A Computational Study on Sentence-based Next Speaker Prediction in Multiparty Conversations

Meng-Chen Lee
mlee45@uh.edu
University of Houston
Houston, TX, USA

Wu Angela Li*
wl71@rice.edu
Rice University
Houston, TX, USA

Zhigang Deng†
zdeng4@central.uh.edu
University of Houston
Houston, TX, USA

ABSTRACT

In this paper we present a computational study to quantitatively examine the task of predicting the next speaker in multi-party conversations using machine learning models. To accomplish this, we create features that accurately represent information relevant to speaker changes in such conversations. We utilize sentence-based models, rather than the widely-used InterPausal Unit (IPU)-based models, and extend the definition of verbal backchanneling to include additional reactions that signify listeners' attention or interest. Through extensive experiments with various machine learning models and inputs, we show that our sentence-based models outperform existing IPU-based models, with the best model achieving 61.39% accuracy. Our study provides design implications and recommendations for the development of virtual agents or humanoid robots with interactive social interaction capabilities.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

Multiparty Conversations, Human-human Communication, Multimodal Interaction, Next Speaker Prediction, Interactive Social Interaction, Machine Learning

ACM Reference Format:
Meng-Chen Lee, Wu Angela Li, and Zhigang Deng. 2024. A Computational Study on Sentence-based Next Speaker Prediction in Multiparty Conversations. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3652988.3673915

1 INTRODUCTION

Various machine learning models have been proposed for turn-change detection and next speaker prediction in multiparty conversations, typically using InterPausal Units (IPUs) [12, 13, 17]. These IPU-based methods involve two stages: detecting a turn change and then predicting the next speaker if a change is detected. Our study approaches this problem at the sentence level to enhance prediction accuracy and eliminate the need for two models, based on the hypothesis that speaker changes are unlikely before a sentence concludes.

Focusing on three-party conversations, our research suggests that sentence-level models may outperform IPU-based models. Using only the last portion of an interval (e.g., the last 1.5 seconds) provides sufficient information for accurate predictions. We also introduce an extended notion of verbal backchanneling, encompassing a wider range of reactions to improve model performance.

Existing studies highlight the role of auditory and visual signals in turn-taking, significantly impacting speech diarization and inspiring the integration of visual features to determine speaker turns [3, 7, 8, 10]. Researchers have explored identifying speakers from visual cues and using speech input to animate virtual conversations [6, 11, 14, 15, 22]. Previous computational models for turn-taking in dyadic conversations informed our approach [24, 26, 27, 32]. Lee et al. developed an IPU-based model using Relatively Engagement Level (REL) [21], while Kawahara et al. [17] and Ishii et al. [12, 13] focused on multiparty settings, addressing class imbalance with two-stage models. Kawahara et al. [17] achieved a first-stage accuracy of 70.6% and a second-stage accuracy of 69.7%, Ishii et al. [12] achieved 76.2% and 55.2%, and Ishii et al. [13] achieved 69.5% and 48.8%. Our approach builds on these insights, focusing on sentence-level data to streamline and enhance turn-change detection and next speaker prediction.

2 DATA COLLECTION

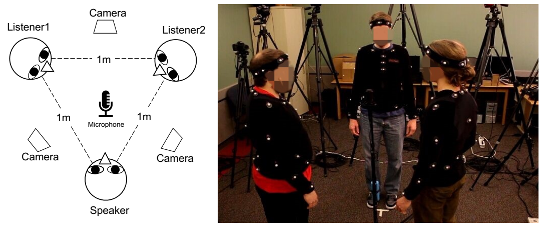

Figure 1: (Left) Illustration of the data acquisition configuration. (Right) A snapshot of our data acquisition process for three-party conversational motion capture.

In this study, we utilized a hybrid data acquisition system consisting of a VICON optical motion capture system, HD video cameras, and a high-quality microphone system to collect a multi-modal, three-party conversational behavior dataset. The dataset includes 12 distinct conversation sessions with six participants (three males

*Meng-Chen Lee and Wu Angela Li contributed equally to this research. Wu Angela Li did this research when did summer internship at University of Houston.
†The Corresponding Author

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673915

and three females) who did not know each other. Participants were randomly assigned to two groups of three, and each group engaged in six sessions discussing various topics. During data capture, participants formed an equilateral triangle with one meter of distance between them [3, 6, 15, 16], remaining stationary but allowed upper body movement. This setup ensured comfort and equal interaction among participants. We captured multimodal data including gaze, head motion, audio, facial expressions, hand gestures, and upper body motion using a ten-camera VICON system, high-quality microphones, and HD video cameras. Eye movements were recorded with close-up HD video cameras [20], and all data were processed at 30 frames per second for consistency (see Figure 1) .

3 FEATURES EXTRACTION

Initially, a language expert transcribed the audio, corrected the system's output, annotated verbal backchanneling, and adjusted timestamps. We then used a forced phoneme alignment tool from automatic speech recognition systems to time each word and sentence in the audio. This process focused on sentence ends rather than IPUs, defining an *interval* as a quadruple I representing start and end times, a speech transcript or backchannel category, and an interlocutor ID. Only speech intervals of at least 1.5 seconds were considered, averaging 4.12 seconds with a maximum of 20.28 seconds, significantly longer than IPUs in previous studies [1], potentially reducing computational expenses. Intervals with interruptions were discarded to maintain data quality.

Prosody Features. We extracted pitch and intensity from acoustic speech using a window size of 33.3 ms with the Praat tool [30], creating a 1×2 *prosodic feature vector* for each frame.

Verbal Backchanneling. We categorized verbal backchannels into non-lexical (e.g., uh-huh, mhm) and phrasal backchannels, considering only those shorter than 1.5 seconds. Our term "verbal backchannel" includes both non-lexical backchannels and reactions like laughing, sighing, or gasping. We identified six common types of reactions: *(hum)*, *(laughter)*, *(gasp)*, *(sigh)*, *(surprise)*, and *(acknowledgment)*, resulting in 1012 instances. We constructed a 1×3 *verbal backchanneling vector* for each frame to indicate whether an interlocutor gave a verbal backchanneling response.

Gestural Backchanneling. Nonverbal gestures like nodding or shaking the head signal listener attention. Using an algorithm by Kawato and Ohya [18], we detected head movements with 86.2% accuracy. Frames were categorized into *stable*, *extreme*, and *transient* states based on yaw angle variations. A headshake is detected between two stable frames if there are at least two extreme frames between them and the yaw angle difference exceeds 3 degrees. We constructed a 1×6 gestural backchanneling vector for each frame to indicate head movements. Figure 2 illustrates this process, which is also used for head nods based on pitch angle.

Gaze Targets. Gaze target features are represented as a 1×6 vector for each frame, indicating whether the i-th interlocutor is looking at the j-th interlocutor. Based on our equilateral triangle setup, we determine gaze targets as follows: an interlocutor is considered to be looking at another if their horizontal gaze angle is between -45 and -15 degrees or 15 and 45 degrees, and their vertical gaze angle is between -30 and 30 degrees. This is illustrated in Figures 3.

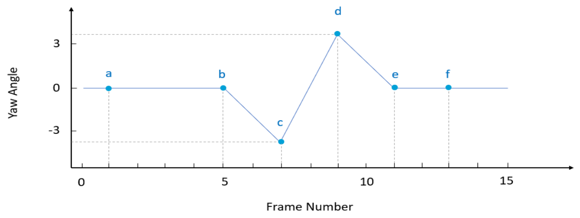

Figure 2: A simplified illustrative example to show the detection of a headshake motion. The states of the labeled points from a to f are: stable, transient, extreme, extreme, transient, and stable, respectively.

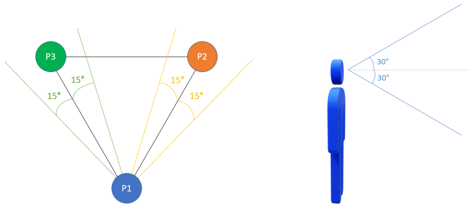

Figure 3: (Left) Illustration of the valid horizontal range of an interlocutor's gaze angle when looking at another interlocutor. Recall that the three interlocutors form an equilateral triangle. (Right) Illustration of the valid vertical range of an interlocutor's gaze angle when looking at another interlocutor.

Current and Next speaker. The current speaker is indicated by a 1×3 binary-valued *current speaker vector* for each frame. After processing speech overlaps, each current speaker vector will have exactly one component with a value of 1. Labels for supervised learning are represented by a 1×3 binary-valued *next speaker vector.*

In our analysis, we focused on the final 45 frames of each data observation, corresponding to 1.5 seconds. This deliberate choice was guided by insights gleaned from prior psycholinguistic investigations, particularly those articulated by Sacks et al. [25]. For each frame, we constructed a 1×20 vector by concatenating the features mentioned above. These vectors were horizontally concatenated across all 45 frames, resulting in a 1×900 vector for each data observation. The labels for these feature vectors were the *next speaker vectors* (1×3).

4 METHODOLOGY

To evaluate the performance of the next speaker prediction task, we conducted experiments using various popular machine learning (ML) models, including Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB), and Random Forest (RF). Our processed data included input feature vectors and corresponding labels. Due to the small dataset size, we used a 90-10 train-test split strategy with ten trials, each with a different split. We fine-tuned all ML models using an exhaustive hyperparameter grid search with 5-fold cross-validation [23]. This approach ensured robust and convincing results, providing insights into model performance variance across splits. Additionally, ensemble learning algorithms like Random Forest were used for their robustness to overfitting [2].

We also developed and evaluated four deep neural networks: one based on Recurrent Neural Networks (RNN), another on Gated Recurrent Units (GRU), a third on Long Short-Term Memory (LSTM) networks [28], and the fourth on the Transformer architecture [31]. The LSTM-based network consisted of two stacked LSTM layers, each with 128 units, 0.2 dropout, 0.2 recurrent dropout, one hidden layer with ReLU activation, and an output layer with softmax activation. The Transformer-based network included a positional embedding layer, 8 transformer blocks with 4 attention heads of size 256, and a multi-layer perceptron with ReLU and dropout layers. Four networks were trained with categorical cross-entropy loss, a batch size of 16, and early stopping with a patience of 3 epochs. Features for deep learning models were vertically concatenated, shifting from a 1 × 900 data format to a 45 × 20 matrix. To mitigate overfitting, we employed hyperparameter tuning, dropouts, and designed generalizable model features [29].

Table 1: Experimental results (average ± standard deviations) by the seven traditional machine learning models and the four deep learning models

Model	Accuracy	Precision	F1-Score
SVM	55.64 ± 3.23	57.15 ± 3.95	56.01 ± 3.35
LR	48.40 ± 6.42	49.10 ± 7.05	48.28 ± 6.70
GB	58.40 ± 5.27	59.45 ± 4.71	58.49 ± 5.23
KNN	39.15 ± 6.71	40.38 ± 7.37	39.15 ± 6.64
DT	56.06 ± 4.09	59.89 ± 4.54	56.41 ± 4.03
NB	51.28 ± 6.22	51.78 ± 8.39	48.36 ± 8.74
RF	57.13 ± 4.51	57.37 ± 7.13	56.64 ± 6.39
simple RNN	52.45 ± 7.20	54.82 ± 6.26	49.08 ± 8.55
GRU	59.57 ± 4.68	61.06 ± 4.82	58.62 ± 5.50
LSTM	**61.17 ± 2.80**	**62.13 ± 3.72**	**61.03 ± 3.13**
TransF	47.02 ± 6.74	42.13 ± 13.03	41.47 ± 11.02
LSTM (Entire)	60.74 ± 2.94	61.22 ± 3.58	59.74 ± 3.31
LSTM (w/ Overlap)	45.00 ± 4.66	46.05 ± 6.01	41.80 ± 5.16
LSTM (Less VBack)	55.85 ± 4.58	56.47 ± 4.77	55.12 ± 4.23
LSTM (No Prosody)	59.89 ± 5.72	60.38 ± 6.60	59.58± 6.00
LSTM (NO GBack)	59.47 ± 3.63	59.68 ± 4.15	59.12 ± 3.77
LSTM (No VBack)	54.15 ± 4.66	54.56 ± 4.94	53.67 ± 4.78
LSTM (No GTV)	47.34 ± 3.22	49.40 ± 4.76	43.95 ± 3.35

5 EXPERIMENTAL RESULTS

We conducted a comprehensive set of experiments, encompassing the comparative analysis of eleven distinct machine learning models, outcomes associated with utilizing the entire interval, results following the truncation of interruptions, effects of employing simpler verbal backchanneling, and ablation studies Our task involves addressing a multiclass classification problem, and we computed key performance metrics, including accuracy, precision, and F-measure, using a weighted-average methodology. This approach ensured that each class or conversational behavior was afforded equal significance during the evaluation process, providing a thorough and equitable assessment of our model's overall performance.

Table 1 shows that the LSTM outperformed all seven traditional models across all metrics, though its performance variance across different trials was less than optimal, possibly due to the small dataset size. The Transformer-based model (TransF) showed notably sub-par performance, particularly in precision, likely affected by the small training set, as LSTM networks tend to perform better than BERT for small datasets in NLP applications [9]. The LSTM model achieved an average accuracy of 61.17%, significantly improving over the random-guess baseline of 33.33% (an increase of **183.53%**). This compares favorably with the best models in previous related works, which achieved accuracies of 48.8%, which is **146.55%** of the baseline of 33.33%, and 69.7%, which is **139.40%** of the baseline of 50.00%, respectively.

We performed additional experiments considering the entire speech interval rather than just the last 45 frames, using zero-prepadding to handle variable input sizes. Our empirical results, aligned with previous research in psycholinguistics and machine learning [4, 5, 16, 19, 25], support that the final portion of an interval is critical for detecting speaker changes. We also explored truncating interrupted speech intervals while leaving interrupting intervals unchanged. Although this increased our dataset size by 20.43%, it significantly degraded model performance, indicating that the trade-off was not worthwhile. We further assessed the impact of the expanded verbal backchanneling definition, showing its importance for model performance compared to traditional definitions. Last, in our ablation studies using the LSTM model, we examined the impact of removing specific features on performance metrics. Removing prosody features led to a modest decline, while excluding gestural or verbal backchanneling features resulted in more significant performance reductions. Omitting gaze target vector (GTV) features caused the most substantial decrease, highlighting the importance of each feature.

6 DISCUSSION AND CONCLUSION

Our experiments with various models and feature formulations identified that LSTM-based models achieve the best performance. This suggests that sentence-level models often outperform IPU-based models in predicting the next speaker. Utilizing only the last portion of an interval, such as 45 frames, provides sufficient information and yields comparable model performance. Additionally, incorporating our extended notion of verbal backchanneling, encompassing a broader range of reactions, enhances machine learning model performance for next speaker prediction. However, our work has some limitations. We did not account for the different heights of participants when classifying gaze targets, which could affect prediction accuracy. Additionally, while interruptions in speech are common in natural conversations, our model's current handling of interruptions may not yield satisfactory performance.

ACKNOWLEDGMENTS

This work was supported in part by US NSF IIS2005430.

REFERENCES

[1] Brigitte Bigi and Béatrice Priego-Valverde. 2019. Search for Inter-Pausal Units: application to Cheese! corpus. In *9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland, 289–293. https://hal.archives-ouvertes.fr/hal-02428485

[2] L Breiman. 2001. Random Forests. *Machine Learning* 45 (10 2001), 5–32. https://doi.org/10.1023/A:1010950718922

[3] Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics* 28, 3 (2017), 449–483. https://doi.org/10.1515/cog-2016-0119

[4] Iwan de Kok and Dirk Heylen. 2009. Multimodal End-of-Turn Prediction in Multi-Party Meetings. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (Cambridge, Massachusetts, USA) *(ICMI-MLMI '09)*. Association for Computing Machinery, New York, NY, USA, 91–98. https://doi.org/10.1145/1647314.1647332

[5] Alfred Dielmann, Giulia Garau, and Herve Bourlard. 2010. Floor Holder Detection and End of Speaker Turn Prediction in Meetings. (01 2010).

[6] Yu Ding, Yuting Zhang, Meihua Xiao, and Zhigang Deng. 2017. A Multifaceted Study on Eye Contact Based Speaker Identification in Three-Party Conversations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3011–3021. https://doi.org/10.1145/3025453.3025644

[7] Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology* 23 (1972), 283–292.

[8] Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns. *Language in Society* 3, 2 (1974), 161–180. https://doi.org/10.1017/S0047404500004322

[9] Aysu Ezen-Can. 2020. A Comparison of LSTM and BERT for Small Corpus. (09 2020).

[10] Charles Goodwin. 1981. *Conversational Organization: Interaction Between Speakers and Hearers.*

[11] Sebastian Gorga and Kazuhiro Otsuka. 2010. Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8.

[12] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2319–2323. https://doi.org/10.1109/ICASSP.2015.7178385

[13] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Transactions on Interactive Intelligent Systems* 6 (05 2016), 1–31. https://doi.org/10.1145/2757284

[14] Aobo Jin, Qixin Deng, and Zhigang Deng. 2022. A Live Speech Driven Avatar-mediated Three-party Telepresence System: Design and Evaluation. *PRESENCE: Virtual and Augmented Reality* 29 (2022), 1–27.

[15] Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A deep learning-based model for head and eye motion generation in three-party conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (2019), 1–19.

[16] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and Turn-Taking Behavior in Casual Conversational Interactions. *ACM Trans. Interact. Intell. Syst.* 3, 2, Article 12 (aug 2013), 30 pages. https://doi.org/10.1145/2499474.2499481

[17] Tatsuya Kawahara, T. Iwatate, and Katsuya Takanashi. 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012* 1 (01 2012), 726–729.

[18] S. Kawato and J. Ohya. 2000. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. 40–45. https://doi.org/10.1109/AFGR.2000.840610

[19] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

[20] Binh H Le, Xiaohan Ma, and Zhigang Deng. 2012. Live speech driven head-and-eye motion generators. *IEEE transactions on visualization and computer graphics* 18, 11 (2012), 1902–1914.

[21] Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 436–444.

[22] Kazuhiro Otsuka. 2011. Multimodal conversation scene analysis for understanding people's communicative behaviors in face-to-face meetings. In *Symposium on Human Interface*. Springer, 171–179.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[24] Antoine Raux and Maxine Eskenazi. 2009. A Finite-State Turn-Taking Model for Spoken Dialog Systems. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, 629–637. https://doi.org/10.3115/1620754.1620846

[25] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (1974), 696–735. http://www.jstor.org/stable/412243

[26] Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. 2002. Learning decision tree to determine turn-taking by spoken dialogue systems. https://doi.org/10.21437/ICSLP.2002-293

[27] David Schlangen. 2006. From reaction to prediction experiments with computational models of turn-taking. *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP* 5. https://doi.org/10.21437/Interspeech.2006-550

[28] Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1 (jan 2014), 1929–1958.

[30] Will Styler. 2013. Using Praat for linguistic research. *University of Colorado at Boulder Phonetics Lab* (2013).

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[32] Nigel Ward, Olac Fuentes, and Alejandro Vega. 2010. Dialog prediction for a general model of turn-taking. *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2662–2665. https://doi.org/10.21437/Interspeech.2010-706