Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation

Kazuyo Onishi
Nara Institute of Science and Technology
Ikoma-shi, Nara, Japan
onishi.kazuyo.oi5@is.naist.jp

Hiroki Tanaka
Nara Institute of Science and Technology
Ikoma-shi, Nara, Japan
hiroki-tan@is.naist.jp

Satoshi Nakamura
Nara Institute of Science and Technology
Ikoma-shi, Nara, Japan
s-nakamura@is.naist.jp

ABSTRACT

Predicting the timing of utterances in dyadic conversations is essential for achieving natural interactions between humans and virtual agents. Since the former often use non-verbal cues to adjust the order of their speech, this study proposes a multimodal model incorporating non-verbal features using a Transformer-based voice activity prediction model. First, in line with previous research, we reproduced a baseline model that utilized audio features (audio waveform, voice activity frame, and voice activity history) as inputs. To this baseline model, we added non-verbal features: gaze direction, action units, head pose, and articular points. We compared our multimodal model with the baseline model to investigate the impact of non-verbal cues on voice activity prediction. We utilized a dyadic expert-novice conversation dataset and evaluated the average outcomes across ten model trainings. Results revealed that our proposed models with all the features improved the accuracy of the next speaker prediction by 2.3% and back-channeled prediction by 1.8% (p-value < 0.025). In particular, action units may contribute significantly to the turn-shift and back-channeled predictions. This study demonstrates that including non-verbal features in Transformer-based turn-taking models enhances the efficacy of models for predicting voice activity in dyadic conversations.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Turn-taking, multimodal, voice activity prediction, transformer

ACM Reference Format:
Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation. In *International Conference on Human-Agent Interaction (HAI '23), December 4–7, 2023, Gothenburg, Sweden.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3623809.3623837

HAI '23, December 4–7, 2023, Gothenburg, Sweden

ACM ISBN 979-8-4007-0824-4/23/12…$15.00
https://doi.org/10.1145/3623809.3623837

1 INTRODUCTION

Turn-taking is a skill employed to ensure smooth communication and mutual understanding, wherein participants alternate speaking roles [44]. Given the inherent difficulty of simultaneously speaking and listening during a conversation, participants must effectively coordinate their roles as speakers and listeners. Humans fluently coordinate such role assignments by quickly switching between speaker and listener roles within 200 ms [28]. This turn-change speed is remarkable, considering it is comparable to human reaction times. Moreover, speaker transitions often occur in a manner that overlaps with the current speaker's speech. Such overlaps are not merely switching errors but are common phenomena in fluent dialogues. Thus, proper conversational switching is very sophisticated, and humans take turns with speakers based on various factors.

However, virtual agents do not yet have sufficient turn-taking capabilities [17]. They frequently demonstrate a tendency to interrupt users or delay their own responses, and the lack of prompt feedback can disrupt the natural flow of conversation. This problem reflects the fact that many virtual agent systems identify a certain period of silence (typically 700 ms) before starting speech. Although setting a shorter silence threshold can hasten response times, it may also result in user interruptions. On the other hand, a long silence threshold might lead the user to mistakenly perceive the system as unresponsive. Moreover, the timing of the speech onset varies depending on the utterance and content, necessitating context-specific assessment. Therefore, turn-taking systems with fixed silence thresholds face such challenges as limited response speed and the inability to adapt to different turn-taking dynamics in varying situations [50, 35].

Another critical aspect of turn-taking is the role of the back-channel. When generating back-channel, since the listener does not intend to take a turn, it is vital to distinguish these actions from those of the current speaker who is attempting to claim the speech role, as previously mentioned. Although back-channel appears to occur at random intervals, humans are adept at recognizing its proper timing and turn shifts. Furthermore, since a back-channel generally does not intend to claim a turn, it is influenced by different factors than those involved in predicting turn shifts [4]. Generating appropriate back-channel for virtual agents is essential to improve user engagement [20, 47].

Hence, turn-taking predictions are pivotal for systems like virtual agents [46, 52]. Incorrect timing obstructs communication and conveys unintended messages to conversation partners [23]. Users are more satisfied with virtual agents when they perceive them as courteous and personable. Since appropriate turn-taking improves such satisfaction, we are working on a model that uses non-verbal

features to achieve a human-like prediction of speech termination and encourages more natural responses. Previous studies have shown that non-verbal cues provide helpful information for turn-taking models, such as shifts of turns and back-channel [26, 25, 16, 29, 5, 22, 21]. However, these studies are merely models that predict specific turn-taking events. In recent years, researchers have moved beyond turn-taking events to general modeling by predicting voice activity itself [12]. Although researchers have actively discussed such models for predicting voice activity regarding verbal and audio cues, few studies have explored non-verbal features.

In this study, we propose a Transformer [48]-based multimodal voice activity prediction model and investigate how adding non-verbal features affects overall voice activity prediction and improves turn-taking performance from a previously used-only audio features model. We integrated gaze direction, action units, head pose, and articular points in addition to the audio features from a previous study [43, 36, 12] and examined the influence of each modality on turn-taking events. Our results show that non-verbal cues provide captivating information for audio features and are effective in voice activity models. The contributions of this study are threefold.

(1) We proposed architecture that incorporates non-verbal features into a voice activity prediction model.

(2) Our model showed that non-verbal features provide essential information for voice activity prediction, and action units may significantly impact them.

(3) Using non-verbal features suggests that virtual agents can engage in natural conversations that faithfully mimic human communication.

2 RELATED WORK

The realization of turn-taking prediction necessitates understanding the cues with which humans anticipate it. Previous studies have identified a variety of predictive signals in dyads and multi-party dialogues. These turn-taking cues mainly fall into three categories: verbal, audio, and non-verbal features.

Verbal features refer to the uttered words and their semantic and pragmatic information, both of which are crucial for a conversation's progression. Completing a syntactic unit is intuitively essential for the current speaker to believe that a turn has been completed. Ward et al. [49] proposed an enhanced recurrent neural network model that obviates the need for lexical annotation. This model delivered promising results in turn-taking prediction for English, Spanish, Japanese, Mandarin Chinese, and French. Ekstedt et al. [11] proposed TurnGPT, a Transformer-based language model that predicts shifts of speakers in spoken language. Their model, trained on various written and spoken dialogue datasets, demonstrated its ability to predict turn-taking by exploiting the context and pragmatic completeness of dialogues.

Audio features, which are also employed in automated speech recognition and speaker identification, are similarly crucial in turn-taking. In turn-hold contexts, speakers maintain an even intonation at the end of the speech; in turn-shift contexts, speakers raise or lower their pitch [7]. Ekstedt et al. explored how prosody is reflected in voice activity prediction, revealing the utilization of various prosodic aspects of speech [10].

Non-verbal features encompass eye gaze and gestures. Speakers look away at the beginning of a turn and shift their gaze toward the listener at the end. Listeners also make eye contact with the speaker for most of the turn, looking away when the turn is over or when it transitions [24]. These patterns sometimes provide vital information for turn-taking events, such as back-channel. In face-to-face interactions, eyebrow movements and mouth openings also function effectively as predictors [27]. Duncan's analysis of turn-taking cues [6] discovered that certain gestures retain a decisive turn and can even override other signals. When speakers use tense hand positions or movements away from the body, listeners rarely attempt to take turns. The relationship between hand gestures and shifts of turns is also evident in other studies [51, 18, 41].

Researchers have recently been experimenting with different turn-taking models in the context of these modalities. Takeuchi et al. proposed a context-dependent response timing model [45] and used decision trees to sequentially decide whether to speak for each analysis frame. Fujie et al. proposed a response timing model that incorporated a first-order delay system and applied it to multi-person conversations [14]. Sakuma et al. proposed an approach that estimates the response timing of a spoken dialogue system by using dialogue act estimation as an auxiliary task [38]. They also used syntactic completion after a specific time, which indicates whether the other party is about to finish speaking [39]. These models have generally addressed various turn-taking problems with separate models for tasks. Turn-taking involves a variety of events; for instance, it should be possible to discern whether the user continues speaking after a brief pause or whether the system responds [42, 13]. It is also important that appropriate moments can be identified at which back-channeling nod and other events can be inserted while the user is speaking [30, 31, 37, 19]. After a user begins to talk, it's also essential to determine whether the utterance is a regular lengthy one or a short listener response (back-channel) [32, 40].

Addressing this situation, Skanzte proposed a unified approach to these turn-taking events in a model that predicts voice activity [43]. Its strength is that it can handle various turn-taking tasks in a generalized manner without separating them. He took prosody as input and predicted a future 3-second voice activity using Long Short-Term Memory (LSTM) [15], a type of recurrent neural network. Roddy et al. [36] also proposed expanded architecture in which separate LSTM subsystems process acoustic and verbal features at different timescales. Ekstedt et al. [12] extended the model to a Transformer and improved its performance by adding innovations to the output window. They also introduced a new evaluation metric: predicting where to transition to different speakers during speech and predicting back-channel locations during speech. They further investigated the impact of prosody on the model and demonstrated that the voice activity prediction model adequately captured prosodic features [10]. However, while the model actively utilizes verbal and audio cues, no method has yet incorporated non-verbal language into a voice activity prediction model.

We investigated how adding non-verbal features affects overall voice activity prediction and improves turn-taking performance from the previously used audio-feature-only model. While non-verbal features have proven effective in models that predict specific turn-taking events [22, 21], this study is novel in that it investigates

their impact on overall voice activity prediction. We incorporated non-verbal cues (gaze direction, action units, head pose, and articular points) into the transformer-based voice activity prediction model and showed that non-verbal features provide essential information for voice activity prediction. Using non-verbal features may move virtual agents closer to human-like communication.

3 TRANSFORMER-BASED VOICE ACTIVITY PREDICTION MODEL

This section describes the voice activity prediction model proposed in previous studies [12, 10]. Fig. 1 shows the flow from voice activity prediction to determine turn-taking events. This model isn't specifically trained for a particular turn-taking event but makes turn-taking decisions based on predictions of future voice activity. With the discrete output window proposed by Ekstedt et al., we establish eight bins, divided into four unequal regions of 0.2, 0.4, 0.6, and 0.8 seconds, which reflect the reality that the further into the future we look, the lower the prediction. We calculate the activity ratio across each bin, and if it exceeds 50%, we regard it as active and create a discrete one-hot representation of size $(2, 4)$. We map the vectors to the indices by treating them as binary numbers. Until now, Ekstedt et al. have treated this model as a projection problem that yields the same output length for the input. However, when implemented in a system such as a virtual agent, we need to predict subsequent voice activity, and so we are moving to a prediction problem that outputs one frame of future voice activity for each input.

As in previous studies, we define four turn-taking events:

- SHIFT/HOLD: predicts the next speaker during mutual silences.
- SHORT/LONG: predicts the length of utterance initiation (short vs. long) during shift events.
- SHIFT-pred: predicts the next speaker during active speech.
- Back-channel (BC) -pred: predicts future back-channels.

We evaluated these metrics through zero-shot classification. These implementations are based on source code [9] posted by previous studies.

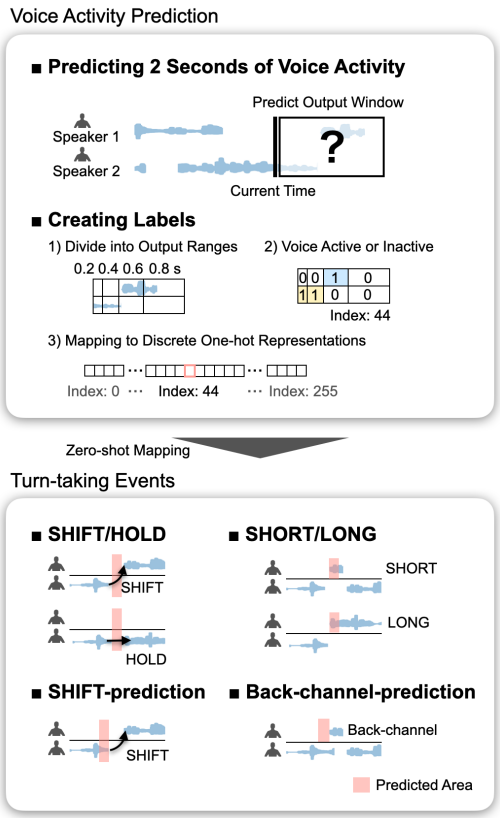

Figure 1: Mapping of turn-taking events from voice activity prediction: Future voice activity is predicted in a non-uniform output window and maps vector to an index. Output classifies and evaluates each turn-taking event with a zero shot.

4 PROPOSED MULTIMODAL MODEL

In this section, we present our proposed multimodal model with non-verbal features as input and extend the Transformer-based voice activity prediction model of Section 3 to include non-verbal cues in addition to the audio features used in previous studies. We describe the architecture of the proposed multimodal model in Section 4.1, the audio features in Section 4.2, and a method for extracting non-verbal features in Section 4.3.

4.1 Model Architecture

We created a multimodal voice activity prediction model (Fig. 2 that uses audio features (audio waveform, voice activity frame, and voice activity history) and non-verbal features (gaze direction, action units, head pose, and articular points). It consists of two conditions: (1) an audio condition, which obtains an embedded representation from audio waveforms and voice activity-related, and (2) a non-verbal condition, which processes non-verbal cues to obtain embedded expressions. Both are described below.

(1) Audio condition The audio condition handles audio waveforms, the voice activity frame, and the voice activity history. As in preceding studies, we used Contrastive Predictive Coding (CPC) [33], a pre-trained model, to derive a sequence representation of the audio waveform. The CPC output is a 100-Hz, 256-dimensional frame representation. But since other features are at 25-Hz, a 1D-convolutional layer that obtains a 25-Hz sequence representation is denoted as $WC \in \mathbf{R}^{256}$. The voice activity frame and history are input into a linear layer to obtain a 256-dimensional sequence representation. We combined these representations to derive sequence representation $VC \in \mathbf{R}^{256}$. The voice activity condition is added to the production of the audio waveform condition, and this sum is processed through the Transformer block to obtain the audio representation, denoted as $AC = (WC + VC) \in \mathbf{R}^{256}$.

(2) Non-verbal condition We employed gaze direction, action units, head pose, and articular points as non-verbal features and combined the data of the two people in the same way as audio waveforms. Each modality is input to a linear layer to obtain a

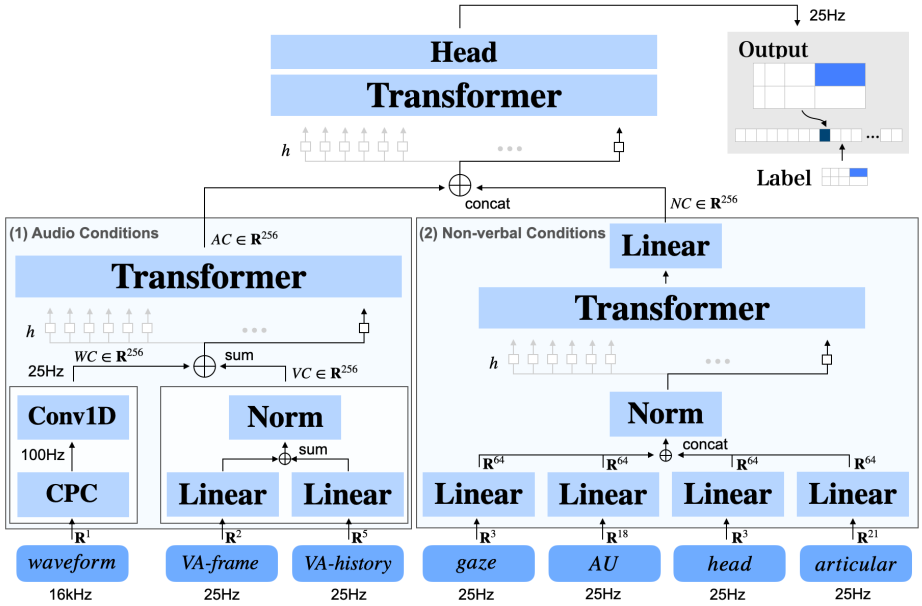

Figure 2: Architecture of proposed multimodal model divided into two parts: (1) processing audio features and (2) processing non-verbal features, which they input to Transformer block to obtain output.

64-dimensional series representation. We concatenated these representations and input them into a Transformer block to derive a non-verbal expression, denoted as $NC \in \mathbf{R}^{256}$.

The obtained audio AC and non-verbal NC representations are concatenated and input into the transformer block, processed in the final linear layer, where the logit is output. The logit outputs a discrete series of voice activity over the next 2 seconds, as in previous studies [12].

4.2 Audio Features

Previous studies employed audio features (audio waveform, voice activity frame, and voice activity history) as predictors [12]. We also adopted them and describe the extraction and input methods below.

Audio waveform The speaker 1 and speaker 2 audio are mixed and treated as a single audio channel. Apart from voice normalization, we did not perform any special preprocessing here because we directly processed the raw waveforms and extracted features from the model.

Voice activity frame We represent the voice activity frame as a 25-Hz frame vector $VA_f(t) \in \{0,1\}^2$, where 1 indicates the interval with voice activity, and 0 denotes it without. Given that the video data, from which we extracted the non-verbal features, are at 25 fps, we also set the voice activity frame at this frequency. Since the input consists of mixed audio waveforms, it plays a crucial role in differentiating between speakers.

Voice activity history We represent the voice activity history as vector $VA_h(t) \in \mathbf{R}^5$, indicating the voice activity's ratio between speakers over a specific time range in the past (-inf:60, 60:30, 30:10, 10:5, 5:0 seconds). Eq. 1 describes the proportion of voice activity in each interval:

$$VA_{history}(section) = \frac{VoiceActiveDuration_{speaker1}}{VoiceActiveDuration_{speaker2}}. \quad (1)$$

Generally, we expect longer speaker-turn retention times to increase the probability of turn-shifting. The voice activity history provides extensive contextual information outside the receptive field of such acoustic models. However, previous research has not discussed how voice activity history contributes to learning.

4.3 Non-verbal Features

This section describes the non-verbal modalities used for learning. Although we are focusing on non-verbal features, it is easy to imagine that they do not function independently but rather complement verbal and audio elements. Therefore, based on the audio features from previous study [12], we extract non-verbal features using OpenFace [1] and OpenPose [3] based on the following non-verbal features: gaze direction, action units, head pose, and articular points. We describe each extraction method below.

Gaze direction We extracted the gaze direction using OpenFace and obtained the average (x, y) coordinates of the right and left eyes in the radians from the video data provided in full HD. We represented the detection accuracy, which depends on the quality of the video and the head's orientation, as $NC_{gaze}(t) \in \mathbf{R}^3$ with an additional confidence value. Eye gaze significantly impacts turn-taking. However, the extent to which it works in a frontal shot dialogue has yet to be determined. Given that many agent systems engage in front-facing dyadic dialogues as well as the presence of webcams and similar devices on monitors, this study's results might provide helpful information.

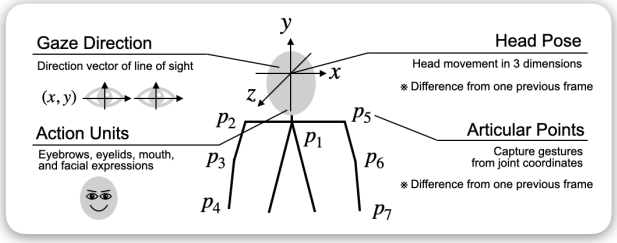

Figure 3: Non-verbal features, including gaze direction, action units, head pose, and articular points

Action units We also extracted action units using OpenFace. These values, which depict facial expressions and movements, are represented by $NC_{au}(t) \in \mathbf{R}^{18}$ with inputs of 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 45, and confidence values. Since no research has employed action units for turn-taking prediction, our study offers new insights. As previously mentioned, lip and eyebrow movements correlate with turn-taking and show promise as an effective source of information for voice activity prediction. We also expect to capture specific facial expressions, such as smiling actions, before back-channel occurrences.

Head pose We also extracted head poses using OpenFace. Head pose detection can capture such behaviors as nodding, which we expect to correlate with shifts of turns and back-channel. We provided the head coordinates (x, y, z), and the difference from one previous frame is extracted as a feature using Eq. (2), represented by $NC_{head}(t) \in \mathbf{R}^3$:

$$head_{x,y,z}(t) = \sqrt{((x, y, z)_t - (x, y, z)_{t-1})^2}. \quad (2)$$

Articular points We extracted the articular points using OpenPose. Considering the inference speed, the video was resized from full HD to 320 x 180 pixels. The coordinates (x, y) of each joint can be obtained, along with the confidence value of each joint point, represented by $NC_{articular}(t) \in \mathbf{R}^{21}$. We used the joint points from No. 1 to No. 7 (Fig. 3), and the difference from one previous frame was extracted as a feature using the following formula (3):

$$Articular^n(t) = \sqrt{\left(Articular^n_t - Articular^n_{t-1}\right)^2}, \quad (3)$$

where n is the number of the indirect points. Researchers have already demonstrated that gestures influence turn-taking and enhance the shifts of turn accuracy and other operations.

5 EXPERIMENTAL EVALUATION

In this section, we investigate how non-verbal features affect the prediction of voice activity through experiments. We describe the data used in Section 5.1, a plan for investigating the impact of non-verbal cues in Section 5.2, and the results in Section 5.3.

5.1 Data

For the model training, we used the NoXi Database [2] and show its recording in Fig. 4. This database is comprised of audio and video recordings of an expert and a novice in separate rooms who are interacting through a screen and actively discussing a single topic. The situation resembles a virtual agent system, where a camera mounted on a monitor captures the interaction through the screen, with the camera positioned in front of the monitor. Thus, eye movements and gestures are also represented similarly to such environments. This corpus was recorded in such different regions as France, Germany, and the UK, with 87 participants in multiple languages, including English, French, and German. We used all these files for our study because we focused on non-verbal rather than verbal features. Since the NoXi database does not contain annotations for voice segments, we annotated them manually by two annotators.

As shown in Table 1, we split this corpus into 16.8 hours for training data, 3.4 for validation data, and 5.3 for test data. The sample size of these data is sufficient, since Skantze et al. [43] used the HCRC Map Task corpus with 10.7 hours of training data and 3.6 hours of test data. We also confirmed that the number of turn-taking events in the test data was sufficient. However, there was a bias in the number of samples for the SHIFT/HOLD and SHORT/LONG events. This bias is a natural outcome, considering the high speaker retention and frequent short nods in actual dialogues. We evenly distributed the test data to files recorded in each region to minimize the bias caused by regional differences in non-verbal features and other factors.

Figure 4: Dialogue between an expert and a novice: Recording shows them in separate rooms, discussing a single topic with counterparts projected on a screen.

5.2 Procesure

We used a multimodal voice activity prediction model to investigate how non-verbal features affect audio features. Generally, non-verbal cues provide helpful information in voice activity at the turn-shift and back-channel points. Therefore, multimodal models may provide more accurate predictions at turn-taking event points than audio feature-only models. We studied the effects of non-verbal features by learning under three main conditions. (1) Only Audio Features: We just learned with audio features, showing the roles of audio waveforms, voice activity frames, and voice activity history. The voice activity frame provides helpful information for speaker identification concerning the mixed audio waveform, and the voice activity history includes information about turn-shifting. We verified that combining these three audio features yields a sufficient baseline. (2) Only Non-verbal Features: We trained using only non-verbal features and verified that gaze direction, action units, head pose, and articular points have information about turn-taking. We expect that non-verbal features will capture information for such

Table 1: Dataset used for training: We used NoXi database from the dyad corpus, whose audio and video files were split into training, validation, and test data.

Datasets	Number of Sessions	Duration [h]	SHIFT/HOLD	SHORT/LONG	SHIFT-pred	BC-pred
Train	54	16.87	1015/11757	3382/1015	1015	3375
Validation	12	3.35	246/2447	595/246	246	592
Test	18	5.29	354/3659	1007/354	354	1005

potent modalities as audio features, and researchers have yet to train such features to be used independently. We are interested in the extent to which non-verbal features are informative for predicting voice activity. (3) Audio + Non-verbal Features: We tested how much audio and non-verbal features improve scores relative to a model with only audio features. We checked the contribution made by each modality by adding one modality at a time and visualized a model's output with only audio features and another with added non-verbal features to see how the non-verbal features acted.

We trained with a 10-second input window, sliding by 0.5 seconds. All the predictors employed a transformed model consisting of a causal encoder with a hidden layer size 256, 2 layers, 4 heads, and a dropout rate of 0.1. We trained our model with a checkpoint of 0.2 and an early stopping criterion of 7 epochs. We used AdamW with a learning rate of 3.63e-4 and a batch size of 128 for optimization. We varied the seed from 0 to 9, trained the model 10 times, and evaluated it with its average. The model and training were implemented in Python using the Py-Torch [34] libraries. We created our model based on codes from previous studies [8] and and are available online[1].

5.3 Results

Table 2 shows the training results for each modality selection. We show cross-entropy loss, a measure of voice activity prediction accuracy, and F1 scores for various measures of turn-taking ability. The numbers 2 in parentheses indicate the standard deviations. We implemented cross-entropy loss using PyTorch's TORCH.NN.FUNCTIONAL.CROSS_ENTROPY function and used the weighted F1 scores for SHIFT/HOLD and negative sampling for SHIFT-pred and BC-pred to ensure equal numbers of positive and negative tasks. The table's top row shows a case that selected only audio features, the middle row shows a case that selected only non-verbal features, and the bottom row shows a case that picked both audio and non-verbal features. The value with the best score in each section in highlighted in bold.

(1) Audio Features Compared to using only audio waveforms, incorporating the voice activity frames increased the scores for all the evaluation metrics. Due to the reduced cross-entropy loss, the voice activity frame facilitated speaker identification in the mixed speech and improved the prediction accuracy of voice activity. The voice activity history significantly improved the SHIFT-pred scores by +2.1% compared to those without a voice activity history, confirmed by a two-tailed T-test($p < 0.025$).

(2) Non-verbal Features Surprisingly, even with only nonverbal features as input, voice activity prediction achieved nearly 70% accuracy in turn-shift prediction and almost 60% in back-channel prediction. This result exceeded the results obtained using only audio waveforms, suggesting that non-verbal features significantly impact turn-taking. However, the cross-entropy of the non-verbal feature-only input exceeded that of the audio waveform-only information. Furthermore, when analyzing the breakdown of non-verbal cues, we found marked differences in back-channel prediction scores; they were higher for gaze direction and action units, with action units standing out.

(3) Audio + Non-verbal Features When individually adding non-verbal features to the audio features, we obtained a significant difference in SHIFT-pred for adding gaze direction and action units ($p < 0.025$). In addition, we observed a significant solid difference ($p < 0.025$) in BC-pred when action units were added. On the other hand, we did not obtain any meaningful differences between SHIFT/HOLD and SHORT/LONG. Table 3 shows the differences between the model with audio features as input and all the proposed non-verbal combinations. The results showed a +2.3% improvement for SHIFT-pred and +1.8% for BC-pred, confirming a significant solid difference ($p < 0.025$). The number of parameters for the baseline model was 5,259,515 (18.15MB), and the number of parameters for the proposed model was 10,642,543 (39.69MB). Also, the inference time for one frame when using NVIDIA A100 NVA100-80G was 1.15E-02 seconds and 1.57E-02 seconds, respectively.

We visualized the output in Fig. 5 to present the changes. The top graph shows the training model with only the audio features added, and the bottom chart shows it with non-verbal features added. The light blue intervals show the expert's speech segments, and the pale yellow intervals show those of the novice. The blue line represents the expert's speech turn probability; the red line is the novice's speech turn probability. A user with a higher chance is more likely to take a speech turn. The horizontal axis is a time series, and the unit is seconds. We classified the intervals from (1) to (4) in the areas of particular variation in the graph. In interval (1), the audio feature model predicts a high probability that an expert's speech will turn immediately after the novice's speech ends; our multimodal model has a gentle probability curve. In interval (2), the prior research model shows a high probability that a speech turn will occur; our model shows a turn-hold. We can see the model's high accuracy, considering the short interval of the expert's speech and classifying it as a novice speech turn. The gaps in interval (3) show breaks where the expert and novice speak to each other. Ideally, the expert and novice speech turn probabilities should be 0.5 each; our model achieves this level. Interval (4) is the section that identified where fast speech turns from novice to expert. Our model immediately shows that the speech turn has shifted, whereas the feature-only model is somewhat unstable and slow to respond.

[1]https://github.com/ahclab/turntaking

Table 2: Training results with each modality selection: We showed cross-entropy loss and F1 score for turn-taking performance. Bolded letters are best scores for each item. Numbers in parentheses indicate standard deviation (SD). Light gray row is baseline model, and dark gray row is proposed model.

Audio Features			Non-verbal Features				Model Performance	Turn-taking Performance (F1 score)			
Waveform	VA-frame	VA-history	Gaze	AU	Head	Articular	Cross Entropy Loss	SHIFT/HOLD	SHORT/LONG	SHIFT-pred.	BC-pred.
✓							3.145 (0.010)	0.734 (0.123)	0.629 (0.008)	0.657 (0.033)	0.650 (0.016)
✓	✓						**2.452** (0.011)	0.885 (0.008)	0.804 (0.022)	0.704 (0.022)	**0.691** (0.010)
✓	✓	✓					2.460 (0.009)	**0.888** (0.006)	**0.811** (0.012)	**0.725** (0.011)	0.687 (0.013)
			✓				3.421 (0.011)	0.814 (0.128)	0.625 (0.035)	0.656 (0.102)	0.530 (0.014)
				✓			3.409 (0.009)	0.837 (0.026)	**0.649** (0.015)	0.676 (0.029)	0.591 (0.013)
					✓		3.435 (0.014)	**0.856** (0.000)	0.643 (0.026)	**0.697** (0.007)	0.428 (0.061)
						✓	3.438 (0.013)	**0.856** (0.001)	0.631 (0.032)	0.692 (0.012)	0.467 (0.053)
			✓	✓	✓	✓	**3.394** (0.012)	0.825 (0.038)	0.638 (0.014)	0.674 (0.023)	**0.598** (0.015)
✓	✓	✓	✓				2.450 (0.010)	0.889 (0.005)	0.810 (0.014)	0.738 (0.007)	0.688 (0.008)
✓	✓	✓		✓			**2.437** (0.005)	**0.897** (0.010)	0.812 (0.010)	0.744 (0.013)	**0.707** (0.010)
✓	✓	✓			✓		2.448 (0.007)	0.892 (0.008)	0.814 (0.007)	0.739 (0.013)	0.693 (0.009)
✓	✓	✓				✓	2.450 (0.010)	0.891 (0.006)	**0.816** (0.018)	0.735 (0.014)	0.691 (0.010)
✓	✓	✓	✓	✓	✓	✓	2.449 (0.006)	0.890 (0.009)	0.812 (0.013)	**0.748** (0.012)	0.705 (0.005)

Table 3: Effect size and significant differences for audio features plus non-verbal features: SHIFT-pred and BC-pred confirmed large significant differences.

	Audio Features, mean of F1 score (SD)	Audio + None-verbal Features, mean of F1 score (SD)	Cohen's *d*	*P* value
SHIFT/HOLD	0.888 (0.006)	0.890 (0.009)	0.144	0.782
SHORT/LONG	0.811(0.012)	0.810 (0.013)	0.035	0.950
SHIFT-pred.	0.725 (0.011)	0.748 (0.012)	1.985	0.000
BC-pred.	0.687 (0.013)	0.705 (0.005)	1.900	0.001

6 DISCUSSION

We examined the roles of the voice activity frame and the voice activity history, which previous studies did not test. Incorporating voice activity frames reduced the cross-entropy loss compared to using only audio waveforms. This result suggests that the voice activity frame facilitates speaker identification in mixed audio and improves the prediction accuracy of voice activity. Furthermore, introducing the voice activity history improved the SHIFT-pred scores, suggesting that it provides valuable information in the turn-shift context. This result supports the notion that voice activity history provides broad contextual information beyond the receptive range of acoustic models.

Next we addressed turn-taking prediction using only non-verbal features as input. Even with non-verbal features only, we achieved nearly 70% accuracy in turn-shift prediction and almost 60% in back-channel prediction. On the other hand, the cross-entropy loss

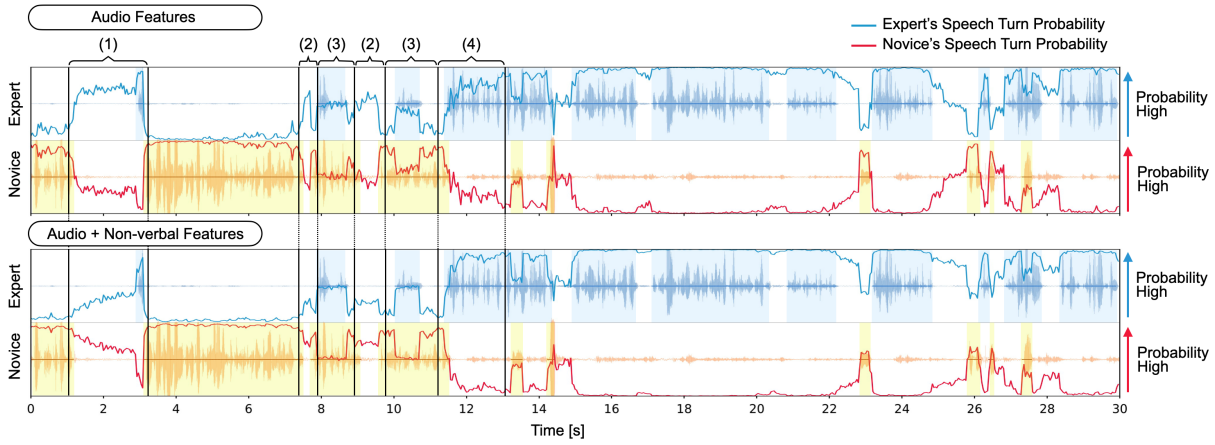

Figure 5: Visualization of turn-taking prediction: Blue line is probability of an expert speech turn, and red line is probability of a novice speech turn. Interval (1) represents a long silent interval, interval (2) represents turn holding, interval (3) represents two people speaking simultaneously, and interval (4) represents a quick turn change.

itself, which represents voice activity prediction, is high. This result does not indicate that non-verbal features contribute highly to voice activity prediction but provide essential information about turn-taking events. Looking at the non-verbal modalities individually, the back-channel scores are notably higher for action units, which give brow and eyelid movements, facial muscles, and lip movements, suggesting that a human might simultaneously smile or blink when nodding, for example.

We added non-verbal features to the audio features and obtained improvements with SHIFT-pred and BC-pred. In particular, when we added gaze direction and action units, a primary and significant difference was obtained for SHIFT-pred. In BC-pred, the addition of action units showed a tremendous difference. This result is consistent with the significantly higher back-channel scores observed when training exclusively with action units, proving that eyebrow and eyelid movements, facial muscles, and lip movements accurately predict back-channel activity. The action units, which were significantly different in both SHIFT-pred. and BC-pred. could be a very important modality. On the other hand, head poses and articular points showed no improvement in turn-taking events, a surprising result because turn-taking is related to head movements and gestures. One possible reason is that the accuracy of the three-dimensional coordinates of the head pose by OpenFace is relatively low compared to the two-dimensional gaze vector, which absorbs the information of the head-lowering motion. Previous studies also observed the relationship between gesture and turn-taking, although we did not find that joint points functioned as features. Since the training data consisted of multiple regions and languages, perhaps unique characteristics were not extractable due to significant cultural and individual differences. In addition, we did not obtain any dominance differences for SHIFT/HOLD and SHORT/LONG in any of the modalities. This result may reflect event bias, which is not surprising given that the sample size for turn-hold is much larger than for turn-shift and more challenging to predict than turn-hold. Perhaps there were also more short turns (back-channel) than long turns, creating a data imbalance that prevented differences. Another factor could be the high performance of the baseline model, where SHIFT/HOLD was at nearly 90% and SHORT/LONG was over 80%.

The question arises whether the models trained on our human-human data can also be used to predict turn-taking in human-computer dialogues. Human-human and human-computer interactions generally look very different, and human-human, multi-turn-taking behavior is not necessarily a role model of the behavior we want from the system. Skantze tested this issue with a human-robot spoken dialogue used for evaluation [43]. We proposed a logistic regression model trained on turn-shifted and turn-hold human-robot conversations using the hidden nodes of the voice activity prediction model as input and found that it achieves very high scores. Thus, the voice activity prediction model we employed will likely show valid results in human-computer dialogues.

7 CONCLUSION

We proposed a Transformer-based multimodal voice activity prediction model for turn-taking. Our results show that non-verbal features are also crucial in voice activity prediction models, particularly action units, which are essential for understanding speaker turns and back-channels. These results offer a fresh perspective for understanding and modeling turn-taking and provide beneficial insights for the developers of virtual agents. Our future work will investigate which numbers of action units are most effective for forecasting. We will also include a new Japanese version of the NoXi database to investigate the effects of cultural and individual differences and seek room for improvement by fine-tuning only in specific regions. We also plan to test this model's effectiveness in a natural virtual agent system.

ACKNOWLEDGMENTS

We would like to thank ANR-CREST-TAPAS Japan-France project. This work was funded by the JST CREST Grant Number JPMJCR19M5.

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[2] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 350–359.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43, 1, 172–186.

[4] Herbert H Clark. 1996. *Using language*. Cambridge university press.

[5] Iwan de Kok and Dirk K. J. Heylen. 2009. Multimodal end-of-turn prediction in multi-party meetings. In *ICMI-MLMI '09*.

[6] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23, 2, 283.

[7] Jens Edlund and Mattias Heldner. 2005. Exploring prosody in interaction control. *Phonetica*, 62, 2-4, 215–226.

[8] Erik Ekstedt. 2022. Continuous conversational ssl. https://github.com/erikekstedt/conv_ssl. (2022).

[9] Erik Ekstedt. 2022. Vap: voice activity projection. https://github.com/ErikEkstedt/vap_turn_taking. (2022).

[10] Erik Ekstedt and Gabriel Skantze. 2022. How much does prosody help turn-taking? investigations using voice activity projection models. *arXiv preprint arXiv:2209.05161*.

[11] Erik Ekstedt and Gabriel Skantze. 2020. Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874*.

[12] Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: self-supervised learning of turn-taking events. *arXiv preprint arXiv:2205.09812*.

[13] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2002. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Seventh international conference on spoken language processing*.

[14] Shinya Fujie, Hayato Katayama, Jin Sakuma, and Tetsunori Kobayashi. 2021. Timing generating networks: neural network based precise turn-taking timing prediction in multiparty conversation. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. International Speech Communication Association, 3771–3775.

[15] Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37–45.

[16] Nishitha Guntakandla and Rodney D. Nielsen. 2015. Modelling turn-taking in human conversations. In *AAAI Spring Symposia*.

[17] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Turn-taking prediction based on detection of transition relevance place. In *INTERSPEECH*, 4170–4174.

[18] Judith Holler, Kobin H Kendrick, and Stephen C Levinson. 2018. Processing language in face-to-face conversation: questions with gestures get faster responses. *Psychonomic bulletin & review*, 25, 1900–1908.

[19] Nusrah Hussain, Engin Erzin, T Metin Sezgin, and Yucel Yemez. 2019. Speech driven backchannel generation using deep q-network for enhancing engagement in human-robot interaction. *arXiv preprint arXiv:1908.01618*.

[20] Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. 2013. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, 181–188.

[21] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. 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*, 131–138.

[22] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2022. Trimodal prediction of speaking and listening willingness to help improve turn-changing modeling. *Frontiers in Psychology*, 13, 774547.

[23] Toshihiko Itoh, Norihide Kitaoka, and Ryota Nishimura. 2009. Subjective experiments on influence of response timing in spoken dialogues. In *Interspeech*.

[24] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26, 22–63.

[25] Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B*, 378, 1875, 20210473.

[26] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*, 226–234.

[27] Chi-Chun Lee and Shrikanth S. Narayanan. 2010. Predicting interruptions in dyadic spoken interactions. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5250–5253.

[28] Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6, 731.

[29] Tomer Meshorer and Peter A. Heeman. 2016. Using past speaker behavior to better predict turn transitions. In *Interspeech*.

[30] Louis-Philippe Morency, Iwan De Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: a probabilistic multimodal approach. In *Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008. Proceedings 8*. Springer, 176–190.

[31] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using neural networks for data-driven backchannel prediction: a survey on input features and training techniques. In *Human-Computer Interaction: Interaction Technologies: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II 17*. Springer, 329–340.

[32] Daniel Neiberg and Khiet P. Truong. 2011. Online detection of vocal listener responses with maximum latency constraints. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5836–5839.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[34] Adam Paszke et al. 2019. Pytorch: an imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

[35] Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskénazi. 2006. Doing research on a deployed spoken dialogue system: one year of let's go! experience. In *Interspeech*.

[36] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 186–190.

[37] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*. Springer, 247–258.

[38] Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2022. Response Timing Estimation for Spoken Dialog System using Dialog Act Estimation. In *Proc. Interspeech 2022*, 4486–4490. DOI: 10.21437/Interspeech.2022-746.

[39] Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2023. Response timing estimation for spoken dialog systems based on syntactic completeness prediction. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 369–374.

[40] Ethan Selfridge, Iker Arizmendi, Peter A Heeman, and Jason D Williams. 2013. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference*, 384–393.

[41] Rein Ove Sikveland and Richard Ogden. 2012. Holding gestures across turns : moments to generate shared understanding. *Gesture*, 12, 166–199.

[42] Gabriel Skantze. 2012. A testbed for examining the timing of feedback using a map task. In.

[43] Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 220–230.

[44] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67, 101178.

[45] Masashi Takeuchi, Norihide Kitaoka, and Seiichi Nakagawa. 2003. Generation of natural response timing using decision tree based on prosodic and linguistic information. In *Eighth European Conference on Speech Communication and Technology*.

[46] Mark Ter Maat, Khiet P Truong, and Dirk Heylen. 2010. How turn-taking strategies influence users' impressions of an agent. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 441–453.

[47] Bekir Berker Türker, Zana Buçinca, Engin Erzin, Yücel Yemez, and T Metin Sezgin. 2017. Analysis of engagement and user experience with a laughter responsive social robot. In *Interspeech*, 844–848.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[49] Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 831–837.

[50] Nigel G. Ward, Anais G. Rivera, Karen Ward, and David G. Novick. 2005. Root causes of lost time and user stress in a simple dialog system. In *Interspeech*.

[51] Margaret Zellers, David House, and Simon Alexanderson. 2016. Prosody and hand gesture at turn boundaries in swedish. *Proc. Speech Prosody 2016*, 831–835.

[52] Ran Zhao, Oscar J Romero, and Alex Rudnicky. 2018. Sogo: a social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on intelligent virtual agents*, 239–246.