6. Interdisciplinary Conference on Electrics and Computer (INTCEC 2026)
24-25 September 2026, Chicago-USA

Real-TurnTurk: A Multimodal Turkish Corpus for Turn-Taking Prediction

Ahmet Tuğrul Bayrak
Data Science and Innovation
Ata Technology Platforms
Istanbul, Turkey
tugrul.bayrak@atptech.com

Fatma Nur Korkmaz
Data Science and Innovation
Ata Technology Platforms
Istanbul, Turkey
fatmanur.korkmaz@atptech.com

Bekir Berker Türker
Data Science and Innovation
Ata Technology Platforms
Istanbul, Turkey
berker.turker@atptech.com

Mustafa Sertaç Türkel
Data Science and Innovation
Ata Technology Platforms
Istanbul, Turkey
sertac.turkel@atptech.com

Alper Kaplan
Digital Operations and Data
Luxembourg National Research Fund
Luxembourg, Luxembourg
alper.kaplan@fnr.lu

Accepted to INTCEC 2026. This is the author's pre-print version. The final authenticated version will be available through the conference proceedings.

***Abstract*—Turn-taking is a basic organizational feature of human conversation and remains difficult to model in natural, synchronous dialog systems. While existing research has explored multimodal approaches and large language models for turn-ending prediction, there is a lack of naturalistic conversational corpora specifically addressing turn-taking dynamics in Turkish. This study introduces a multimodal Turkish conversational dataset of unscripted dyadic interactions, comprising synchronized front-facing video, per-speaker audio channels that allow overlapping speech to be attributed to individual speakers, and time-aligned transcriptions. Turn-taking prediction is formulated as a binary classification problem, and a Genetic Algorithm (GA) is employed to optimize interpretable decision rules derived from visual, acoustic, and linguistic features. A hybrid AND-OR rule representation is adopted in the proposed framework to represent the alternative cue combinations that precede a turn transition.**

***Index Terms*—turn-taking prediction, predictive modeling, multimodal, data generation, rule optimization, genetic algorithms**

I. Introduction

Spoken dialogue systems built on Large Language Models (LLMs) are now widely deployed, however several interaction-level problems remain unresolved. Many current systems rely on silence detection to determine turn completion: the system assumes the user has finished speaking after a predetermined period of silence. The core limitation is the variability of human speech: some speakers proceed rapidly with minimal pauses, while others require longer intervals to formulate their thoughts without yielding the turn. When a system misjudges these cues and transitions prematurely, speech overlaps occur and dialogue synchronization degrades.

Sacks et al. [1] introduced turn-constructional units (TCUs) and transition relevance places (TRPs), a framework that still guides computational and empirical work. Silence durations and turn offsets carry paralinguistic meaning and reflect individual conversational styles, and simple acoustic thresholds cannot reliably capture the hold/shift distinction [2], [3]; conversational context and task type further alter turn-taking dynamics, indicating that metrics derived from structured elicitation tasks may not generalize to naturalistic dialogue [4].

In computational modeling, standard metrics often fail to capture the real-time trade-off between response latency and false cut-ins [5], and reviews stress that the absence of standardized multilingual benchmarks remains a significant limitation [6], [7]. Continuous, frame-level formulations have increasingly been used in place of silence thresholds: multi-scale recurrent models predict upcoming speech activity from multimodal streams [8], and Voice Activity Projection learns turn-taking events self-supervised from future voice activity, with real-time variants now embedded in incremental dialogue systems [9], [10]; full-duplex speech foundation models integrate turn management into the generative model itself [11]. While language models and lexico-syntactic features improve turn-ending prediction in text settings [12], [13], multimodal fusion of audio, text, and gesture is reported to improve on unimodal baselines [13], [14]. Social dynamics such as participatory profiles and user competency further affect conversational equity, which motivates adaptive models [15], [16].

Turkish conversational resources have expanded to sentiment analysis and synthetic turn-taking datasets [17], [18]; however, a gap remains for naturalistic Turkish corpora with primary turn-taking annotations. This study introduces such

TABLE I
STATISTICS BY CONVERSATION

Speaker Set	Dur. (min)	Speech (min)	Non-sp. (min)	Non-sp. (%)	# Words
user_01 & user_02	26.5	22.7	3.7	14.1	3,267
user_03 & user_04	16.0	13.7	2.3	14.4	2,236
user_01 & user_05	16.4	13.6	2.8	17.2	2,219
user_01 & user_04	24.4	20.9	3.5	14.5	3,290
user_03 & user_06	17.1	14.1	3.1	17.9	2,252
user_03 & user_07	25.5	22.9	2.6	10.1	3,856
user_03 & user_08	28.5	24.6	3.9	13.7	3,960
user_01 & user_06	17.1	14.5	2.5	14.8	2,174
user_03 & user_09	25.1	23.1	2.0	8.1	3,917
user_03 & user_10	25.7	23.0	2.7	10.5	3,952
user_01 & user_11	31.4	29.0	2.4	7.6	4,605
Total	**253.6**	**222.1**	**31.5**	**12.4**	**35,728**

a corpus of unscripted dyadic interactions, with synchronized video, per-speaker audio channels in most interactions, and time-aligned transcriptions, and optimizes rules over features from each modality with a genetic algorithm to detect turn change points.

II. DATA

The dataset comprises 11 unscripted dyadic interactions with 4.23 hours (253.6 min) of synchronized VTT text, WebM audio, and MP4 video. The interactions were recorded with one 48 kHz WebM file per speaker, which allows overlapping speech to be attributed reliably to a single speaker. Video is 1920×1080 at 16 fps, with the two participants occupying the left and right halves of the frame. The transcripts contain 5,383 speech segments and 35,728 words with millisecond-precision timestamps and speaker labels.

Of the total duration, 222.1 min (87.6%) is speech and 31.5 min (12.4%) is non-speech. Because overlapping speech is counted for both participants, the summed speaker talk time is 231.4 min; this is the reference total against which the percentages of Table II are computed, whereas the word percentages of the same table are computed against the 35,728-word total. Under the criteria of Section II-A, 1,750 turn-taking instances were annotated, excluding backchannels from positives, against 3,500 sampled negatives. Per-conversation statistics are listed in Table I and per-speaker statistics in Table II.

The speakers are not evenly represented. As the *# Conv.* column of Table II shows, two act as recurring interlocutors: user_01 in 5 conversations and user_03 in 6. These two never converse with each other, and every conversation contains exactly one of them; together they account for 60.8% of words and 60.1% of speaking time. Of the rest, user_04 and user_06 appear twice and the others once. This recording design, in which two speakers act as recurring interlocutors, produces the speaker-level imbalance and directly constrains the cross-validation protocol (Section IV-C). A subset of the dataset is available on Hugging Face at https://huggingface.co/datasets/tugrulbayrak/Real-TurnTurk.

A. Turn Labelling

A turn change is the moment when the current speaker ends and the next begins. Turn changes are identified by a semi-automatic procedure over VTT transcripts: parsing cues, merging consecutive cues that carry the same speaker label into a single continuous turn, detecting speaker transitions between merged turns, and applying a rule-based filter that separates genuine turns from backchannels. Merging is necessary because sequential cues from one speaker are a segmentation artifact and must not be counted as transitions. As shown in Fig. 1, a transition is a turn change only when four criteria hold concurrently:

$$\text{TurnChange}(t)=1 \iff \begin{array}{l}(S_i \neq S_{i+1}) \\ \wedge\,(d_{i+1} \geq 0.5s) \\ \wedge\,(T_{i+1} \notin \text{FillerWords}) \\ \wedge\,(\text{len}(T_{i+1}) > 3)\end{array} \tag{1}$$

Where t denotes the onset time of segment $i{+}1$; S_i and S_{i+1} represent the speakers of the current and subsequent segments; d_{i+1} is the duration of the subsequent segment; and T_{i+1} is its textual content, with $\text{len}(\cdot)$ measured in characters.

- *Speaker*: the subsequent segment must come from a different speaker.
- *Duration*: a 0.5 s threshold is applied, since minimal feedback signals in Turkish (e.g., "hıhı", "mm") typically last under 0.5 s while genuine turn initiations are longer.
- *Content*: segments whose text is a Turkish filler token, or whose surface form is three characters or fewer, are excluded as acknowledgments.
- *Temporal*: the timestamp is fixed at the onset of the incoming speaker's segment, giving a precise reference point for feature extraction.

Two labelling criteria (duration and filler status) overlap with linguistic features (*word_duration*, *is_filler*) available to the optimizer, which could allow a rule to reproduce the labelling function rather than conversational structure. The two are computed on disjoint speech, however: (1) is evaluated on the *incoming* segment $i+1$, whereas features cover the 2.0 s window *preceding* t, i.e. the outgoing speaker. This is also why *is_filler* can act as a positive predictor although fillers are excluded from positive labels: the occurrences refer to different speakers.

3,500 negatives were chosen uniformly at random from all points that satisfy the feature-extraction requirements yet are not annotated turn changes, with a ±1.0 s exclusion buffer around every positive, ensuring that adjacent windows cannot leak across classes. Sampling was stratified per conversation in proportion to duration, seed 42. Backchannel onsets are excluded from positives by (1); they are eligible as negatives. Because 12.4% of the corpus is non-speech, most negatives fall inside ongoing speech rather than in pauses. The 1:2 ratio was fixed a priori and not tuned.

Overlapping speech is not excluded from the analysis. 81.4% of annotated transitions contain some overlap, mea-

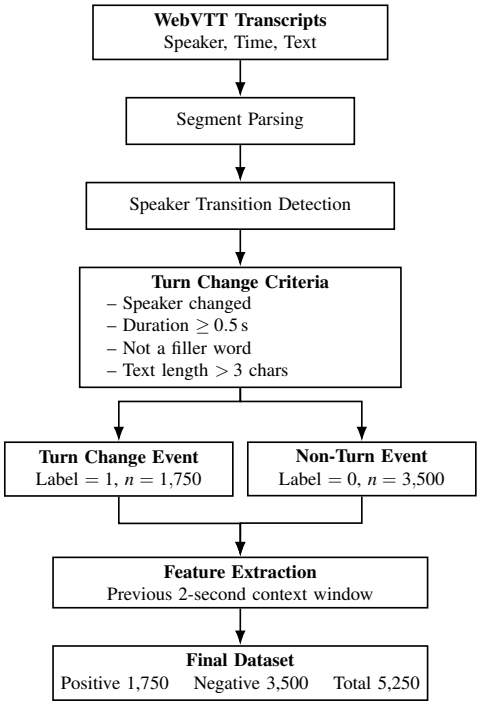

Fig. 1. Turn labelling flow.

surable per speaker in the nine interactions with separated channels.

III. Feature Generation

In total, 28 features are extracted per 2-second analysis window preceding each candidate turn change point. Table III lists all of them with their units and observed ranges. All 28 features describe the current speaker, that is, the participant holding the turn over the 2.0 s window preceding t.

A. Visual Features

9 features capture facial activity, gaze behavior, head motion, and expressiveness: landmark motion magnitude, action unit (AU) intensity change, blink rate, gaze direction change rate, mouth aperture variation, head rotation, head translation, eyebrow displacement, and lip corner pull. Speakers frequently display anticipatory signals such as looking away, nodding, or a drop in facial activity shortly before a turn is yielded.

TABLE II
Speaker Contribution Statistics

Speaker	# Words	% Words	Speak. (min)	% Speak. time	# Conv.
user_01	9,523	26.7	66.2	28.6	5
user_02	1,044	2.9	8.2	3.5	1
user_03	12,172	34.1	72.9	31.5	6
user_04	1,937	5.4	12.8	5.5	2
user_05	1,196	3.3	7.7	3.3	1
user_06	1,376	3.9	9.8	4.2	2
user_07	1,241	3.5	8.5	3.7	1
user_08	1,759	4.9	11.5	5.0	1
user_09	1,784	5.0	11.5	5.0	1
user_10	1,922	5.4	11.3	4.9	1
user_11	1,774	5.0	11.0	4.8	1
Total	**35,728**	**100.0**	**231.4**	**100.0**	**22**

TABLE III
Feature Details

Feature	Unit (range)
Visual (9), MediaPipe Face Mesh + Py-Feat, 16 fps	
landmark_movement	norm. coord. (0–100)
au_intensity_change	AU intensity (0–1)
blink_rate	blinks per 2 s (0–10)
gaze_changes	normalized (0–1)
mouth_opening_change	MAR (0–50)
head_pose_rotation	degrees (0–180)
translation_changes	norm. coord. (0–100)
eyebrow_movement	normalized (0–1)
lip_corner_pull	AU12 intensity (0–1)
Acoustic (12), eGeMAPSv02 via openSMILE, 10 ms frames	
f0_mean	Hz (80–400)
f0_variation	ratio (0–1)
f0_variation_scaled	ratio (0–0.5)
energy_mean	norm. RMS (0–1)
energy_std	norm. RMS (0–0.5)
energy_rate	RMS s^{-1} (unbounded)
onset_rate	onsets/s (0–20)
spectral_centroid	Hz (0–8000)
spectral_rolloff	Hz (0–8000)
spectral_bandwidth	Hz (0–8000)
spectral_contrast	dB (0–100)
voiced_ratio	ratio (0–1)
Linguistic (7), from time-aligned transcripts	
word_duration	s (0.2–5.0)
syllable_count	count (1–10)
is_filler	binary $\{0, 1\}$
is_question	binary $\{0, 1\}$
is_agreement	binary $\{0, 1\}$
word_repetition	binary $\{0, 1\}$
sentence_completeness	score (0–1)

B. Acoustic Features

12 features represent prosodic and spectral properties: mean F_0 plus its coefficient of variation and range-scaled variant; mean, standard deviation, and mean absolute per-second change of RMS energy; spectral centroid, 85% rolloff, bandwidth, and contrast; voiced-frame ratio; and onset rate. A sudden drop in energy or a flattening pitch contour is commonly associated with a completed turn.

C. Linguistic Features

7 features capture lexical, syntactic, and discourse properties: mean word duration and mean syllables per word in the window, three binary indicators for whether the final token is a filler, whether the window contains an interrogative token, and whether it contains an affirmative token, a binary indicator of immediate word repetition, and a syntactic completeness score. Fillers and incomplete structures typically accompany continuation, whereas interrogatives and completed units mark potential transitions.

D. Extraction Tools and Windowing

Each frame is split into left and right halves corresponding to the two participants and each crop is processed indepen-

Fig. 2. Simplified chromosome representation containing two features for visualization purposes.

dently. Landmarks, head pose, eye aspect ratio, and mouth aspect ratio are obtained with MediaPipe Face Mesh [19] and AU intensities with Py-Feat [20]; frame values at 16 fps are aggregated by mean and standard deviation. Visual features are computed from the current speaker's crop only. Not using the listener's crop is a limitation, since listener gaze and nodding are established transition cues. Acoustic features are the eGeMAPSv02 set [21] extracted with openSMILE [22] at 10 ms frames on each speaker's own channel, which also supplies F_0; onsets and offsets are obtained from Silero VAD [23] and word timings from wav2vec 2.0 forced alignment [24]. Every window covers the 2.0 s immediately preceding t; no information at or after t enters the feature vector. No feature scaling is applied. Thresholds are sampled directly from each feature's observed range on the training partition, which keeps test-partition statistics out of the search. The GA is implemented with DEAP [25].

IV. Methods

A. Problem Formulation

The task is binary classification: given the feature vector of the two-second window preceding a conversational event, decide whether the current speaker is about to yield the turn (1) or continue (0). The framework targets rules that are interpretable and cheap enough to evaluate for use in real-time systems, and prioritizes transparency over benchmarking against black-box classifiers. Turn transitions are not assumed to follow a single cue: a speaker may yield after completing a question, after a characteristic prosodic pattern, or following a specific visual behavior. The model must therefore represent several alternative pathways, which motivates the hybrid AND–OR representation below.

B. Genetic Algorithm Based Rule Optimization

Genetic algorithms are population-based evolutionary search methods [26], [27] well suited to the combinatorial space of feature–threshold–operator combinations induced by rule learning. Each individual encodes a complete rule as a sequence of conditions; each condition contributes a feature, a numeric threshold, a comparison operator ($\geq$, $\leq$, $=$), and a logical operator ($\wedge$, $\vee$) linking it to the next, giving rules of the form *if* $(f_1\ c_1\ \theta_1)$ *[op]* $(f_2\ c_2\ \theta_2)$ *… then* $y = 1$*, else* $y = 0$. The number of conditions per chromosome is drawn uniformly between 3 and 7. Fig. 2 illustrates a two-condition chromosome, where F is feature, C comparison operator, T threshold, and L logical operator.

TABLE IV
Genetic Algorithm Configuration

Parameter	Value
Population size / max. generations	500 / 900
Selection	tournament, size 5
Crossover	single-point on conditions, rate 0.85
Mutation	rate 0.10 per offspring, 1/3 each for threshold, feature, logic operators
Elitism	top 5% carried over unchanged
Conditions per rule	$\mathcal{U}\{3, \ldots, 7\}$
Fitness	F_1 on the training partition
Runs per fold / seeds	5 / 1–5

Algorithm 1 GA-based AND–OR rule discovery

Require: Training dataset; Feature pool
Ensure: Best performing rule
Initialize population: generate random rules containing 3 to 7 conditions
for generation = 1 to 900 **do**
  Calculate F1-score for all rules in the current population
  Carry over the top 5% of rules to the next generation $\{\text{Elitism}\}$
  while next generation size < 500 **do**
    Select two parent rules using tournament selection
    Generate an offspring via crossover (85
    Mutate the offspring's threshold, feature, or logic (10
    Add the offspring to the next generation
  end while
  if maximum F1-score has not improved for 100 generations **then**
    break
  end if
end for
return Rule with the highest F1-score

Table IV lists the configuration and Algorithm 1 the procedure. Tournament selection controls selection pressure, crossover recombines condition blocks from two parents, the three mutation operators (threshold, feature, logic) perturb numeric boundaries, replace features, and replace logical connectives, and elitism carries the current best rule into the next generation unchanged. The AND–OR encoding allows one rule to express multiple independent pathways to a transition, which a single threshold cannot represent. Candidate rules are scored by F_1 on the training partition: optimizing precision alone yields conservative rules that miss true transitions, optimizing recall alone produces a high false-positive rate, and F_1 balances the two. A prediction counts as a true positive when it falls within 0.5 s of an annotated turn change, which prevents temporal imprecision in the reference timestamps from being charged as an error; unmatched predictions are false positives and unmatched references false negatives. The same tolerance is applied to all baselines in Table VI.

TABLE V
EVOLUTION OF MODEL PERFORMANCE ACROSS GENERATIONS (5-FOLD AVERAGES)

Generations	Precision	Recall	F_1
0	32.4	49.3	39.1
100	38.6	60.4	47.1
300	45.1	66.0	53.6
500	45.9	71.2	55.8
700	46.0	73.6	56.6
900	46.1	74.6	57.0

TABLE VI
PERFORMANCE RESULTS

Model	Precision	Recall	F_1
Always predict turn change	33.3	100.0	50.0
Random, $p = 0.5$	33.3	50.0	40.0
Stratified random	33.3	33.3	33.3
Silence threshold 1.0 s	10.5	40.3	16.7
Silence threshold 2.0 s	20.4	35.0	25.8
Silence threshold 3.0 s	22.2	10.7	14.4
GA rule (proposed)	46.1	74.6	**57.0**

C. Cross-Validation Strategy

A 5-fold cross-validation is applied with partitioning at the conversation rather than the sample level, ensuring that samples from one conversation do not appear in both training and test. Because 11 conversations do not divide evenly into five parts, the folds hold three, two, two, two, and two conversations respectively. For each fold, four folds train and one tests; GA optimization runs on the training data only and the discovered rule is evaluated on the unseen test conversations. Each fold serves as test partition exactly once, and precision, recall, and F_1 are averaged across folds.

Conversation-level partitioning eliminates conversation-specific leakage, while speaker-specific leakage remains. Because every conversation contains user_01 or user_03 (Section II), a fully speaker-disjoint 5-fold split cannot be realized here and the same speakers necessarily occur in training and test folds. A fully speaker-disjoint estimate would require a leave-one-speaker-out protocol over the nine non-recurring speakers, which is left to future work; the reported figures should be read as an upper bound on generalization to unseen speakers.

V. RESULTS

Table V reports the evolution of test-fold performance and Table VI compares the final rule against baselines; values are means over the 5 folds. The silence baselines predict a turn change at t whenever the pause immediately preceding t, taken from the Silero VAD offsets, exceeds $\tau \in \{1.0, 2.0, 3.0\}$ s; no other information is used. The three reference baselines follow analytically from the 1,750:3,500 class ratio.

Because the class ratio is 1:2 (1,750 positives to 3,500 negatives across the corpus, and preserved within every fold), always predicting a turn change attains precision 33.3, recall 100.0, $F_1 = 50.0$. The always-positive baseline is the primary reference point used below. Against it the GA rule improves F_1 by 7.0 points and raises precision from the 33.3% base rate to 46.1% at 74.6% recall. All three silence thresholds fall below it. This follows from the corpus composition described in Section II-A: non-speech is rare (12.4%) and 81.4% of transitions involve overlap, which leaves pause duration weakly informative here. The gap to the silence baselines is therefore not interpreted as evidence of high predictive accuracy, and given the 11 conversations in the corpus, no claim of statistical significance is made for the 7.0-point margin over this baseline.

After the evaluation phase, the GA was re-trained on the entire dataset to derive one interpretable rule, given in (2) and reported for qualitative discussion. Thresholds are given in the units listed in Table III:

$$\begin{aligned}&(\text{word\_duration} \geq 0.60\,\text{s} \wedge \text{energy\_rate} \geq 1.00)\\ \vee\,&(\text{gaze\_changes} \geq 0.35 \wedge \text{f0\_mean} \geq 120\,\text{Hz})\\ \vee\,&(\text{is\_filler} = 1 \wedge \text{word\_duration} \geq 0.80\,\text{s})\end{aligned} \tag{2}$$

In the rule, longer word duration combined with energy change acts as a prosodic closing cue, gaze shifts paired with mean F_0 form a visual-acoustic transition signal, and fillers, although insufficient alone, become informative when prolonged. Only 5 of the 28 features are selected and the visual modality contributes one condition; the three modalities do not contribute equally in the final rule.

VI. CONCLUSION

This study introduced a multimodal Turkish turn-taking corpus of 4.23 hours of synchronized video, per-speaker audio, and time-stamped transcripts with 1,750 filtered turn-change events, together with a rule optimization procedure for predicting turn transitions from these signals. Instead of a black-box classifier, prediction is formulated as binary classification over interpretable rules evolved with a hybrid AND–OR representation, on the premise that prosodic, visual, and linguistic mechanisms can independently trigger a transfer. The resulting rule reaches $F_1 = 57.0\%$, above both the silence thresholds and the always-positive baseline of $F_1 = 50.0$, and it can be inspected directly. The margin over that baseline is modest. Future work will use a balanced, non-hub recording design with speaker-normalized acoustic features, compare against Random Forest, XGBoost, and transformer-based classifiers, and integrate the rules into real-time LLM-based agents.

REFERENCES

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] R. A. Patamia, H. P. T. Dinh, M. Liu, and A. Cosgun, "Turn-taking modelling in conversational systems: a review of recent advances," *Technologies*, vol. 13, no. 12, art. no. 591, 2025.

[3] C. Threlkeld, M. Umair, and J. de Ruiter, "Using transition duration to improve turn-taking in conversational agents," in *Proc. SIGDIAL*, 2022, pp. 193–203.

[4] S. Watson, A. J. M. Sørensen, and E. MacDonald, "The effect of conversational task on turn taking in dialogue," in *Proc. ISAAR*, vol. 7, 2020, pp. 61–68.

[5] D. Lala, K. Inoue, and T. Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios," in *Proc. ACM ICMI*, 2018, pp. 78–86.

[6] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Comput. Speech Lang.*, vol. 67, 101178, 2021.

[7] G. Castillo-López, G. de Chalendar, and N. Semmar, "A survey of recent advances on turn-taking modeling in spoken dialogue systems," in *Proc. IWSDS*, 2025, pp. 254–271.

[8] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale RNNs," in *Proc. ACM ICMI*, 2018, pp. 186–190.

[9] E. Ekstedt and G. Skantze, "Voice activity projection: self-supervised learning of turn-taking events," in *Proc. Interspeech*, 2022, pp. 5190–5194.

[10] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Real-time and continuous turn-taking prediction using voice activity projection," in *Proc. IWSDS*, 2024.

[11] A. Défossez et al., "Moshi: a speech-text foundation model for real-time dialogue," arXiv:2410.00037, 2024.

[12] S. Z. Razavi, "Dialogue management and turn-taking automation in a speech-based conversational agent," Ph.D. dissertation, Univ. of Rochester, 2021.

[13] M. J. Pinto and T. Belpaeme, "Predictive turn-taking: leveraging language models to anticipate turn transitions in human-robot dialogue," in *Proc. IEEE RO-MAN*, 2024, pp. 1733–1738.

[14] Y. Lin, Y. Zheng, M. Zeng, and W. Shi, "Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals," arXiv:2505.12654, 2025.

[15] L. Hu and G. Chen, "Exploring turn-taking patterns during dialogic collaborative problem solving," *Instr. Sci.*, vol. 50, no. 1, pp. 63–88, 2022.

[16] V. Chattaraman, W.-S. Kwon, J. E. Gilbert, and K. Ross, "Should AI-based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults," *Comput. Hum. Behav.*, vol. 90, pp. 315–330, 2019.

[17] E. N. Polat, C. Demiroğlu, O. T. Yıldız, and N. Kafescioğlu, "Decoding emotional dynamics: a comparative analysis of contextual and non-contextual models in sentiment analysis of Turkish couple dialogues," *IEEE Access*, vol. 12, pp. 172648–172695, 2024.

[18] A. T. Bayrak, M. S. Türkel, and F. N. Korkmaz, "Syn-TurnTurk: a synthetic dataset for turn-taking prediction in Turkish dialogues," arXiv:2604.13620, 2026.

[19] C. Lugaresi et al., "MediaPipe: a framework for building perception pipelines," arXiv:1906.08172, 2019.

[20] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, "Py-Feat: Python facial expression analysis toolbox," *Affect. Sci.*, vol. 4, pp. 781–796, 2023.

[21] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[23] Silero Team, "Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," GitHub repository, 2024. [Online]. Available: https://github.com/snakers4/silero-vad

[24] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12449–12460.

[25] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: evolutionary algorithms made easy," *JMLR*, vol. 13, pp. 2171–2175, 2012.

[26] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Univ. of Michigan Press, 1975.

[27] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.