Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction

Sam O'Connor Russell and **Naomi Harte**
ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
{russelsa,nharte}@tcd.ie

Abstract

Turn-taking is richly multimodal. Predictive turn-taking models (PTTMs) facilitate naturalistic human-robot interaction, yet most rely solely on speech. We introduce MM-VAP, a multimodal PTTM which combines speech with visual cues including facial expression, head pose and gaze. We find that it outperforms the state-of-the-art audio-only in videoconferencing interactions (84% vs. 79% hold/shift prediction accuracy). Unlike prior work which aggregates all holds and shifts, we group by duration of silence between turns. This reveals that through the inclusion of visual features, MM-VAP outperforms a state-of-the-art audio-only turn-taking model across all durations of speaker transitions. We conduct a detailed ablation study, which reveals that facial expression features contribute the most to model performance. Thus, our working hypothesis is that when interlocutors can see one another, visual cues are vital for turn-taking and must therefore be included for accurate turn-taking prediction. We additionally validate the suitability of automatic speech alignment for PTTM training using telephone speech. This work represents the first comprehensive analysis of multimodal PTTMs. We discuss implications for future work and make all code publicly available.

1 Introduction

There is 200 ms of silence on average between speaking turns in a two-party interaction (Stivers et al., 2009; Levinson and Torreira, 2015), yet language formation takes at least 600 ms (Indefrey, 2011). Turn-taking is therefore *predictive*: listeners plan their next turn while the speaker is still speaking (Levinson and Torreira, 2015; Garrod and Pickering, 2015). Multimodal cues including syntax, prosody and gaze support this process (Holler et al., 2016), enabling speakers to *hold* the floor or to *shift* to another speaker (Skantze, 2021).

An important goal of human-robot interaction (HRI) is to develop machines capable of real-time conversation (Marge et al., 2022). Turn-taking models are central to this objective. Current turn-taking models in consumer systems are *reactive*, deciding whether to speak after a human stops speaking. This results in dialogue that is less spontaneous than human interaction (Li et al., 2022; Woodruff and Aoki, 2003). Predictive turn-taking models (PTTMs) aim to overcome these issues (Skantze, 2021). Inspired by human turn-taking, PTTMs are neural networks trained to continually make turn-taking predictions, e.g. the probability of an upcoming shift (Skantze, 2017). Most work on PTTMs has been conducted using corpora of two-party human interaction (Skantze, 2021).

Almost all PTTMs rely on speech to make predictions; visual information, such as facial expression, is ignored. This may suffice for PTTMs trained on telephone speech (Skantze, 2017; Li et al., 2022), but when interlocutors can see one another, might visual cues be useful predictors? The psycholinguistics literature strongly suggests so. Barkhuysen et al. (2008) cropped short segments from recordings of questions and asked participants if they belonged to the middle or the end of the question. They were more accurate when presented with both audio and video than in audio-only and video-only conditions. In a recent study Nota et al. (2023) found that listeners were faster at recognising questions containing eyebrow frowns. Such studies underline the importance of visual cues in turn-taking and multimodal interaction more generally (Holler and Levinson, 2019).

Aims Despite the essential role visual cues play in turn-taking, state-of-the-art PTTMs rely only on speech. It is therefore unknown whether visual cues can improve PTTM performance. We therefore consider the following research questions:

1. Can visual cues improve the performance of predictive turn-taking models?

Findings of the Association for Computational Linguistics: ACL 2025, pages 209–221

2. If so, which aspects of turn-taking benefit most from the inclusion of visual cues?

Overview We introduce multimodal VAP (MM-VAP), a transformer-based PTTM which combines speech with visual features, including facial expression. We show our model outperforms the voice activity projection (VAP) model, a state-of-the-art audio-only PTTM (Ekstedt and Skantze, 2022b).

PTTMs are typically trained using accurate phonetic alignments (Ekstedt and Skantze, 2022b; Li et al., 2022). We use automatic speech recognition (ASR) to reflect real-word conditions. We re-implement and re-train the VAP model on the Switchboard corpus of telephone speech (Godfrey et al., 1992). We find performance falls slightly due to automatic alignment errors, but the impact is minimal. Next, we train the VAP model using audio from the Candor corpus of videoconferencing interaction (Reece et al., 2023), again using ASR. We find good performance in distinguishing *shifts* from *holds*. We conduct a facial action unit analysis, which reveals that the next speaker displays enhanced mouth, lip, jaw and chin movements before speaking (i.e. before a shift). These are visual cues which multimodal PTTM could exploit. We therefore combine speech with facial expression, gaze, and head pose in our new multimodal PTTM and demonstrate superior performance over the VAP model (79% vs 83% balanced accuracy for hold/shift prediction). The performance increase is most notable in the F_1 score for shifts, with a 6-10% increase in the F_1 score.

In prior work, PTTM performance is reported as a single figure for all holds and shifts. In a more comprehensive analysis, we group holds and shifts by duration of silence between turns. This reveals the strength of our multimodal model across all durations of silence between turns (gaps) and overlapping speech. The performance of both MM-VAP and VAP is worse when there are longer gaps in the Candor corpus, however, MM-VAP outperforms the audio-only model across all durations of gaps considered (83% vs 79% for all gaps >0 ms, 78% vs 75% for all gaps >750 ms). We discuss our working hypothesis that interlocutors utilise visual cues *when they can see one another*, and they must therefore be included to maintain robust performance.

Finally, we conduct a detailed ablation study. This reveals that facial action units, which encode facial expression, are the biggest contributors to the increased accuracy over the audio-only turn-taking model.

We conclude with a discussion of our findings, highlighting the vital importance of non-verbal communication, which researchers in human-robot interaction must consider going forward. We make our code publicly available for future research[1].

2 Background

Turn-taking In a conversation, the current speaker either *holds* the turn or *shifts* to another interlocutor (Sacks et al., 1974). The time between turns is the floor-transfer offset (FTO), which is positive for a gap and negative for an overlap (Heldner and Edlund, 2010). A typical two-party interaction has a mean FTO of +200 ms (Stivers et al., 2009). There is no single definition of a turn. In Conversation Analysis, a turn is defined by social action (a question, agreement, etc.) (Kasper and Wagner, 2014). In predictive turn-taking, a turn is identified from voice activity (VA), which indicates the presence or absence of speech at any moment in time (Ekstedt and Skantze, 2022b; Li et al., 2022).

Turn-taking is aided by multimodal cues. The words that we speak as well as how we speak (prosody) play important roles. Bögels and Torreira (2015) found prosody is essential for listeners to correctly determine the end of the turns with multiple completion points e.g. "*are you a student* \ *here* \ *at this university* \.". Visual cues include gaze. Speakers look away at the start and back toward the listener at the end of a turn (Kendon, 1967). Certain gestures and facial expressions are associated with faster shifts between speakers (Trujillo et al., 2021; Nota et al., 2023).

Predictive turn-taking PTTMs continually predict upcoming speaker changes (Skantze, 2021). PTTMs are either RNN (Skantze, 2017; Li et al., 2022) or transformer-based (Ekstedt and Skantze, 2022b). They can be trained to predict the VA in the next 2 seconds (Ekstedt and Skantze, 2022b), enabling turn-taking predictions to be made, e.g. predict a shift if the VA probability of the listener exceeds that of the speaker. The start time of the next turn can also be predicted (Li et al., 2022). PTTMs are trained on corpora of human-human interaction and may be deployed to human-robot interaction later. Corpora used to train PTTMs include Switchboard (telephone speech; Godfrey

[1] https://github.com/russelsa/mm-vap

et al. (1992); Li et al. (2022); Ekstedt and Skantze (2022b)), MapTask (in-person, video unavailable; Anderson et al. (1991); Roddy et al. (2018b,a)) and the Mahnob corpus (in-person, with video recordings; Bilakhia et al. (2015); Roddy et al. (2018b)). Earlier PTTMs use engineered features e.g. part-of-speech tags and the GeMAPs acoustic feature set (Skantze, 2017; Roddy et al., 2018a), though feature engineering has now been replaced by pre-trained feature extractors (Ekstedt and Skantze, 2022b).

Multimodal predictive turn-taking Almost all PTTMs for two-party interaction rely on speech alone, though a limited number of multimodal models have been reported. Roddy et al. (2018b) demonstrated that incorporating gaze vectors alongside speech in a PTTM resulted in a performance improvement on the 11 hour Mahnob corpus (Bilakhia et al., 2015) (0.86 vs 0.85 F_1 hold/shift prediction). Kurata et al. (2023) found visual features improved the classification of 5 second segments of speech located at the end of turns into hold and shift categories. As the end of the turn must already be known for the model to run, it is therefore not a PTTM, although the findings are promising.

Onishi et al. (Onishi et al., 2024) proposed to extend a recent state-of-the-art audio-only turn-taking model, the VAP model (Ekstedt and Skantze, 2022a), to include visual features. Like Roddy et al., they found that the inclusion of visual cues boosted performance however, the model was only tested on between 1.5-2.0 hours of data in 4 languages. As the audio of two speakers was down-mixed into a mono audio channel, ground-truth knowledge of the current speaker (the voice activity) is required for inference (Onishi et al., 2024). The model therefore 'knows' exactly when each speaker change occurs.

The inclusion of the visual modality is therefore under-explored, and it is unclear if visual information can enhance performance at scale. Though beyond the scope of this work, visual features have been considered in PTTMs for multiparty interaction (Malik et al., 2020).

The use of manual alignments The use of time-aligned transcriptions to extract the voice activity for training and testing is widespread in the PTTM literature (Roddy et al., 2018a; Ekstedt and Skantze, 2022b; Onishi et al., 2024; Inoue et al., 2024). A manual approach limits the amount of data that can be used for training and testing. Automatic speech recognition (ASR) is a promising tool for conversational speech transcription (Russell et al., 2024). However, it introduces errors, particularly on aspects of conversational speech such as filled pauses and disfluencies (Russell et al., 2024). To the best of our knowledge, the use of ASR transcribed interaction for training and testing PTTMs has not been considered in the literature to date, despite more closely mirroring a real-world deployment and enabling much larger quantities of data to be used for training and testing.

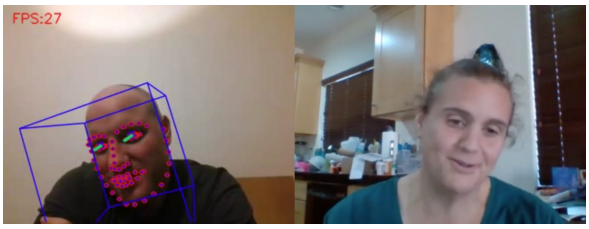

Figure 1: Still from Candor with OpenFace features shown for the left participant

3 Methodology

Corpora We use the Candor corpus (Reece et al., 2023) of 1,656 two-party videoconferencing (VC) interactions to train and validate PTTMs. Each interaction consists of casual conversation in US English e.g. sports teams (mean session length 34 mn). Its large size is ideal for deep learning. Although not in-person, interlocutors can see one another. We are unaware of a suitable in-person corpus of this size. VC provides a front-facing camera angle that is ideal for visual feature extraction (Figure 1). We use a 710 hour subset of the full 850 hour corpus where visual feature extraction is optimal, detailed further on. We obtain a time-aligned transcription using Speechmatics, Ltd. (2024) ASR previously validated on VC speech (Russell et al., 2024). We also use the 260 hour Switchboard corpus of two-party US English interactions (Godfrey et al., 1992). Participants discuss a prescribed topic, e.g. office attire (mean session length 6 min 23 s). Prior PTTMs rely on accurate phonetic transcriptions like those provided with Switchboard (Ekstedt and Skantze, 2022b). Using ASR introduces an inevitable alignment error of approximately 480 ms (Russell et al., 2024), though this reflects real-world conditions. Both corpora contain stereo audio (one channel per speaker), which we downsample to 16 kHz. Candor has 320x240 resolution, 30 fps mp4 video.

Identifying turn-taking events We extract `shifts` and `holds` from the transcriptions by identifying silences greater than +250 ms where only one speaker is active 1 second before and after the silence Ekstedt and Skantze (2022b). If the speaker remains the same, it is a `hold` and if the speaker changes, it is a `shift` (see Figure 2). We repeat the above procedure for different FTOs to assess performance across longer and shorter holds and shifts, and overlapping speech. This is a more comprehensive analysis than prior work which aggregates all holds and shifts together (Ekstedt and Skantze, 2022b). In Table 1 we show the complete set of holds and shifts in each corpus. Note that the ASR and ground-truth alignments result in different numbers of holds and shifts in Switchboard, due to ASR's inherent alignment error.

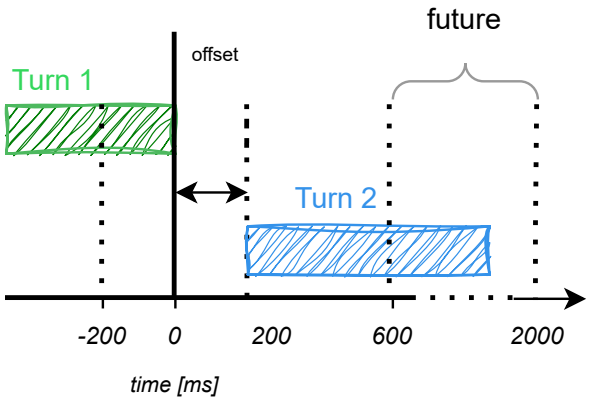

Figure 2: Schematic depicting a shift between speakers

Table 1: Statistics for shifts and holds. Floor-transfer offset (FTO) is the time between the end of the previous turn and the start of the next turn (negative overlap).

Corpus	Minium FTO [ms]	# Shifts	# Holds	Shifts proportion	# Shifts per minute	# Holds per minute
Candor	-250	23515	62483	0.38	0.42	1.11
(ASR)	0	115975	414666	0.28	2.05	7.34
710 hours	250	83158	206830	0.40	1.47	3.66
	500	53360	115475	0.46	0.94	2.04
	750	32831	65804	0.50	0.58	1.16
	1000	19699	35697	0.55	0.35	0.63
	1250	11438	19831	0.58	0.20	0.35
	1500	6148	10250	0.60	0.11	0.18
Switchboard	-250	6599	9419	0.70	0.42	0.61
(ground-truth	0	26883	150407	0.18	1.73	9.67
alignment)	250	16267	74510	0.22	1.05	4.79
260 hours	500	9909	43793	0.23	0.64	2.82
	750	6087	27348	0.22	0.39	1.76
	1000	3893	14180	0.27	0.25	0.91
	1250	2407	6958	0.35	0.15	0.45
	1500	1385	3348	0.41	0.09	0.22
Switchboard	-250	10595	7761	1.37	0.68	0.50
(ASR	0	34886	210386	0.17	2.24	13.53
alignment)	250	20097	143182	0.14	1.29	9.21
260 hours	500	10302	74666	0.14	0.66	4.80
	750	5269	33299	0.16	0.34	2.14
	1000	2702	14952	0.18	0.17	0.96
	1250	1383	6599	0.21	0.09	0.42
	1500	663	2786	0.24	0.04	0.18

Visual feature extraction We use OpenFace (Baltrušaitis et al., 2016) to extract visual features. Facial action units (FAUs, 17 in total) numerically describe facial movements e.g. jaw drop on a scale from 0.0 to 5.0. There is one gaze vector per eye which is a 3 dimensional unit vector. We also extract head position (X, Y, Z in mm) and head rotation (roll, pitch and yaw in radians). We select 15 facial landmarks located on the brow, jaw, nose and lips (x and y in pixels) and a confidence score. In total, there is one 60 dimensional vector per frame. We scale all features to 0, 1 by max min scaling at participant level. Histograms of eye gaze and head pose are unimodal with different modes per participant, reflecting the differing setup of participant devices. We therefore zero mean the head pose and eye gaze vectors at participant level. We show a sample visualisation of OpenFace features in Figure 1. Head pose is depicted as a blue cube centred on the head pointing in the estimated direction of head pose. Eye gaze vectors are depicted in green. Facial landmarks are shown in red, though we only use a subset of 15 in this work. It is not straightforward to visualise FAUs so these are omitted.

OpenFace fails to completely track participants in 238 of the sessions. We investigate and find failure is due to various issues e.g. in one session, a participant gets up and leaves to fetch something. In the overwhelming majority of the corpus (1,418 sessions, 710 hours) OpenFace runs without issue. We manually check tracking performance on a subset and observe good performance. We therefore conduct all our work on these sessions.

For analysis, we compute maximum FAU intensity 200 ms before shifts and holds where the FTO > + 250 ms, a common time frame in PTTM evaluation (Ekstedt and Skantze, 2022b; Roddy et al., 2018a). We select an equal number of 200 ms random periods of silence and speech located far (1 second) away from the start or end of a turn. We then compare the median FAU intensity.

Audio-only turn-taking model We establish the performance of the state-of-the-art voice activity projection (VAP) model (Ekstedt and Skantze, 2022b), an audio-only turn-taking model which has been used in several studies e.g. (Ekstedt and Skantze, 2022a; Ekstedt et al., 2023; Inoue et al., 2024). The VAP model is a transformer (Vaswani, 2017) based neural network trained to predict who will speak in the next 2 seconds: the *training objective*, shown in Figure 3. At each point in time there are four bins per speaker representing the next 2 seconds. The bins span the next 0-200, 200-600,

600-1200 and 1200-2000 milliseconds. If 50% of frames within a bin contain speech, we set the bin label to 1, otherwise the bin is labelled 0. This gives $2^8 = 256$ possible VAP states. The model has 5.8 M parameters, and full details are in Inoue et al. (2024); Ekstedt and Skantze (2022a).

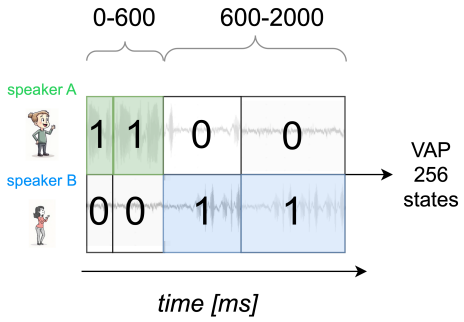

Figure 3: The VAP training objective introduced by Ekstedt and Skantze (2022b) which captures speaking activity in the next 2 seconds in a two-party interaction.

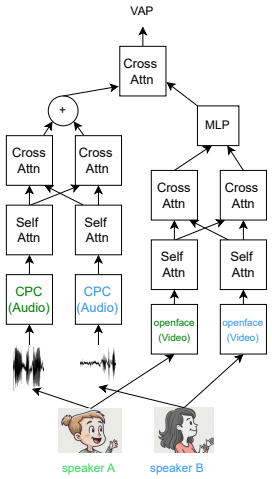

Figure 4: Schematic of our transformer-based multimodal predictive turn-taking model (late fusion version), incorporating audio and video from both speakers.

Multimodal turn-taking model (ours) Like the VAP model, our model, multimodal VAP (MM-VAP, Figure 4) consists of self- and cross-attention blocks. A self-attention block is N stacked transformer decoder layers, where query (q), key (k) and value (v) are identical:

$$\text{SELF-ATTN}(x) = \text{TRANSFORMER}\,(q = x, k = x, v = x) \quad (1)$$

A transformer layer learns attention or the 'compatibility' of output with the query via the key and value (Vaswani, 2017). In self-attention blocks, as $q = k = v$, the model learns temporal patterns in the input e.g. audio or video. In cross-attention blocks, we stack N layers with two inputs x_1 and x_2. In each layer, we compute two transformer layers with shared weights:

$$\text{CROSS-ATTN}(x_1, x_2) = \sigma\left(\begin{matrix}\text{TRANSFORMER}(q{=}x_1, k{=}x_2, v{=}x_2) + \\ \text{TRANSFORMER}(q{=}x_2, k{=}x_1, v{=}x_1)\end{matrix}\right) \quad (2)$$

where $\sigma(\cdot)$ is a layer normalisation operation followed by a GeLU activation. In these layers, the model learns temporal patterns between x_1 and x_2 e.g. from audio to video.

The model in Figure 4 is a *late fusion* model as the audio and video modalities are combined just prior to the output (Baltrušaitis et al., 2018). The initial cross-attention layers learn patterns between each speaker's audio and video separately. The final cross-attention block learns temporal patterns between the video and audio modalities. We also design an early fusion model where audio and video from each participant is fed into a cross-attention layer just after the feature extraction. Both versions have 8.7 M parameters. As each speaker's visual and non-verbal signals are modelled separately and combined with cross-attention, the voice activity signal is not required at inference.

Like the VAP model, we use a pre-trained audio feature extractor (Riviere et al., 2020) yielding a 256 dimensional feature vector at 50 Hz. We freeze the extractor layers in training. We upsample visual features from 30 to 50 Hz to match using linear interpolation. Within the model, we use single hidden layer multilayer perceptrons (MLPs) to project the 60 dimensional vector to 256 dimensions. Self-attention blocks consist of 3 stacked self-attention layers and 1 cross-attention block. We apply a causal masking to ensure that the model can only make predictions from past audio and video frames. We use layer normalisation (Lei Ba et al., 2016) and the GeLU activation function (Hendrycks and Gimpel, 2016) throughout. The model outputs a 256 dimensional vector via the softmax function, learning a probability distribution over all 256 VAP states with a cross-entropy loss (Figure 3).

Training We train all models as follows. We withhold 5% of sessions for testing. We conduct a 5-fold cross-validation on the remaining sessions (80% training, 20% validation). We segment audio and video into 20 second segments with a 2 second overlap and randomly shuffle all segments. We use the cross-entropy loss function to train the model output (256 dimensions) to learn VAP labels (Figure 3). Based on an initial hyperparameter sweep we set the batch size to 16 and the learning rate to

0.005. We train the model for all 5 folds with an Nvidia RTX 6000 GPU for 10 epochs. The GPU hours are: VAP Switchboard 21 hrs, VAP Candor 21 hrs, early fusion 110 hrs, late fusion 113 hrs.

Evaluation We evaluate trained models as follows. At each hold or shift in the validation set, we sum model probabilities in a 200 ms window. The window is located either before the end of a turn, the start of an overlap, or during mutual silence between speaking turns. We sum the shift probability, defined as the marginal probability of all VAP states where the non-active speaker is 1 in both bins in the 600-2000 ms period (Figure 3). Like the original VAP paper (Ekstedt and Skantze, 2022b) we only consider the latter half of the VAP objective for hold/shift evaluation. We find that predictions are less reliable in the first 0-600 ms. We then set a threshold: if the summed probability is greater than this threshold, a shift is predicted. We choose thresholds which maximise the weighted F_1 and balanced accuracy scores on the validation set. Finally, we report performance on the unseen test sessions with these thresholds. We compare F_1 and balanced accuracy with a dummy baseline model which always outputs a hold.

We report performance with the F_1 and balanced accuracy. The F_1 is (Powers, 2020):

$$F_1 = \frac{2 \times \#\text{TP}}{2 \times \#\text{TP} + \#\text{FP} + \#\text{FN}} \quad (3)$$

where TP is the number of true positives, FP false positives and FN false negatives. We compute the F_1 score by arbitrarily assigning a positive 1 label to a shift and a negative 0 to a hold (F_1 shift). We then re-compute with a positive 1 = hold and a 0 = shift (F_1 hold) and report the weighed F_1:

$$F_{1weighted} = \rho_s F_1\text{SHIFT} + \rho_h F_1\text{HOLD} \quad (4)$$

where ρ_s and ρ_h are the proportion of shifts and holds. The weighted F_1 accounts for the presence of more holds than shifts (Table 1). We also report balanced accuracy (Brodersen et al., 2010):

$$\text{BAL. ACCURACY} = \frac{1}{2}\left(\frac{\#TP}{\#P} + \frac{\#TN}{\#N}\right) \quad (5)$$

where TP and TF are the number of true positives and false negatives, and P and N are the numbers of positives (shifts) and negatives (holds). We compare F_1 and balanced accuracy of models on the common test set with the paired t-test (Ross and Willson, 2017). We compare facial action units with the non-parametric Mann-Whitney U (MWU) test (Mann and Whitney, 1947).

Ablation study We conduct an ablation study by training 4 different versions of the MM-VAP model. We train and validate using the same procedure, but each of which receives audio and a subset of the visual features as input. We divide the visual features into gaze, head pose, facial action unit and facial landmarks groups (6, 6, 17, 30 dimensions receptively, excluding the scalar confidence score).

4 Results

4.1 Audio-only turn-taking

We begin by assessing the performance of the audio-only VAP predictive turn-taking model (PTTM). We first consider hold/shift prediction during periods of silence greater than 250 ms, as in Ekstedt and Skantze (2022b). We report the 5-fold average performance in Table 2.

Table 2: Average shift/hold prediction performance of the VAP model on Candor (CND, videoconference) and Switchboard (SWB, telephone) corpora evaluated during silence between turns (FTO > +250 ms).

Corpus	Alignment	F_1 (Weighted)	F_1 (Hold)	F_1 (Shift)	Accuracy (Balanced %)
SWB	ground-truth	0.82	0.89	0.47	67
SWB	ASR	0.81	0.89	0.45	65
CND	ASR	0.83	0.89	0.71	79

All models trained on Switchboard outperform a baseline model which always predicts hold (weighted F_1: 0.74 for SWB ASR and 0.70 for Candor, $p < 0.01$). The Switchboard ground-truth alignment results are comparable with those of Ekstedt and Skantze (2022b), verifying our re-implementation. The F_1 and balanced accuracy scores are slightly higher using the ground-truth alignment, reflecting the ASR alignment error. In the Candor corpus, using audio-only cues, the VAP model performs well above the baseline and the balanced accuracy is higher than Switchboard. There is a higher F_1 shift and a similar F_1 hold and weighted F_1.

4.2 Visual feature analysis

As turn-taking literature details the importance of visual cues (Section 1), we investigate facial expression in Candor interactions. In Figure 5 we show the median peak FAU intensity at key turn-taking events (method in Section 3). We compare

with random speech and random silence with the MWU test and omit comparisons where $p > 0.05$. We exclude outer brow raiser, upper eyelid raiser and nose wrinkler as $p > 0.05$ in all comparisons.

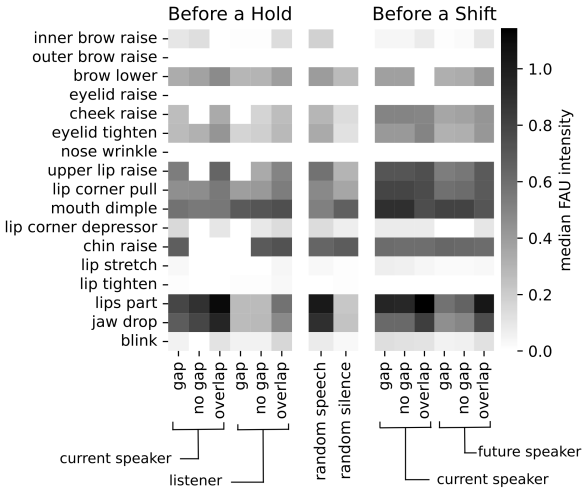

Figure 5: Median FAU intensity in Candor during random speech and silence, and before holds and shifts.

As a sanity check, we observe that when an interlocutor speaks, there is more cheek, eyelid, brow, lip and jaw activity than when they do not speak (*random speech* and *random silence*). At the end of a turn, the FAU activity of the current speaker largely resembles random speech. During a hold, the listener remains silent and FAU activity resembles random silence (*listener* vs *random silence*). We observe subtle differences in the FAU activity of shifts when comparing cheek, lip, and jaw FAUs of *future speaker* with *random silence*. This suggests the presence of a distinct facial expression partially resembling speech ahead of an upcoming transition. This could be exploited in a predictive turn-taking model. We repeated the analysis for eye gaze and head pose. These did not reveal meaningful patterns, so we omit heat maps. These features may still be useful in a neural network which captures patterns across much long periods of time.

4.3 Multimodal turn-taking

We continue by investigating the performance of our proposed early and late fusion turn-taking models (MM-VAP). We train on the Candor corpus (Section 3). For completeness, we also train a VAP model only using video features. This model is identical to the original VAP model with the removal of the audio feature extractor and a dimensionality of 60, reflecting the dimensionality of the visual feature vector.

Table 3: Average shift/hold prediction across 5 folds on the Candor corpus (best in bold). a = audio only, v = video-only, and a+v = multimodal; e = early fusion, l = late fusion. Percentage change relative to the audio-only model is provided as ↑ / ↓, with '-' indicating $p > 0.05$.

Evaluation point	Model	F_1 (Weighted)		F_1 (Hold)		F_1 (Shift)		Accuracy (Balanced %)	
during mutual silence	a	0.83		0.88		0.70		79	
(FTO > +250 ms)	v	0.72	↓14%	0.82	↓8%	0.47	↓33%	68	↓15%
	a+v (e)	0.84	-	0.90	-	0.74	↑5%	82	↑3%
	a+v (l)	**0.86**	↑3%	**0.90**	↑2%	**0.74**	↑6%	**83**	↑4%
before end of turn	a	0.81		0.87		0.67		76	
(FTO > +250 ms)	v	0.70	↓13%	0.80	↓8%	0.45	↓33%	66	↓14%
	a+v (e)	0.83	↑2%	0.88	-	0.69	↑4%	79	↑3%
	a+v (l)	**0.83**	↑3%	**0.89**	↑2%	**0.70**	↑6%	**80**	↑4%
before end of turn	a	0.86		0.91		0.66		77	
(FTO > 0 ms)	v	0.78	↓10%	0.87	↓5%	0.43	↓35%	70	↓10%
	a+v (e)	0.87	↑2%	0.92	-	0.69	↑6%	80	↑3%
	a+v (l)	**0.87**	↑2%	**0.92**	-	**0.71**	↑6%	**83**	↑4%
before overlap	a	0.78		0.85		0.57		70	
(FTO < -250 ms)	v	0.72	↓8%	0.82	↓3%	0.42	↓25%	64	↓8%
	a+v (e)	0.79	↑3%	0.87	↑2%	0.60	↑7%	72	↑3%
	a+v (l)	**0.80**	↑4%	**0.87**	↑2%	**0.62**	↑10%	**74**	↑5%

Multimodal vs. audio-only models We report the average performance over 5 folds of the Candor corpus in Table 3. As before, we consider model performance 200 ms before a shift/hold (FTO > 250 ms). Our multimodal models significantly outperform the audio-only and video-only models ($p < 0.01$). The late fusion model shows a 3% relative increase in the weighted F_1 score and a 6% relative increase in balanced accuracy over the audio-only model. The video-only model has the worst overall performance with a 33% relative reduction in the F_1 shift score.

Expanding the analysis Next, we remove the requirement for a minimum period of silence between turns, increasing the number of shifts and holds considered (FTO > 0 vs FTO > 250 ms in Table 1). We also move the evaluation window to 200 ms before the end of a turn, ensuring the next speaker has not started to speak during evaluation. This hence captures the capacity of the model to predict upcoming shifts while the previous speaker is still speaking, akin to human turn-taking (Section 1). For this more comprehensive set of speaker transitions, our multimodal models outperforms both video-only and audio-only models. Best performance is achieved with the late fusion model, with a 6% relative increase in the F_1 shift score and a 4% relative increase in balanced accuracy (Table 3). We also evaluate hold/shift prediction before overlapping speech. Note that for the purposes of this evaluation, we exclude overlapping speech where there is no change in speaker 1 second after the overlap has concluded. This ensures that brief periods of overlap are excluded (i.e. backchannels).

We find a 6% increase in balanced accuracy and a 10% increase in the F_1 shift scores for the late-fusion MM-VAP model (Table 3).

Performance during longer transitions Our multimodal models are the best-performing overall. We assess if this performance benefit is uniform across all types of transitions. We group subsets of holds and shifts by varying the minimum FTO from 0 to 1500 ms in 250 ms increments (Table 1). We then compute model performance on each subset. We use the balanced accuracy because, unlike the $F1$ weighted score, it equally weights holds and shifts. This is important as proportions of holds/shifts by group (Table 1). We plot the mean balanced accuracy over 5 folds in Figure 6.

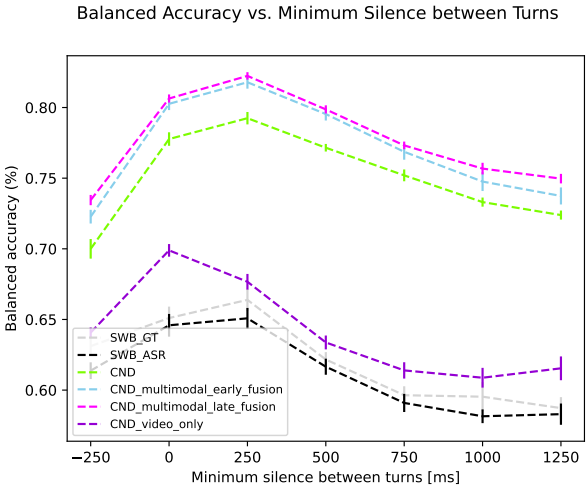

Figure 6: Balanced accuracy of models averaged over 5 folds, ± standard error in the mean, grouped by a minimum period of silence between turns; i.e. the FTO.

The balanced accuracies in Tables 2 and 3 correspond with the 0 and 250 ms points in Figure 6. The gap in performance of Switchboard and Candor is consistent with our prior finding of a lower F_1 shift score on Switchboard (Table 2). Performance falls for both audio-only VAP and multimodal MM-VAP models as the duration of silence between speaking turns increases, indicating that these turns are more challenging to predict. However, we find that the performance of the MM-VAP models is significantly better than the audio-only VAP models across all FTOs ($p < 0.01$ in all cases). Thus, the inclusion of visual information leads to improved performance across the complete range of holds and shifts in the corpus.

Ablation study Finally, we conduct an ablation study, training the late fusion version of the MM-VAP model on audio and gaze, head pose, facial landmarks and facial action unit subsets. We report the 5-fold average performance in Table 4.

Table 4: Ablation study results, 5-fold average. a=audio, v=video, a+v (l) = multimodal, late fusion. Percentage change relative to the a+v (l) model which uses all visual features is shown as ↑ / ↓, with '-' indicating $p > 0.05$.

Evaluation Point	Model	F_1 (Weighted)		F_1 (Hold)		F_1 (Shift)		Accuracy (Balanced %)	
during mutual silence	a + v (l)	0.86		0.90		0.74		83	
(FTO > +250 ms)	a + gaze	0.74	↓14%	0.83	↓8%	0.51	↓31%	67	↓20%
	a + pose	0.84	↓2%	0.89	-	0.72	↓3%	80	↓3%
	a + lmks	0.79	↓8%	0.86	↓4%	0.62	↓17%	74	↓11%
	a + faus	0.85	-	0.89	-	0.73	↓2%	81	↓2%
before end of turn	a + v (l)	0.83		0.88		0.74		80	
(FTO > 250 ms)	a + gaze	0.72	↓13%	0.82	↓8%	0.48	↓35%	65	↓19%
	a + pose	0.82	-	0.87	-	0.68	↓8%	78	↓2%
	a + lmks	0.78	↓7%	0.85	↓4%	0.59	↓20%	72	↓10%
	a + faus	0.83	-	0.88	-	0.69	↓6%	79	↓2%
before end of turn	a + v (l)	0.87		0.92		0.71		83	
(FTO > 0 ms)	a + gaze	0.78	↓10%	0.88	↓5%	0.43	↓39%	65	↓20%
	a + pose	0.87	-	0.92	-	0.67	↓3%	79	↓2%
	a + lmks	0.83	↓5%	0.90	↓3%	0.57	↓18%	73	↓10%
	a + faus	0.87	-	0.92	-	0.68	↓2%	80	↓4%
overlap	a + v (l)	0.80		0.87		0.62		74	
(FTO < 250 ms)	a + gaze	0.68	↓15%	0.81	↓7%	0.35	↓44%	58	↓22%
	a + pose	0.79	↓2%	0.86	-	0.58	↓6%	71	↓3%
	a + lmks	0.74	↓8%	0.84	↓4%	0.49	↓22%	66	↓11%
	a + faus	0.79	↓2%	0.86	-	0.59	↓4%	72	↓2%

The best-performing model is the model which includes all visual features (a + v (l), Table 4). Considering the models trained on the four subsets of the visual features, we note reduced performance, which is most notable when considering the F_1 shift score. We compare the weighted $F1$ score, and find that the gaze and landmark trained models perform significantly worse than the model trained on all visual features in all cases ($p < 0.01$ when comparing a+v (l) with gaze and landmarks, Table 4). There is no significant difference in the performance of the facial action unit and head pose trained models on certain metrics (e.g. $p = 0.09$, before end of turn FTO > 0ms Table 4). However, on the $F1$ shift score, the facial action unit trained model achieves the best performance of all models trained on the reduced visual feature sets ($p < 0.05$). The best-performing model trained on a subset of visual features is therefore the facial action unit model.

5 Discussion

Our re-implementation of the audio-only VAP PTTM (Ekstedt and Skantze, 2022b; Inoue et al., 2024) performed well on the Candor videoconferencing (VC) corpus. However, our new multimodal PTTM, MM-VAP - which uses speech along with facial action units, landmarks, gaze, and head pose - significantly outperformed the VAP model. The performance increase was most notable on the F_1

shift metric, with a 6-10% relative increase in performance (Table 2). Returning to our first research question, our findings show that visual cues improve PTTM performance. This echoes the psycholinguistics literature underlining the importance of visual cues in turn-taking (Section 1).

We also investigated the use of ASR, as to date PTTM models have used manual alignments. We found performance dropped slightly but remained broadly similar, supporting the use of ASR in PTTM development (Table 2).

How does visual information help? Our second research question concerned the aspects of turn-taking which benefit from visual information. As prior work considered all holds and shifts together, we conducted a more fine-grained analysis. This revealed that shifts, or transitions between speakers, benefit the most from visual information. Our working hypothesis is that non-verbal behaviours provide vital turn-taking cues when interlocutors can see one another, and our model benefits from this. Notwithstanding slight differences in the corpora (session length, topic, see Section 3), the most notable difference is that in Candor interlocutors can see one another, whereas in Switchboard, they cannot. Thus, as our results show, worse performance is achieved when audio alone is used to model turn-taking in Candor. Our multimodal model which incorporates visual cues improved performance, and we demonstrated that this is consistent across the complete range of holds and shifts in the corpus. A detailed analysis using Conversation Analysis methodology should be conducted to uncover the exact role visual cues play here, however we found some evidence of enhanced lip jaw and cheek movement before speaking (Figure 5). This indicates the presence of visual cues from the listener prior to the onset of speech (a shift). This is consistent with our finding that visual cues most benefited turn-taking by improving the performance on shifts (F_1 shift, Table 3).

Visual signalling As the role of visual cues in turn-taking is well-supported by the psycholinguistics literature (Sections 1 and 2), we believe visual cues such as gaze aversion are exploited in our multimodal model. Future work is needed to establish the exact role of visual cues and their impact on model performance, but the literature outlines why visual cues are particularly important during speaker transitions. For instance, the more complex a question is, the longer the response (Strömbergsson et al., 2013) e.g. open-ended versus yes/no questions (Walczyk et al., 2003). Extended silences are also associated with gaze aversion in interaction (Walczyk et al., 2003) and hence these may rely more on non-verbal cues.

Finally, we conducted a thorough ablation study, which showed that facial action units are the biggest contributors to model performance. The ablation study demonstrated that not all features contributed to model performance in isolation, most notably gaze (Table 4). This is despite the role of gaze in turn-taking being well-supported in the turn-taking literature (Kendon, 1967). The videoconferencing setup of the corpus (Reece et al., 2023) might be impacting performance here, as videoconferencing is known to impact gaze (Sellen, 1995). Alternatively, gaze extraction is challenging and can be impacted by factors such as lighting levels (Zhang et al., 2021), which are uncontrolled in the Candor corpus. However, we did find that when combined together, the model which incorporated all visual features performed the best (Table 4), suggesting the model does leverage information from all of visual features.

Future work MM-VAP has shown promising results, but it can be improved. OpenFace could be replaced by a more powerful learnable visual front-end, e.g. from the audio-visual speech recognition literature (Fenghour et al., 2021). The model should be updated to handle audio and video at different sample rates (Section 3).

6 Conclusion

We presented the first comprehensive analysis of multimodal predictive turn-taking, introducing MM-VAP a new multimodal model which uses speech along with visual features. We found a strong improvement in performance above the audio-only state-of-the-art in videoconferencing speech. We therefore encourage researchers to incorporate multimodal cues into models for predictive turn-taking and multimodal interaction more generally. We make all code publicly available.

7 Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Trinity College Dublin. Speechmatics provided academic access to their ASR platform.

8 Limitations

Our paper has a number of inevitable limitations, which we discuss below.

Data limitations We compared telephone and videoconferencing (VC) interaction and did not include any in-person corpora. Videoconferencing interaction has a number of key differences to in-person interaction, including limited use of body language and difficulties in knowing when to speak (Sellen, 1995; Isaacs and Tang, 1993; O'Conaill et al., 1993; Boland et al., 2021). Nevertheless, we have shown that including visual features improves performance in PTTMs trained on VC data. We expect that this would also be the case in in-person settings, where interlocutors are not hindered in their use of non-verbal communication by technology, though this should be confirmed. A comparison of a model trained on a corpus of in-person interaction would be useful. However, this is not straightforward due to the lack of availability of a suitable public dataset. Available corpora of dyadic interaction are not large, e.g. the 11-hour Mahnob mimicry corpus (Bilakhia et al., 2015). Furthermore, varying camera angles can complicate visual feature extraction, unlike in VC where the angle is front-facing. The use of ambient microphones in in-person settings introduces issues such as background noise and speaker diarisation. These issues are resolved through the advanced audio processing algorithms in modern VC platforms.

Feature extraction limitations A further limitation is the use of OpenFace (Baltrušaitis et al., 2016) for visual feature extraction. OpenFace has the benefit of being easily interpretable by humans, through the extraction of high-level features like head pose, but this may not be the most useful representation to use in a neural network. However, the model does benefit from these features as shown by our results. As we have suggested for future work, the audio-visual speech recognition literature is a good starting point for visual front-end architectures (Fenghour et al., 2021; Ivanko et al., 2023). OpenFace is also not robust as it failed on a minority of sessions. This is due to issues beyond our control, which are an inherent part of data captured *in-the-wild*. Excluding these sessions leaves 710 hours; more than sufficient for deep learning. Future front-ends should be made more robust to missing data arising from issues such as participants leaving the frame. We did not assess how OpenFace tracking may also be hampered by variable lighting conditions, glasses, facial hair, etc. but we observed that tracking was good in sessions we verified manually. Again, these artefacts will be unavoidable in data captured in naturalistic interaction. The Candor and Switchboard corpora, though captured in naturalistic settings, are free from any specific background noises, and the audio quality is good. In less controlled settings, the ASR transcription could be of lower quality than the one used here. This is due to the impact of background noises, echo, poor microphone quality, all of which degrade ASR performance (Agrawal and Ganapathy, 2019; Alharbi et al., 2021).

Model limitations Our multimodal model itself has its own limitations. It is unable to handle video and audio features at different frame rates (30 Hz / fps for the video and 50 Hz for the audio after feature extraction). We resolved this by upsampling the visual features in time to 50 Hz (i.e. a factor of 1.67). We applied causal masking to ensure the model could not use future information to make predictions, though the upsampling does introduce a slight future 'bleed' as present frames are modified by the next frame through interpolation. A future version of the model could overcome this by ensuring that visual and acoustic features are handled at different temporal rates as done with an RNN in (Roddy et al., 2018b). The acoustic features are processed by passing them through a pre-trained feature extractor (Riviere et al., 2020), whereas the visual features are high-level descriptors, e.g. angles in radians. This issue could be resolved in a future iteration of the model, replacing the OpenFace front-end, as discussed. Nevertheless, the model introduced in this paper shows considerable performance improvements over the state-of-the-art audio-only approach.

References

Purvi Agrawal and Sriram Ganapathy. 2019. Modulation filter learning using deep variational networks for robust speech recognition. *IEEE journal of selected topics in signal processing*, 13(2):244–253.

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, 9:131858–131876.

Anne H Anderson, Miles Bader, Ellen Gurman Bard,

Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

Pashiera Barkhuysen, Emiel Krahmer, and Marc Swerts. 2008. The interplay between the auditory and visual modality for end-of-utterance detection. *The journal of the Acoustical Society of America*, 123(1):354–365.

Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. 2015. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61.

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Julie E Boland, Pedro Fonseca, Ilana Mermelstein, and Myles Williamson. 2021. Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, 151(6):1272–1282.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? investigations using voice activity projection models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice activity projection: Self-supervised learning of turn-taking events. In *Interspeech 2022*, pages 5190–5194.

Erik Ekstedt, Siyang Wang, Éva Székely, Joakim Gustafson, and Gabriel Skantze. 2023. Automatic evaluation of turn-taking cues in conversational speech synthesis. In *INTERSPEECH 2023*, pages 5481–5485.

Souheil Fenghour, Daqing Chen, Kun Guo, Bo Li, and Perry Xiao. 2021. Deep learning-based automated lip-reading: A survey. *IEEE Access*, 9:121184–121205.

Simon Garrod and Martin J Pickering. 2015. The use of content and timing to predict turn transitions. *Frontiers in psychology*, 6(751):1–12.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Judith Holler, Kobin H Kendrick, Marisa Casillas, and Stephen C Levinson. 2016. *Turn-taking in human communicative interaction*. Frontiers Media SA.

Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Peter Indefrey. 2011. The spatial and temporal signatures of word production components: a critical update. *Frontiers in psychology*, 2:255.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11873–11883, Torino, Italia. ELRA and ICCL.

Ellen A Isaacs and John C Tang. 1993. What video can and can't do for collaboration: a case study. In *Proceedings of the first ACM International Conference on Multimedia*, pages 199–206. ACM Press.

Denis Ivanko, Dmitry Ryumin, and Alexey Karpov. 2023. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 11(12).

Gabriele Kasper and Johannes Wagner. 2014. Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34:171–212.

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

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. In *INTERSPEECH 2023*, pages 2658–2662.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *ArXiv e-prints*, pages arXiv–1607.

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

Siyan Li, Ashwin Paranjape, and Christopher D Manning. 2022. When can i speak? predicting initiation points for spoken dialogue agents. *arXiv preprint arXiv:2208.03812*.

Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. 2020. Who Speaks Next? Turn Change and Next Speaker Prediction in Multimodal Multiparty Interaction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 349–354, Baltimore, MD, USA. IEEE.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech Language*, 71:101255.

Naomi Nota, James P Trujillo, and Judith Holler. 2023. Conversational eyebrow frowns facilitate question identification: An online study using virtual avatars. *Cognitive Science*, 47(12):e13392.

Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. 1993. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. *Human–Computer Interaction*, 8(4):389–428.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2024. Multimodal voice activity projection for turn-taking and effects on speaker adaptation. *IEICE Transactions on Information and Systems*.

David MW Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018a. Investigating speech features for continuous turn-taking prediction using lstms. *arXiv preprint arXiv:1806.11461*.

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

Amanda Ross and Victor L. Willson. 2017. *Independent Samples T-Test*, pages 13–16. SensePublishers, Rotterdam.

Sam O'Connor Russell, Iona Gessinger, Anna Krason, Gabriella Vigliocco, and Naomi Harte. 2024. What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics*, 3(3):100163.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Abigail Sellen. 1995. Remote Conversations: The Effects of Mediating Talk With Technology. *Human-Computer Interaction*, 10(4):401–444.

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*, pages 220–230.

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

Speechmatics, Ltd. 2024. Speechmatics ASR. [Online; accessed Dec 2024].

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Sofia Strömbergsson, Anna Hjalmarsson, Jens Edlund, and David House. 2013. Timing responses to questions in dialogue. In *Interspeech*, volume 2013, pages 2584–2588.

James P Trujillo, Stephen C Levinson, and Judith Holler. 2021. Visual information in computer-mediated interaction matters: Investigating the association between the availability of gesture and turn transition timing in conversation. In *Human-Computer Interaction. Design and User Experience Case Studies: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part III 23*, pages 643–657. Springer.

A Vaswani. 2017. s. *Advances in Neural Information Processing Systems*.

Jeffrey J Walczyk, Karen S Roper, Eric Seemann, and Angela M Humphrey. 2003. Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(7):755–774.

Allison Woodruff and Paul M Aoki. 2003. How push-to-talk makes talk less pushy. In *Proceedings of the 2003 ACM International Conference on Supporting Group Work*, pages 170–179.

Xucong Zhang, Seonwook Park, and Anna Maria Feit. 2021. Eye gaze estimation and its applications. *Artificial Intelligence for Human Computer Interaction: A Modern Approach*, pages 99–130.