.48)**	44.93 (0.45)	52.95 (0.53)	50.10 (0.50)	**56.33 (0.56)**
LR	46.68 (0.47)	50.96 **(0.51)**	46.68 (0.47)	**50.88 (0.51)**	46.09 (0.46)	56.95 **(0.56)**	46.14(0.41)	**56.96 (0.56)**

the models predict that the turn changes and select the next speaker, the thread stops.

Even if the proposed turn change and next speaker prediction models outperform existing baselines, there is still room for improvement. For instance, in addition to the features used in the proposed models, existing works show that prosody, head and hand gestures can also be exploited. Thus, adding these features can further improve the performance of turn change and next speaker prediction models. However, currently, none of the existing datasets contains these features along with the features exploited in this research work.

REFERENCES

[1] J. Allwood, "An activity based approach to pragmatics," *Abduction, belief and context in dialogue: Studies in computational pragmatics*, pp. 47–80, 2000.

[2] N. Guntakandla and R. Nielsen, "Modelling turn-taking in human conversations," in *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction, Stanford CA*, 2015.

[3] T. Meshorer and P. A. Heeman, "Using past speaker behavior to better predict turn transitions." in *Interspeech*, 2016, pp. 2900–2904.

[4] I. De Kok and D. Heylen, "Multimodal end-of-turn prediction in multi-party meetings," in *ICMI*, 2009, pp. 91–98.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, no. 4, p. 696735.

[6] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, 1969.

[7] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *ICASSP*, vol. 1, 1992, pp. 517–520.

[9] Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task," in *ICASSP*, 2018, pp. 6159–6163.

[10] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[11] V. Petukhova and H. Bunt, "Whos next? speaker-selection mechanisms in multiparty dialogue," in *Workshop on the Semantics and Pragmatics of Dialogue*, 2009.

[12] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *Interspeech*, 2012.

[13] R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato, "Predicting next speaker and timing from gaze transition patterns in multi-party meetings," in *ICMI*, 2013, pp. 79–86.

[14] R. Ishii, S. Kumano, and K. Otsuka, "Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings," in *ICMI*, 2015, pp. 99–106.

[15] R. Ishii , S. Kumano, and K. Otsuka, "Predicting next speaker based on head movement in multi-party meetings," in *ICASSP*, 2015, pp. 2319–2323.

[16] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and J. Tomita, "Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation," *Multimodal Technologies and Interaction*, vol. 3, no. 4, p. 70, 2019.

[17] K. Funakoshi, "A multimodal multiparty human-robot dialogue corpus for real world interaction," *LREC*, pp. 35–39, 2018.

[18] D. C. O'Connell and S. Kowal, *Dialogical genres: Empractical and conversational listening and speaking*. Springer Science & Business Media, 2012.

[19] E. E. Hilbrink, M. Gattis, and S. C. Levinson, "Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction," *Frontiers in psychology*, vol. 6, p. 1492, 2015.

[20] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *MB*, vol. 88, 2005, p. 100.

[21] H. Bunt, "The dit++ taxonomy for functional dialogue markup," in *TSMLEDA@AAMAS09*, 2009, pp. 13–24.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *in SIGKDD*. ACM, 2016, pp. 785–794.

[23] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held, "Multi-layer perceptrons," in *Computational Intelligence*, 2013, pp. 47–81.

[24] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[25] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 2013, vol. 398.

[26] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[27] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *GRC*, vol. 2, 2005, pp. 718–721.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.