INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multimodal Turn-Taking Model Using Visual Cues for End-of-Utterance Prediction in Spoken Dialogue Systems

Fuma Kurata[1], *Mao Saeki*[1], *Shinya Fujie*[2], *Yoichi Matsuyama*[1]

[1]Waseda University
[2]Chiba Institute of Technology

kurata@pcl.cs.waseda.ac.jp

Abstract

In this study, we propose a multimodal model for predicting the end-of-utterance probability in spoken dialogue systems, highlighting the unique role of visual cues in addition to acoustic and linguistic information. Although the effectiveness of visual cues, such as gaze, mouth, and head movements, has been suggested, few studies have fully incorporated them into turn-taking models, and the relative importance of these visual cues has also been underresearched. To address these issues, we first conducted an ablation study on visual features, showing the larger contribution of eye movements than mouth and head movements. Additionally, an end-to-end visual feature extraction model utilizing 3D-CNN is employed to comprehensively capture these visual cues. By combining visual features with acoustic and verbal information, AUC score for end-of-utterance prediction improved from 0.896 to 0.920, demonstrating the effectiveness of incorporating these visual cues in turn-taking models.

Index Terms: spoken dialog systems, turn-taking, multimodal machine learning

1. Introduction

This study proposes a turn-taking model incorporating visual, acoustic, and linguistic features. Traditionally, turn-taking research in spoken dialogue systems has extensively relied on linguistic and acoustic information as turn-taking cues, such as syntactic completion, semantics, intonation, speaking speed, and other voice characteristics [1]. However, visual cues, such as eye gaze, mouth movements, nodding, and gestures, have also been found to be effective turn-taking cues in conversation analysis studies. For instance, Kendon [2] found that in face-to-face conversations, speakers often look away at the beginning of their speech and look back at the end. Vincent et al. [3] demonstrated a strong correlation between mouth-opening movements and speech production, while Maynard [4] showed that in conversations in Japanese, speakers use vertical head movements to indicate the end of a clause or turn. Furthermore, Duncan [5] found that the end of the turn can be indicated by completing a hand gesture. These previous studies suggest that by incorporating these visual cues, turn-taking models have the potential to improve accuracy. This study aims to investigate two research questions.

1. How effective are visual cues, such as gaze, mouth, and head movements, as turn-taking cues?
2. To what extent can turn-taking model performance be improved by incorporating visual information in addition to audio and verbal cues?

To answer these questions, we build an end-of-utterance prediction model using facial feature points that incorporate motion features of the eyes, mouth, and head. We then conduct an ablation study to examine the effectiveness of these motion features. We also extract visual features using 3D-CNN and fuse them with acoustic and language features to build a multimodal model for predicting end-of-utterance. This study focuses on online interview dialogues, and gestures are not included during validation as they are often not captured in this setting. The results of this study have the potential to improve turn-taking model accuracy by demonstrating the effectiveness of visual cues and the importance of incorporating comprehensive visual information in turn-taking models.

2. Related work

2.1. Classification of turn-taking models

Skantze categorized turn-taking models in spoken dialogue systems into three groups [1]. The first category is represented by the Silence-based model, which utilizes the speaker's silence longer than a certain threshold to indicate the end of utterances. However, it is not realistic to suppress interruptions while minimizing response delay [6]. The second category is the IPU-based model, where IPUs (inter-pausal units) are speech segments without silence for a specific duration (e.g., 200 ms). The end of the IPUs are detected using VAD or ASR, at which point TRPs (transition-relevant places) are predicted from the turn-taking cues from the speaker. TRPs are defined as the point where a turn change can occur. Accurate prediction of TRPs allows the system to take turns with minimal gaps and avoid interrupting the user at non-TRP points. The third category is the Continuous model, which continuously processes user speech without relying on VAD or ASR. This model offers the benefit of predicting not just TRPs, but also BRPs (backchannel-relevant places) and intentional system interruptions during user speech. This paper adopts the IPU-based turn-taking model, which separates the optimization problem of turn-taking timing from the problem of TRP prediction. By focusing on the problem to just TRP prediction, we assess the effectiveness of visual cues in turn-taking by comparing TRP prediction performance with and without visual information.

2.2. Multimodal turn-taking model using visual cues

The literature on turn-taking models for spoken dialogue systems is vast, with many studies focusing on linguistic or acoustic cues or both [7, 8, 9, 10, 11, 12, 13]. Meanwhile, some studies have also incorporated visual cues. De Kok et al. [14] used a sequential probabilistic model that incorporated the head gestures of the participants to predict the end of a speaker's turn in a multi-person conversation. Roddy et al. [15] proposed a

10.21437/Interspeech.2023-578

turn-taking model that incorporates different timescale modalities and demonstrated the inclusion of gaze features along with language and acoustic features. These studies, however, did not fully utilize eye, mouth, and head movements, which have been established as valid visual turn-taking cues in previous research. In contrast, Ishii et al. [16] utilized comprehensive visual cues by extracting high-level representations of visual information as turn-taking cues from images of speakers using Resnet. However, they extracted visual information end-to-end, making it unclear which specific visual features were effective. To the best of our knowledge, this study is the first to investigate in detail the effectiveness of visual cues in a turn-taking model. Accordingly, we conduct an ablation study to identify practical visual features for turn-taking cues.

3. Dataset

To create a dataset of turn-taking samples, we adopted online interviews that were originally designed to assess an English speaking ability [17]. The data comprises 210 dialogues of a 10-minute online interview conversation between a Japanese learner of English and an English teacher. Figure 1 illustrates the process of creating video clips. The audio and video of both the interviewer and interviewee were segmented into IPUs, and a 5-second video clip was created by cutting from the end of each IPU. In this study, each video was segmented by a silence longer than 300ms. Each video clip included a 0.3-second silent interval after the end of the IPU. A total of 21,728 video clips were created and the samples were divided into training, validation, and evaluation sets: the dataset contained 18,910 training samples, 1,911 validation samples, and 1,046 evaluation samples.

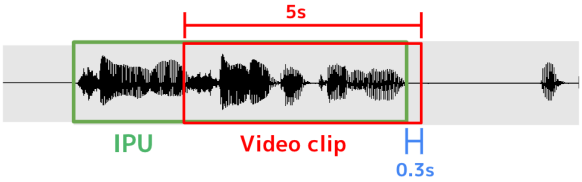

Fig. 1. *Dataset creation process: Clip out the last 5 seconds of IPUs to create video clips for the dataset. Each video clip contains a 0.3-second silence interval after the IPU.*

To systematically generate training data, we used multiple rules to automatically label the training and validation data. During the interview, the speaker either continues speaking after IPU or gives the turn to the interlocutor. We assigned the label of ”continue utterance” in the former case and that of ”end utterance” in the latter case. Multiple conditions were set using voice activity detection (VAD) and transcription of dialogue sentences to execute this process. The training data consisted of 8,109 ”continue utterance” samples and 10,801 ”end utterance” samples. To accurately measure the model performance, the test data were labeled by three annotators. In cases where a judgment cannot be made, the label ”unknown” was assigned, and the ground truth was determined based on the majority vote. Nineteen samples without a majority were excluded from the evaluation data. The inter-rater agreement was 0.767 using Krippendorf’s alpha. The test data consisted of 530 ”continue utterance” samples and 497 ”end utterance” samples.

4. Ablation study on visual cues

The literature of conversation analysis suggests that gaze [2], mouth [3], and head movements [4] are useful cues for turn-taking. However, no study has investigated the effectiveness of individual cues in the turn-taking model. We conducted an ablation study on visual cues to evaluate their effectiveness.

4.1. Design of visual features

We used the speaker’s face landmark coordinates to extract visual features of gaze, mouth, and head movements. To obtain the coordinates, we used MediaPipe’s FaceMesh[1] and collected 478 landmark points from the speaker’s face. We calculated three types of feature, as shown in Table 1, using the coordinates of the eyes, mouth, and facial contours. The first feature is the pupil position, which captures the relative position of the pupil concerning the eye size in the x and y-axis directions. The second feature is the mouth opening, which indicates the extent of mouth opening relative to the face size in both x and y directions. The third feature is the head direction, which indicates the inclination of the speaker’s head in roll, pitch, and yaw directions relative to the camera’s front direction. Figure 2 shows the appearance of these features. We standardized these 7 features to have a mean of 0 and a variance of 1.

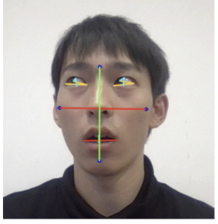

Fig. 2. *Visual feature extraction: The feature points used for extracting pupil positions, mouth opening, and head direction features, along with the line segments connecting each of them.*

4.2. Visual feature extraction model

We used a 5-layer unidirectional LSTM as a feature extraction model to capture the temporal changes in gaze, mouth, and head using the aforementioned features. The number of intermediate dimensions of LSTM was set to 15, and 3 fully connected layers were added on the top of the intermediate layer at the final timestep to obtain an output dimensionality of 2. The final output value was passed through the softmax function to generate the probabilities of ”end utterance” and ”continue utterance”. For the model input, we utilized the final 2-second visual features from the end of the 5-second video clip. We used the dataset described in Section 3 and cross-entropy loss as the objective function, with a learning rate of 5×10^{-3}.

4.3. Result of ablation study

This section explores how effective eye gaze and mouth and head movements are as visual cues for turn-taking. We performed an ablation study on the aforementioned model to evaluate the effectiveness of these features. The experiment set out to examine the effects of removing the pupil position, mouth opening, and head direction features from the input of LSTM end-of-utterance prediction model on its performance.

[1] https://google.github.io/mediapipe/solutions/face_mesh.html

Table 1. *Extraction method for each visual feature*

Features	Extraction method
X-axis - pupil position	(x-axis distance between the center of the pupil and left edge of eye)/(x-axis distance between the right and left edge of eye)
Y-axis - pupil position	(y-axis distance between the center of the pupil and the upper edge of the eye)/(y-axis distance between the top and bottom of the eye)
X-axis - mouth opening	(x-axis distance between the right and left edges of the mouth)/(x-axis distance of the left and right reference feature points of the face)
Y-axis - mouth opening	(y-axis distance between top and bottom of the mouth)/(y-axis distance of the upper and lower reference feature points of the face)
Roll-head direction	(x-axis coordinates of feature points in the upper part of the face)-(x-axis coordinates of feature points in the lower part of the face)
Pitch - head direction	(z-axis coordinates of feature points in the upper part of the face)-(z-axis coordinates of feature points in the lower part of the face)
Yaw - head direction	(z-axis coordinates of feature points in the left part of the face)-(z-axis coordinates of feature points on the right part of the face)

Table 2. *Ablation study on visual features*

	AUC	Accuracy	Precision	Recall	F1
All features	**0.801**	**0.797**	**0.739**	**0.899**	**0.811**
w/o eye feature	0.684	0.678	0.619	0.870	0.723
w/o mouth feature	0.739	0.733	0.660	0.924	0.770
w/o head pose feature	0.758	0.759	0.700	0.875	0.778

Table 2 shows that the model with all features provides the best performance for end-of-utterance prediction, indicating the essential role of pupil position, mouth opening, and head direction features in accurate turn-taking. Furthermore, the worst performance was observed when the pupil positional features were removed, indicating that gaze movement is the most crucial signal for turn-taking among the three movements of gaze, mouth, and head. Additionally, the performance was comparable when removing the mouth opening and head direction features, with a slightly lower performance when excluding the mouth opening feature. This suggests that mouth movement is a relatively more critical cue for turn-taking than head movement.

To further explore the effectiveness of the visual feature extraction model, we qualitatively observed misclassified samples in predicting the end-of-utterance with all visual features. Several samples were found to have incorrect predictions due to the failure to capture movements. Due to the limited number of dimensions in the visual features, the visual cues extraction capability may be inadequate. To enhance the performance of the visual feature extraction model, we explored an end-to-end approach that utilizes higher-dimensional visual features in the second experiment.

5. End-to-End visual feature extraction

5.1. End-to-End model using 3D-CNN

Ishii et al. [16] combined ResNet and GRU for visual feature extraction. ResNet extracted spatial image features while GRU captured their temporal changes. For our visual feature extraction model, we used X3d [18], a 3D-CNN that can extract both spatial and temporal features. X3d is expected to capture visual turn-taking cues more accurately than a 2D-CNN-LSTM model given its superior performance in video object recognition tasks [18, 19]. We selected X3d-S as our model size for this study.

Before being inputted into the X3d model, the image sequence of the video clips underwent the following preprocessing steps: (1) extracting the video frame by frame and retaining the last 2 seconds, (2) extracting the facial region based on the facial landmark coordinates using FaceMesh of MediaPipe, (3) resizing the image and adjusting the length of the long side to the specified length for X3d-S, (4) making the image square by zero-padding the shorter sides, (5) down-sampling the number of frames in the time direction to match the specified length of X3d-S, and (6) normalizing the image colors for image recognition preprocessing. The resulting preprocessed image tensors were then used as input for X3d-S.

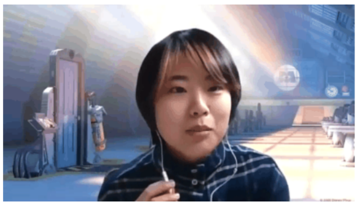

Fig. 3. *Image before processing*

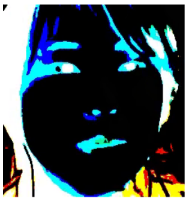

Fig. 4. *Image post processing*

The model was trained using the same process as in Section 4.2. A learning rate of 10^{-4} was used during training.

5.2. Evaluation of End-to-End model

To evaluate the end-to-end model using X3d, we compared it to the non-end-to-end model based on Facemesh and LSTM previously described in Section 4.2. The results are presented in Table 3.

Table 3. *Evaluation of E2E model*

approach	AUC	Accuracy	Precision	Recall	F1
Non-E2E(LSTM)	0.801	0.797	0.738	**0.899**	0.811
E2E(X3d)	**0.830**	**0.831**	**0.805**	0.857	**0.830**

The end-to-end model using X3d-S demonstrated higher AUC, accuracy, and F1 metrics in comparison to the non-end-to-end model using LSTM, indicating a more accurate prediction of the end-of-utterance using the former approach. This is possibly due to the precise capture of gaze information through the utilization of higher-dimensional visual features. Nonetheless, the end of a speaker's utterance depends not only on visual cues but also on prosody and the content of the utterance [1]. The next section examines to what extent incorporating acoustic and linguistic information alongside visual information can enhance predictive performance.

6. Multimodal end-of-utterance prediction

6.1. Overview

Our proposed multimodal end-of-utterance prediction model extracts acoustic, linguistic, and a comprehensive set of visual features from the speaker's face, which are then fused for prediction. Figure 5 shows a whole architecture of the proposed model. The proposed model comprises three submodules that extract acoustic, linguistic, and visual features, fused using five fully connected layers. The model outputs the probability of ”end utterance” and ”continue utterance” by passing through the softmax function. This model can work with a single or a combination of submodules as it is fused by simple concatenation. The following sections will explain the details of each submodule.

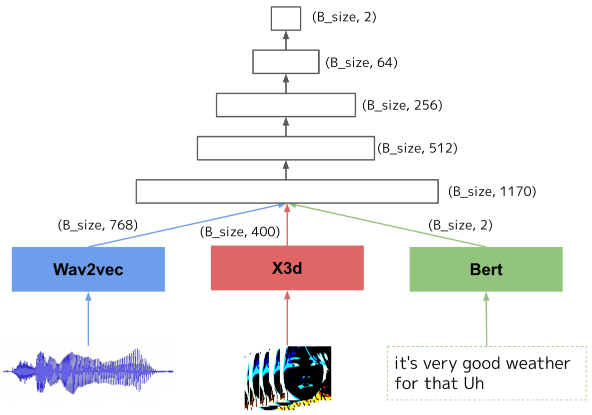

Fig. 5. *The proposed multimodal model architecture: The numbers in parentheses in the figure indicate the dimensionality of the tensor output from each layer. B_size represents the batch size.*

6.2. Feature extraction modules

We utilized Wav2vec 2.0 [20] to extract acoustic turn-taking cues from the speaker's voice. This pre-trained self-supervised learning framework uses a large amount of speech data as a feature extractor to learn these cues. We finetuned the "wav2vec2-base" pre-trained model with the dataset described in Section 3, which was augmented with noise. Visual cues from the speaker were extracted using the method described in Section 5.1, and language features were extracted using Bert [21]. Bert used the transcribed utterances of the interviewer and interviewee as dialogue history, with up to 512 tokens before the end of the video clip. We used the "bert-base-uncased" learned model for Bert.

6.3. Model Training

The proposed multimodal end-of-utterance prediction model was trained in two steps. In step 1, we trained the unimodal end-of-utterance predictors for the acoustic feature extraction model, visual feature extraction model, and language feature extraction model. The problem setup, dataset, and loss function are identical to the model in the ablation study section 4.2. During training, we set the learning rate to 2×10^{-5} for the acoustic feature extraction model, 10^{-4} for the visual feature extraction model, and 5×10^{-6} for the language feature extraction model. In step 2, we trained the fully connected layer of the multimodal end-of-utterance prediction model while keeping the parameters of each submodule fixed. Only the fully connected layer was trained with a learning rate of 10^{-4}.

6.4. Model evaluation

In the experiment, we compared the performance or end-of-utterance prediction for all combinations of acoustic, linguistic, and visual features. The results are presented in Table 4. Among the unimodal models, the model that utilized acoustic information achieved the highest AUC score of 0.887, followed by the model that incorporated visual information with an AUC score of 0.830, and the model that relied solely on linguistic information with a slightly lower AUC score of 0.827. Furthermore, the performance of the multimodal model improved significantly, from an AUC score of 0.896 to 0.920, by combining visual information with acoustic and linguistic information, indicating the usefulness of visual cues in turn-taking. Conversely, the addition of linguistic information to acoustic and visual information only resulted in a marginal improvement in the AUC score, from 0.917 to 0.920. Future research should explore methods for aligning linguistic information with acoustic and visual information in an organic manner.

Table 4. *Evaluation of proposed multimodal model*

Input Feature	AUC	Accuracy	Precision	Recall	F1
Audio	0.887	0.885	0.833	0.953	0.889
Vision	0.830	0.831	0.805	0.857	0.830
Language	0.827	0.826	0.794	0.863	0.827
Audio+Vision	0.917	0.918	**0.896**	0.940	0.917
Audio+Language	0.896	0.896	0.874	0.918	0.895
Vision+Language	0.836	0.835	0.827	0.852	0.833
Audio+Vision+Language	**0.920**	**0.919**	0.891	**0.950**	**0.919**

7. Conclusion

This study confirms the effectiveness of visual information as turn-taking cues and proposes a multimodal end-of-utterance prediction model using comprehensive visual cues. We developed a model that extracts motion features of the eye, mouth, and head based on facial landmark coordinates and predicts the end of utterance. Our ablation study showed that these three motion features contributed to the performance of predicting the end-of-utterance, with eye movements being the most significant factor. We then employed 3D-CNN to extract comprehensive visual features end-to-end, resulting in better performance compared to the non-end-to-end model. Finally, by integrating these visual features with acoustic and linguistic features, we demonstrated that incorporating visual cues improved the end-of-utterance prediction performance, showing the effectiveness of using visual cues in turn-taking models.

8. Acknowledgement

This paper is based on results obtained from a project, JPNP20006 ("Online Language Learning AI Assistant that Grows with People"), subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

9. References

[1] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[2] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, pp. 22–63, 1967.

[3] V. L. Gracco and A. Lofqvist, "Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements," *Journal of Neuroscience*, vol. 14, no. 11, pp. 6585–6597, 1994.

[4] S. K. Maynard, "Interactional functions of a nonverbal sign head movement in japanese dyadic casual conversation," *Journal of pragmatics*, vol. 11, no. 5, pp. 589–606, 1987.

[5] S. Duncan, "Some signals and rules for taking speaking turns in conversations." *Journal of personality and social psychology*, vol. 23, no. 2, p. 283, 1972.

[6] N. G. Ward, A. G. Rivera, K. Ward, and D. G. Novick, "Root causes of lost time and user stress in a simple dialog system," 2005.

[7] E. Ekstedt and G. Skantze, “Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog,” *arXiv preprint arXiv:2010.10874*, 2020.

[8] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, “Prediction of turn-taking using multitask learning with prediction of backchannels and fillers,” *Listener*, vol. 162, p. 364, 2018.

[9] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, “Turn-taking predictions across languages and genres using an lstm recurrent neural network,” in *2018 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2018, pp. 831–837.

[10] J. Yang, P. Wang, Y. Zhu, M. Feng, M. Chen, and X. He, “Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 7747–7751.

[11] D. Lala, K. Inoue, and T. Kawahara, “Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 78–86.

[12] J. Sakuma, S. Fujie, and T. Kobayashi, “Response timing estimation for spoken dialog systems based on syntactic completeness prediction,” in *2022 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2023, pp. 369–374.

[13] R. Masumura, T. Tanaka, A. Ando, R. Ishii, R. Higashinaka, and Y. Aono, “Neural dialogue context online end-of-turn detection,” in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 2018, pp. 224–228.

[14] I. De Kok and D. Heylen, “Multimodal end-of-turn prediction in multi-party meetings,” in *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 91–98.

[15] M. Roddy, G. Skantze, and N. Harte, “Multimodal continuous turn-taking prediction using multiscale rnns,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 186–190.

[16] R. Ishii, X. Ren, M. Muszynski, and L.-P. Morency, “Multimodal and multitask approach to listener’s backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling?” in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021, pp. 131–138.

[17] M. Saeki, Y. Matsuyama, S. Kobashikawa, T. Ogawa, and T. Kobayashi, “Analysis of multimodal features for speaking proficiency scoring in an interview dialogue,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 629–635.

[18] C. Feichtenhofer, “X3d: Expanding architectures for efficient video recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 203–213.

[19] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.