Turn Taking

Ravi M H
SIGNALS AND RULES

• Communication mechanism for managing taking of turns in face-to-face interaction
• Signals, cues and rules are used
• Three basic signals
  – Turn-yielding signals
  – Attempt-suppressing signals by the speaker
  – Back channel signals by the auditor.
• Interactional rules
  – Prohibition of interruption
  – Requirement for properly timed signals
  – Acknowledge understanding
  – Confirm continued attention of the listener
Sources of data

- Interviews
  - Strong intrinsic motivation
  - Avoids more artificial topic for discussion
- Videotaping
  - Single camera placement
  - Recording less subtle expressions, smiles and grimaces
  - High quality monophonic audio track was obtained.
- Transcripts (more in next slide)
  - Requirements- max behavioral breadth and continuity
  - Depends on view of wealth of communication engaged
Transcripts -

• Phonemes
  • Segmental phonemes - pronouncing syllables within framework of English sound system
  • Supra segmental phonemes - intonation, stress, pitch and juncture

• Para language
  • Vocal behaviors not part of sound system of language.
  • Intensity, pitch height, extent.

• Body motion
  • Head gestures and movements, Shoulder movements
  • Facial expressions Hand gestures and movements
  • Foot movements Leg movements
  • Postures and posture shifts
  • Use of Artifacts, pipe, Kleenex, papers and clipboard.
Turn taking mechanism

• Rules delimit appropriate responses by participant
  – Rules and signals establish empirical expectations w.r.t turn taking activities.
• Speaker – claims speaking turn any moment
• Auditor – does not claim speaking turn.
• Simultaneous turns – simultaneous talking, and no contention
  – Turn taking Mechanism is not designed to resolve simultaneous turns
• Turn yielding
  – Rules
    • Turn-yielding signal is often met with turn taking signal
    • But if not, then results in contention, or simultaneous talking.
• Signal
  – Turn yielding signal, 6 discrete behavioral cues
    • Intonation : any pitch level terminal junction combination.
    • Paralanguage : Drawl : drawl on final syllable.
    • Body motion
    • Socio centric sequences,- appearance of one of several stereotyped expressions, Ex: “but uh”, “or something” or “you know”
    • Paralanguage : pitch/loudness
    • Syntax: completion of grammatical clause
• Attempt-Suppressing signal
• Rules
  • Speaker maintains turn for him
  • Speakers hands being engaged in gesticulation
  • Self and object adaptors
  • Dropping of the gesticulating hand
• Back-Channel communication
  – Rules
    • Messages as “mm-hmm”
    • Head nods from auditor.
  • Signals
    – Has large and complex set of signals.
    – Sentence completions.- auditors complete a sentence speaker begins
    – Brief requests for clarification
    – Restatement in few words of an preceding thought by a speaker
Coordinating through gaze

- Formalizing common pattern in a computational model of gaze-coordinated turn taking
- Dialogue acts
- Linguistic actions – realization of acts through observable communicative behaviours
• Two sequences useful for turn-taking
  – Mutual-break – after finishing conversation with utterance.
    • Used when conversation proceeds smoothly
  – Mutual-hold – recipient begins speaking without immediately looking away.

• Gaze actions signal their intention to give or keep the turn

• Gaze is one of many indicators of turn-taking behavior
Utterance segmentation and turn-taking

- Utterance is when user speaks and ends when user ceases to speak.
- Detection is on triggers on silence/non-speech.
- Nailon - online real-time prosodic analysis tool
- Finding the end of utterance is important in turn-taking mechanism.
Application of utterance detection

- Spoken language understanding
- Topic detection
- Information retrieval
- Interaction control
- Turn-taking
- Back-channeling
- System barge-in
• **Utterance units**
  – Humans use prosodic boundaries to delimit speech during speaking and listening
  – End-of-utterance (EOD) detectors rely on silence threshold
  – Problem – during spontaneous speech – silent pauses also within segments, called utterance units.
    • Humans can distinguish, but how would the machine ???
• Implementation was done by Jens Edlund, Heldner and Gustafson.
• Segmenting speech into pause bounded utterance units
  – Augmented end-of-utterance detection along with boundary tones.
  – mid level boundary tones used to single out internal pauses from final ones.
  – Voice Activity Detection (VAD) used to discriminate speech from non-speech.
• Limitations
  – Acoustic analysis was not tuned for children.
  – Children speech is different than grown-ups.

• Future scope
  – Add more features, and test on greater variety of data
  – Merge information by prosodic analysis with other semantic interpretations, dialogue context, etc.
3. Conversation acts in task oriented system

- Grounding - establish a mutual understanding or common ground of conversation content
- Ex: TRAINS Project, having manager (M) talking to system(S) to accomplish a task in that domain.
- The domain knowledge is a must in task oriented acts.
- Pervasiveness of acknowledge and agreement signals – striking features of task-oriented dialogues.
Assumptions and tradeoffs

• Each Utterance encodes a single speech act
• Speech acts are single agent plans. Listener is passively present
• Utterances are heard and understood correctly
• Tradeoffs: utterances can be misunderstood
  » Visual cues, head nodding, and continued eye contact are not taking into account, which is general in human nature.
  » Presence of multiple agents - as this can't be stand alone system and has to be part of a framework.
Conversation acts

- 4 levels of actions for maintaining the coherence and content of conversation.
- 1. Core speech acts – Discourse Unit acts
  - Inform
  - Request
  - Promise

- 2. Argumentation Acts
  - Elaborate
  - Summarize
  - Clarify
  - Q&A
  - Convince
  - Find-Plan
Argumentation acts
• 3. Grounding Acts
  • Initiate
  • Continue
  • Acknowledge
  • Repair
  • ReqRepair
  • ReqAck
  • Cancel
4. Turn-taking Acts

- Keep-turn
- Release-turn (assign-turn)
- Take –turn

- Certain sound patterns, such as "uhh", seem to carry no semantic content beyond keeping the turn.
- Pauses are opportunities for anyone to take the turn.
- Filling pauses with “uhh”, signals desire to keep turn, instead of “release-turn”
How to recognize my conversation acts??

- Monitoring speech input for certain set of features
- Can be overridden by evidence from other features
- Can be done through **context** saved through states accumulated in tracking a discourse unit through a sequence.
• Conversations are primarily concerned to
  – solve a problem
  – Gather Information
  – Social goals.

• Recognizing Turn-taking acts.
  – Dependent on social setting
  – Determined using the same channel as that of system that is regulating.
• Local initiative – other party is under obligation to speak or respond to
  • Questions
  • Requests
• If no obligations, then higher level goal and expectative will derive the local control
Finite state turn-taking model

• Model to control turn taking behavior of conversation agents
  • Uses cost matrix
  • Decision theoretic principles

• Knowing *what* to say is equally important as to knowing *when* to say

• Most of models caters to
  – hand-coded expert knowledge
  – And not data-driven optimization
In this model....

• Builds upon previous work on finite-state model of conversational floor
• Simple and general
• Grounded in decision theory
• Lends well to data-driven optimization
Jaffe and Feldstein studied

- Mean duration of pauses
- Switching pauses
- Simultaneous speech
- Vocalization in recorded dyadic conversations.

- Proposed first order Markov models to capture alternation of speech and silence in dialog.
6 finite state model
• Turn transitions with gap

\[ \text{SYSTEM}^{(R,W)} \rightarrow \text{FREE}_S^{(W,G)} \rightarrow \text{USER} \]

\[ \text{USER}^{(R,W)} \rightarrow \text{FREE}_S^{(W,G)} \rightarrow \text{USER} \]

• Turn transitions with overlap

\[ \text{SYSTEM}^{(K,G)} \rightarrow \text{BOTH}_S^{(R,K)} \rightarrow \text{USER} \]

\[ \text{USER}^{(G,K)} \rightarrow \text{BOTH}_U^{(K,R)} \rightarrow \text{SYSTEM} \]
• Failed interruptions

\[\text{USER}^{(G,K)} \rightarrow \text{BOTH}_U^{(R,K)} \rightarrow \text{USER}\]

\[\text{SYSTEM}^{(K,G)} \rightarrow \text{BOTH}_S^{(K,R)} \rightarrow \text{SYSTEM}\]

• Time outs

\[\text{SYSTEM}^{(R,W)} \rightarrow \text{FREE}_S^{(G,W)} \rightarrow \text{SYSTEM}\]

\[\text{USER}^{(W,R)} \rightarrow \text{FREE}_U^{(W,G)} \rightarrow \text{USER}\]
Cost of Turn-Taking Actions

- **CS** is the cost of interrupting a system prompt before its end when the user is not claiming the floor (false interruption)
- **CO(t)** is the cost of remaining in an overlap that is already **t** ms long
- **CU** is the cost of grabbing the floor when the user is holding it (cut-in)
- **CG(t)** is the cost of remaining in a gap that is already **t** ms long
Decision Theoretic Action Selection

- the optimal decision at any point in time is the one that yields the lowest expected cost, here the expected cost of action $A$ is:

\[ C(A) = \sum_{S \in \Sigma} P(s = S|O) \cdot C(A, S) \]

- Where $\Sigma$ is the set of states, $O$ are the observable features of the world, and $C(A, S)$ is the cost of action $A$ in state $S$
• Finite state Turn-Taking Machine relies on three core elements
  • A non-deterministic finite-state machine
  • A cost matrix that models the impact of different system actions
  • A decision-theoretic action selection mechanism
Should I bid to take my turn.....

• Turn-taking framework for spoken dialogue system which bid for turn.
  – Observation in other models
    • Position for Possible turn release is predictable
    • Emphasis on release detection for transitions.
    • Use pause threshold, and predict release-turn using lexical cues.
    • Other uses Reinforcement learning to tailor onset of system utterances.
  – But, Question is who will resolve turn-conflicts among these None
• Bidding approach, for turn-conflict resolution, is negotiative framework – turn-bidding model
  – Useful, when amount of time to speak was reduced, conflicts grew.
  – This markedly increases value of utterance and urgency to speak.
  – Utterances overlap due to competing interest of conversant in taking turn
• Common point- people do not wait for their turn-release cues.
  • Common among women :P
  • Opponents are suppressed due to insufficient conversational importance. Ex: “the house on fire” as conversation would be more important than opponent taking turn
Other approaches

1. Keep or release turn Approach
   - Works well when they have smooth transition of release-turn
   - Using non-syntactic cues such as gestures, falling pitch, changing gaze direction
   - Usually some system uses pause-threshold in some models.
   - Some system attempt to estimate probability of TRP occurrence, on syntactic and prosodic cues.
   - Goes awry when listener attempts to take tht run without turn-yielding cues.
• 2. Turn Resolution
  – Detect turn-conflict.
    • Discrepancy of mutual belief between conversants
    • Easy when previous turn-holder immediately releases the turn.
    • But can’t predict when both compete for turn ........
  – Turn-Conflict resolution
    • Increasing strength of cues,
    • relatively increasing the volume
    • Gives more turn-taking cues than turn-releasing cues
Turn-Bidding model

• 3. Psycholinguistic Framework
  – Modelled on
    • Individuals interest in having turn
    • Importance driven approach
    • To remedy some understanding discrepancy.
    • Pauses – first one to speak at the onset time of utterance.
• Computational framework
  – Implementing the psycholinguistic concept above
  – Bids of various levels are used
    • Shorter, short, mid, long, longer (strongest to weakest)
    • If the bid is tied, winner is decided randomly
  – Reinforcement learning is used
  – Develop dialogue system – to assign importance to utterances, depending on system’s experience than on some possibly arbitrary designer decision.
• Turn bidding approach
  – Is efficient
  – Is crucial for conversational success
  – Couple high-valued utterances with short onsets, uses Reinforcement learning to global dialogue importance.
  – Can model agents with more intricate patterns of interactions
  – Entire cost of dialogue is optimized, not just factors for turn transition.
  – Improvements can be made in multi-agent environ
Turn-taking failures

• Usually system have fixed, turn-taking strategy
• Depends on length of pause after utterance.
• Others have, turn based on content and prosody of user utterance.
• What if these fails and we don’t know what to speak immediately?
• Usually humans stop utterances in certain situation – this helps.
• Errors in speech interval detection
• Misrecognition of users intention to release a turn.
• Overlap user and system utterances
• *Discontinuations* – hard to detect, as it has only word fragments, and not grammatical.
Remedial measures

• Discontinuation detection
  – Word error rates are very high
  – Out of grammar utterances.
  – Prosody recognition

• Overlapping information used for predicting speech recognition performance degradation
Turns in Multiparty conversation

• Challenges
  – To reason about source and target of utterances
  – State and dynamics of the floor
  – Participant may speak to another participant, or system, or contribute in discussion, and wait for others to pitch in.
  – System cannot take floor- though it can predict end-of-turn
  – Inferences to be made about
    • multiparty conversation
    • Reasoning under uncertainty about possible outcomes
    • Tradeoff between acting and waiting for additional info
    • Take into account its own delay in perception and rendering pipelines of system.
Any remedies

• Decision theoretic approach
  – Highlight opportunities and directions
  – Moving from heuristics to principles decision policies
  – Inferences about conversation dynamics and system processing delays.
  – Explore behaviors of expected-utility policies and compare with heuristic procedures.
Various two-parties approach

- Turn constructional units in basic model, separated by transition relevance places.
- Gaze, gesture and non-verbal communication channels for turn talking regulation.
- Usually have dyadic settings, usually with 2 ppl.
- Machine learning employed, plus prosodic, syntactic and semantic features to predict turns.
- Bidding approach (as seen earlier).
- Finite state machine and decision theoretic approach for grabbing floor (as seen earlier).
Reference model – Turn-taking model

- Components
  - Conversational floor - participant ratified to speak owns it.
  - 4 Floor management actions
    - Hold action – performed by one who currently has floor.
    - Release – releasing the floor to someone else
    - Take – those who don’t own floor, can take it.
    - Null - participants indifference, for no claim

- Sensing component - estimate the current speaker
- Decision component – select floor management actions
- Dialog management layer – generates systems semantic contributions
- Turn-taking behavioral control component – renders actions into synchronized gaze, gesture and speech.
Heuristic approach (on turn taking model)

- Based on heuristics for sensing, decision making and behavioral control.
- Current speaker identified via handcrafted models, using sound source localization information from microphone array and from visual scene analysis.
- Additional rules – non understanding, and utterances.
- Also has non-verbal gestures,
  - avatars face turning towards other participant
  - Lifting eyebrows
- Drew incorrect inferences – floor released
  - Led to turn taking problems, as *floor transition battles* and *turn-initial overlaps*
Decision - Theoretic approach

• Making turn-taking decisions
• Continuously deliberate about key uncertainties and resolve tradeoffs bw waiting and taking floor
• To reduce floor battles and minimize gaps in conv
• Restrict subspace of turn-taking policies
  – System takes floor only if no one else speaks
  – Always takes floor once a silence longer than a specified duration is observed.
• Stochastic input processing delay(ID) also calculated by system
Decision - Theoretic approach....

- Takes into account
  - System processing delays
  - Cost function
  - Probabilistic model for tracking uncertainties
    - Potential outcomes
      - FloorTransitionToSystem
      - FloorTransitionToOther
      - FloorTransitionBattle
  - Cost Modelling
    - Two case: one where floor was released to system
    - Another, where floor wasn’t released to system
- Uses a learnt model for better floor release actions [See note]
Thank you!!!!