Acoustic and Visual Cues of Turn-Taking Dynamics in Dyadic Interactions

Bo Xiao[1], *Viktor Rozgić*[1], *Athanasios Katsamanis*[1]
Brian Baucom[2], *Panayiotis G. Georgiou*[1], *Shrikanth Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Psychology, University of Southern California, Los Angeles, CA, USA
`http://sail.usc.edu`[1], `baucom@usc.edu`[2]

Abstract

In this paper we introduce an empirical study of multimodal cues of turn-taking dynamics in a social interaction context. We first identify pauses, gaps and overlapped speech segments in the dyadic conversation dataset. Second, we define two types of measurements, Mean Equalized Energy (MEE) and Animation Level (AL) on the audio and video channels, respectively. Then, we verify the hypothesis that the speaker with higher MEE or AL is more likely to take the floor after silence or overlapped speech. The results suggest that both the vocal and visual movement energy offer useful cues towards inferring the intention of the interlocutor to grab the floor.

Index Terms: turn-taking, cues, equalized energy, motion vector

1. Introduction

Human behavioral signal processing is an emerging research domain which considers not only the physical properties of various signals, but also their meaning in a social and emotional context [1]. For instance, human dyadic conversations are one of the most common scenarios of social interaction that exemplifies an intricate choreography of give and take behavior between the interlocutors. Researchers in linguistics have studied turn-taking behavior for a long time. In a seminal work, Sacks *et al* [2] proposed a model for turn-taking with two components, namely the turn-constructional component and turn-allocation component. The first one deals with the construction of a turn with a certain linguistic unit-type, i.e, sentential, clausal, phrasal *etc.* The second one suggests two possible cases of next speaker allocation: either the current speaker selecting the next speaker or self-selection by any speaker. This model considers the intention of interlocutors to get the right to speak (called the floor). Furthermore, Yule suggested [3] that in a conversation, people compete for the floor just as in markets, where the floor is the scarce commodity. Having control of the floor is called a turn, and if the control is not fixed in advance, anyone can attempt to get control, which gives an interpretation of turn-taking. Researchers in social signal processing suggested [4] that turn-taking is the key to understanding conversational dynamics. Nevertheless, turn-taking in spontaneous conversation involves intricate timing, as pointed out by Shriberg [5]. Listeners project the end of the current speaker's turn and often begin speaking before the current speaker is finished. This results in a considerable amount of overlap in speech, and also supports the above assumption of people competing for floor. As we will show in the data we analyzed, turn transitions are facilitated by mainly pauses (intra-speaker), gaps (inter-speaker) and overlaps.

From an engineering point of view, considerable work has been done towards understanding turn-taking, typically with the goal of designing automated spoken dialog systems. The efforts are mainly in three directions. The first one focuses on the durational aspects. Heldner and Edlund [6] studied the duration of pause, gap and overlap in several corpora, and found that the timing of turn-taking is less precise and more distributed, disagreeing with traditional "no-gap-no-overlap" claims. The second one explores the multimodal cues for predicting the speaker's or listener's intention of taking or yielding the turn. Cassell *et al* [7] studied the relation of the gaze behavior of interlocutors with turn-taking integrated with information structures. It was shown that the beginning of "themes" (what the utterance is about) are frequently accompanied by a look-away from the listener, and the beginning of "rhemes" (the contribution to the pool of knowledge in the conversation) are frequently accompanied by a look-toward the listener. Gravano and Hirschberg [8] first defined Inter-Pausal Unit (IPU) as a maximal sequence of words surrounded by silence longer than 50 ms, then studied seven different communication cues of the speaker showing turn-yielding attempt, including a falling or high-rising intonation at the end of the IPU, an increased speaking rate, a lower intensity level, and *etc.* Finally, they showed that the likelihood of a turn-taking attempt from the interlocutor increases linearly with the number of cues mentioned above conjointly displayed by the speaker. The third direction is to design spoken dialog agents on the system level where the goal is to model the turn transitions. Raux and Eskenazi [9] proposed a Finite State Turn-Taking Machine, where the behavior of the user and system is modeled by six states. The novelty relies on the non-deterministic transition of states, a cost matrix that models the impact of different system actions in different states and a decision-theoretic action selection mechanism. The design enables data-driven learning of the model. Bohus and Horvitz [10] designed a multi-party dialog system with three components: sensing conversational dynamics, real-time turn-taking decisions and rendering decisions into appropriate behaviors. The model was aimed towards a collaborative conversation setting. And the proposed floor shift was carried out with four management actions namely Hold, Release, Take and Null. To sum up, the research in this area studies how a human conduct conversations and designs system that mimics human behavior.

In this paper, we study two multimodal cues, speech energy and body movement, in dyadic conversations. We found empirically that during overlap, the speaker with higher energy is more likely to get the floor afterwards. Moreover, during

This work is supported in part by NIH & NSF & DOD.

overlap, pause and gap, if body movement is observed, the one with higher degree of body movement is also more likely to get the floor afterwards. In other words, the data we analyzed support the model that has suggested that people compete for the floor, and this kind of competition could manifest multimodally. We show that in contrast to Cassell *et al*'s work [7], although the motion vector extracted from the video recording as a body movement measure is not as precise as gaze, it provides valuable information about the intention of the interlocutors, and is quite easy to access. Earlier work [11] had shown the usefulness of motion features directly derived from videos in characterizing fluency in spoken child-machine interactions.

The dataset we adopted for this study is a multimodal dyadic interaction database recently collected at USC, which includes audio, video and motion capture modalities. The interaction scenario concerned conflict between couples or friends in which college students role-played. This distinguishes our work from [8] where visual information was absent and the scenario was collaborative.

In the following we introduce the dataset and manual annotation in Section 2, then the method of feature extraction is described in Section 3. The results of analysis are given in Section 4, and we conclude with future work in Section 5.

2. Dataset

In this section we introduce the dataset [12]. The sessions are dyadic interactions based on unscripted role-playing on conflictual topics, such as cheating in relationships, arguing over a drinking problem *etc.* Two groups of nine topics were designed to fit same sex participants acting as friends, and opposite sex participants acting as couples. In order to get interactions that are as realistic as possible, participants were provided with the topics a few days before the recording. On the site of collection, participants were allowed several minutes to prepare for a topic, where they could exchange ideas and make up a story. They were also encouraged to bring life experiences to make the conversation vivid. The total length of collection is about 3 hours of several 3 to 10 minutes sessions. During the collection the two participants sat on a couch side by side.

The hardware in the collection included 10 HD Flea 2 cameras (30fps), 12 sensor Vicon motion capture system (120fps), three 4-microphone T-arrays, two lapel microphones and one shotgun microphone (48kHz) without dropped frames. Two cameras were near-view of the subjects and the rest were far-view ceiling cameras, all in resolution 1024×768. Cameras were synchronized by Firewire bus, while microphones were connected to two daisy-chained 8-channel MOTU-896 devices. Audio and video timelines were aligned to the precision of a single video frame.

In order to get a reliable ground truth of turn-taking, we manually segmented the speaker activity on the two lapel microphone channels using the Transcriber software. The *joint state* of the dyadic conversation is obtained by combining the segmentation on two channels. Four possible states are assigned to every 10 ms sample on the time axis, namely silence (SIL), speaker one active (S1), speaker two active (S2) and overlapped speech (OVL). We reject inactive segments on the lapel channel shorter than 200 ms, because it becomes unreliable and fuzzy due to hesitation and stop between words. Nevertheless for the final joint states, the duration of SIL might be small.

For this study we analyzed 35 sessions with total duration of 160 minutes. The data streams from the two lapel microphones and the two near view cameras were used for feature extraction.

3. Feature Extraction

3.1. Extraction of Equalized Energy

The audio signals from the two lapel microphones are used for computing short time energy. For both channels, we apply a rectangular window with 20 ms length and 10 ms shift to the original audio signal, and compute L_2 norm of the sample vector within the window. This gives us the raw energy E_1 and E_2 for speaker 1 and speaker 2 respectively. However, due to the gain difference of the two devices, E_1 and E_2 are not directly comparable. In order to equalize the energy, we assume that when the joint state is SIL, the two microphones would pick up environmental noise (e.g. from air-conditioners, computers) at the same energy level. Then the mean value of E_1 and E_2 upon SIL segments can be computed as N_1 and N_2. Dividing E_1 and E_2 by N_1 and N_2 respectively, we achieve equalization in the sense that the mean values of the environmental noise on the two channels are equal. Note that potentially microphones can exhibit a DC bias that can affect measured energy levels however we analyzed the collected signals and for our microphone setup these biases are negligible. The equalized energies are denoted by $\tilde{E}_i(t) = \frac{E_i(t)}{N_i}$ and $i = 1, 2$ for time index t.

3.2. Extraction of Motion Vectors

We extract Motion Vector (MV) features from MPEG 4 Xvid coded videos of the two front view cameras associated with the two subjects using the FFmpeg library. The MVs are computed as a part of the MPEG 4 video encoding procedure and embedded in the bit-stream. For our dataset, MVs are estimated on every 16×16 pixels block, giving a 64×48 grid of MV field for a 1024×768 image frame. Since the MVs are mostly zeros, we store the result sparsely with each entry being a 4 dimensional vector (X, Y, dX, dY), where $0 \leq X < 64$ and $0 \leq Y < 48$ are the column and row indices of the MV, and dX, dY are integer-valued speed on the horizontal and vertical directions. Intuitively the MVs are results of block matching among consecutive frames, so they best represent the motion if the object's shape does not change, and the distance of movement is adequate to fit the block size. Plotting MVs on top of the video demonstrates that common hand and head movements are well captured by the MVs in terms of the location and direction, and to a certain degree of the velocity. However, the MV is insensitive to motion of small objects like eyes and lips. Also it is not able to distinguish specific objects of interest from others, like on a region of clothes and background.

It is necessary to filter the MVs to get better representation of body movement. First, empirically we eliminate MVs that are with too small or too large velocity. If $|dX| < 2$ and $|dY| < 2$, the entry is removed from the set of MV. And if $|dX| \geq 10$ or $|dY| \geq 10$, the entry is also removed. Small MVs are usually caused by lighting and camera noise, while large MVs usually exceed reasonable range of body movement and are subject to accidental matching of blocks. Next, we apply a simple background-foreground separation to eliminate "false" motion caused by noise. The foreground is obtained by thresholding the difference of a new image and the background, while the background is updated by a running average of new image and old background. The percentage of foreground pixels on every motion vector block is counted, and motion vector of a block with a percentage lower than another threshold is rejected. Finally we had to setup a rectangular "region of interest" (ROI) because sometimes a part of the other subject appears crossing the boundary of the scene. These movements are excluded to

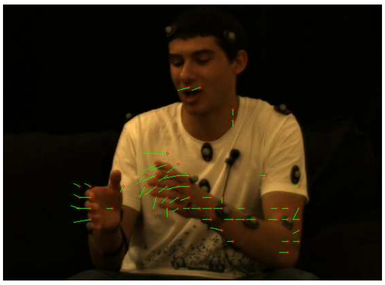

Figure 1: Example of MV on one frame (The image is clipped).

avoid interference of the two subjects. For a relatively small number of sessions, these ROI are manually set. One example of MV field is demonstrated in Figure 1.

When the filtering is done, we get a sequence of sets of MVs for the sequence of frames, although the set could be empty. For this paper, we only use the count of MVs on each frame as a measure of how animated the subject is, denoted $M_i(t) = |\{\text{MV of speaker i on frame t}\}|$ and $i = 1, 2$.

4. Analysis and Results

4.1. Distribution of Overlap, Pause and Gap

Based on the segmentation data of the joint states (defined in Sec. 2), we find that turn transition is facilitated mostly by gap and overlap, as argued in [6]. The counts of transition from one segment to the next are shown in Table 1 where the rows correspond to the current state and columns correspond to the next. Note that transitions directly between S1 and S2 are rare, as are direct transitions between SIL and OVL. We can see that in a conflictual scenario, occurrences of overlap are more prevalent.

We selected the OVL segments that are not proceeded or followed by SIL. In other words we only look into the OVL segments where the previous and next segments are either S1 or S2. The rest of the cases are too scarce to analyze. As a result, 1523 occurrences of OVL segments are collected. 568 samples correspond to the case that the speaker before and after the overlap are the same (OVL-Intra), and 955 samples have a different speaker before and after the overlap (OVL-Inter).

Similarly, all SIL segments (either pause or gap) are picked out, excluding silence at the start and end of each session, pauses shorter than 200 ms, as well as those segments having OVL as either previous or next segment. As a result, 2959 samples are collected with 1623 pauses and 1336 gaps. The histogram of duration of pause, gap and overlap is plotted in Figure 2. As we can see, OVL-Intra is not rich in short duration because normally the other speaker cannot make very short utterances. For the other 3 groups, frequencies of occurrence almost always decrease with the increase of duration. These observations are also consistent with [6].

	SIL	S1	S2	OVL
SIL	0	1535	1556	15
S1	1439	0	37	816
S2	1655	27	0	725
OVL	9	731	815	0

Table 1: Count of speaker state transition.

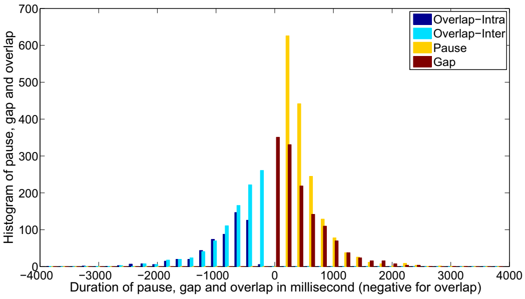

Figure 2: Duration histogram of pause, gap and overlap.

4.2. Multimodal Turn-Taking Cues

In order to compare the energy level of two speakers during overlapped speech, we define the Mean Equalized Energy (MEE) value for speaker i on the kth OVL segment as $\text{MEE}_i^k = \text{mean}(\log\{\tilde{E}_i(t)|t \in T_{\text{OVL}}^k\})$, where T_{OVL}^k is the corresponding time index set. To verify the hypothesis that the speaker with higher energy is more likely to continue speaking, we compare the speaker index having larger MEE with the speaker index after OVL segment. As a result, the two indices are equal with a chance of 0.66. This suggests that the speaker with higher energy during overlapped speech is approximately twice as likely as the other one to take the floor. In addition, MEE of the speaker getting the floor after the current segment (SPK-WIN) versus that of the other speaker (SPK-LOS) is plotted in Figure 3. As we can see, more samples are placed under the line connecting $(0, 0)$ and $(1, 1)$.

For the visual modality, Animation Level (AL) is defined as the log scale of average number of motion vectors in the segment of interest. Specifically, let $\text{AL}_i^k = \log(\text{mean}(\{M_i(t)|t \in T^k\}) + 1)$, where $i = 1, 2$ and T^k is the time index in interest. Note that video is in a different frame rate from audio, so the index range is computed separately. As their may be no motion in a frame, 1 is added to the mean to avoid taking logarithm of zero. Normalized histograms of AL for both interlocutors combined on SIL and OVL segments are plotted in Figure 4. It shows that during OVL segments, the AL is distributed more to the higher end, meaning that the interlocutors tend to be more animated during OVL then SIL.

To investigate the relation of having higher AL and turn-grabbing after OVL segments, we plot the AL of SPK-WIN versus that of SPK-LOS. The resulting plot in Figure 5 is split to 6 disjoint regions: *(i)* At the origin, neither speaker exhibits animation; *(ii)* SPK-WIN has non-zero AL while SPK-LOS has

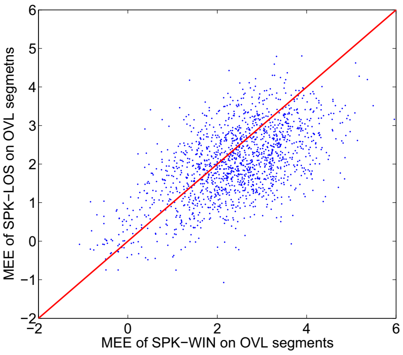

Figure 3: MEE of SPK-WIN vs. SPK-LOS on OVL segments.

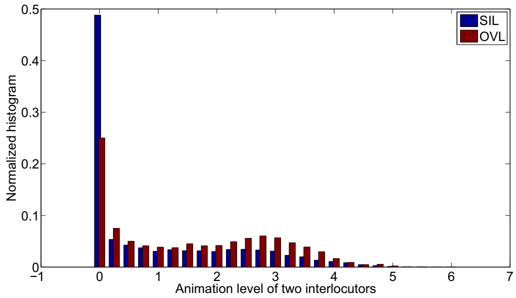

Figure 4: Normalized histogram of AL during SIL and OVL segments.

zero AL; *(iii)* The opposite case of (ii). *(iv)* Both speakers have non-zero AL, and the SPK-WIN's AL is higher. *(v)* The opposite case of (iv). *(vi)* Both speakers have equal non-zero AL. The counts for these cases are labeled in the figure. We note that region (i) is quite small, and region (ii) and (iv) both outnumber region (iii) and (v). We see that region (ii) and (iv) are 60% of the overall occurrences (excluding cases (i) and (vi)), hence the higher-animated speaker is more likely to win the floor.

Similarly, the analysis is applied to SIL segments. The AL of SPK-WIN versus that of SPK-LOS is plotted in Figure 6. The resulting figure is also split in the same way. Here the size of region (i) is relatively larger than in OVL segments. And a similar trend, with a ratio of 0.64, is presented for regions (ii) and (iv) versus overall cases (excluding (i) and (vi)).

The findings mentioned above support the hypothesis that the interlocutor with a higher energy level or higher animation level is more likely to get the floor. Moreover, when the two cues are observed jointly, we found that the interlocutor with both higher MEE and higher AL gets the floor with a probability of 0.74.

5. Conclusions

In this paper we conducted an empirical study on turn-taking behavior, focused on the audio and visual cues that represent the intention of the speaker to take the turn. The Mean Equalized Energy and Animation Level measures were defined, and comparison of these cues of the two interlocutors shows that the one with higher energy or animation level is more likely to get the floor. Based on these findings, we suggest that in human dyadic conversations, higher vocal energy and visual movement energy are a means of conveying an attempt to grab the floor. By monitoring these cues, a behavioral computing system shall have better understanding of the intention or mental states of the subject. This could be beneficial not only to automated spoken dialog system design but also other behavioral informatics applications trying to infer the mental state of the subject, or understand the interaction between humans.

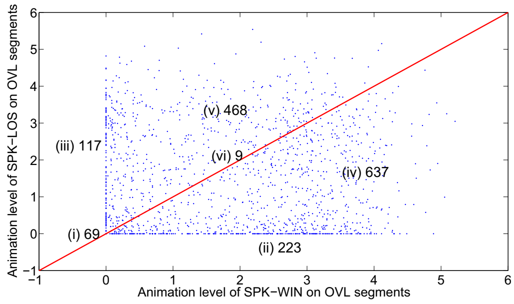

Figure 5: AL of SPK-WIN vs. SPK-LOS for OVL segments

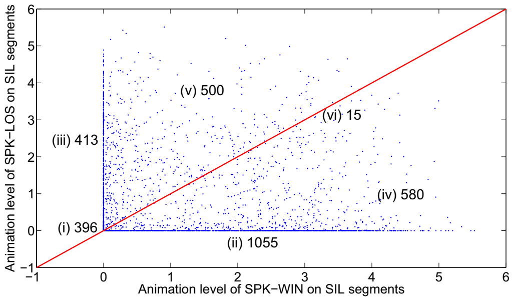

Figure 6: AL of SPK-WIN vs. SPK-LOS for SIL segments

The study is limited by the amount of the data and lacking of rich annotation on psychological state of the subject. Moreover, body movement and hand gestures are culturally related. In our dataset, the subjects are mainly American English speaking college students, so the claim might not extend to cross-cultural conversations.

For future work, finer annotation of the intention of speakers over pause, gap and overlap regions might be helpful to inform the observation-based behavioral computing for interpersonal interactions.

6. References

[1] M. Black, A. Katsamanis, C. Lee, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proc. InterSpeech*, Makuhari, Japan, Sep. 2010.

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

[3] G. Yule, *Pragmatics*. Oxford University Press, 1996.

[4] A. Vinciarelli, H. Salamin, and M. Pantic, "Social signal processing: Understanding social interactions through nonverbal behavior analysis," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, 2009, pp. 42–49.

[5] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. Interspeech*, 2005.

[6] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, 2010.

[7] J. Cassell, O. Torres, and S. Prevost, "Turn taking vs. discourse structure: How best to model multimodal conversation," *Machine Conversations*, pp. 143–154, 1999.

[8] A. Gravano and J. Hirschberg, "Turn-yielding cues in task-oriented dialogue," in *Proceedings of the SIGDIAL*. ACL, 2009, pp. 253–261.

[9] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in *Proc. NAACL-HLT*. Association for Computational Linguistics, 2009, pp. 629–637.

[10] D. Bohus and E. Horvitz, "Computational models for multiparty turn-taking," MSR-TR-2010-115, Microsoft Research, Tech. Rep., 2010.

[11] S. Yildirim and S. Narayanan, "Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio–Visual Information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 2–12, 2009.

[12] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, and S. Narayanan, "A new multichannel multimodal dyadic interaction database," in *Proc. InterSpeech*, Makuhari, Japan, Sep. 2010.