GETTING FROM A TO B: COMPLEXITIES OF TURN CHANGE AND RETENTION IN CONVERSATION

Emer Gilmartin[1,2], Marcin Włodarczak[3]

[1]ADAPT Centre, Trinity College Dublin, Ireland; [2]INRIA Paris, France

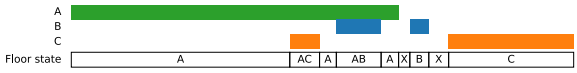
[3]Department of Linguistics, Stockholm University, Sweden
gilmare@tcd.ie, wlodarczak@ling.su.se

ABSTRACT

Spoken interaction proceeds as a series of contributions or utterances from participants. The question of who 'has the turn' and how turns are managed is fundamental to understanding of all genres of spoken interaction, and to conversation in particular, as turntaking is managed locally by participants. We use floor state, descriptions of who is speaking or not at any time, to study between and within floor state transitions (BST and WST), sequences of activity from one substantial stretch of single party speech to the next, analagous to turn change or retention. We explore the patterns of speech and silence in the Switchboard corpus of over 2000 casual two-party telephone conversations using an annotation and analysis paradigm that has previously exposed interesting features of multiparty talk. We find that over half of the transitions observed in Switchboard involve more than one interval of silence, a phenomenon first noted in multiparty talk. We also find that although the transition distribution in Switchboard's dyadic conversations broadly follows patterns found in multiparty talk, there are fewer complex transitions observed. **Index Terms**: dialogue, corpus studies, turntaking, human-computer interaction

1. INTRODUCTION

The arrangement of participant contributions to spoken interaction, particularly during turn change or retention, is of fundamental interest to studies of speech and dialog, and has become a topic of interest in speech and language technology with increasing uptake of natural language interfaces. Consequently, patterns of speech and silence in interaction have been widely used in large scale analyses of dialogue, often involving machine learning, to infer a wide range of information about spoken interaction – examples include predicting speaker activity from conversation history [1, 2, 3], inferring speaker and context characteristics [2, 4], or extracting personality traits of speakers in dyadic and multiparty interaction [5, 6, 7]. However, in much of this work, the focus is on optimising models' predictive power over entire conversations rather than on elucidating the specifics of how conversation works locally.



A . . . cameras. You're just not normally as aware of it.
C Yeah
B It's true.
B Yeah.
C There's two on every bus.

Figure 1: An excerpt from a casual conversation in English corresponding to a between-speaker transition, **A_AC_A_AB_A_X_B_X_C**, from speaker A to C with seven intervening intervals (**AC**, **A**, **AB**, **A**, **X**, **B**, **X**). *Top*: A diagram showing temporal organization of individual speakers' contributions (represented as color bars) and the resulting floor states. *Bottom*: A simplified transcript. Speakers' contributions are color-coded for consistency.

In this paper we examine how interaction progresses in the Switchboard corpus [8], a large collection of dyadic phone conversations, and compare the results to patterns previously observed in multiparty conversation. To do this we use an annotation and analysis paradigm developed and previously employed to analyse a range of smaller corpora of multiparty dialogue. The analysis is based on the concept of *floor state* - who is speaking or silent at any moment during interaction. By annotating *floor state intervals*, stretches of time during which a particular floor state holds, we can analyse *floor state transitions* or sequences of contiguous floor states. While these transitions can record any sequence of conversational activity, we are particularly interested in sequences of activity between longer stretches of single party speech, as these transitions can throw light on how turn-taking is accomplished. In order to track changes in floor possession, we further categorize the sequences as

either *between speaker transitions* (BST) or *within speaker transitions* (WST). In WST, the speaker on either side of the transition is the same, analogous to a turn retention, while in BSTs, the single party speech bounding the transition is produced by different speakers, analogous to a turn change. Figure 1 shows an example of a short exchange from a three-party conversation to illustrate the labelling scheme and concepts. The example involves nine floor states – solo speech (**A**, **B**, **C**), overlaps (**AC**, **AB**) and general silence / nobody speaking (**X**). Existing data-driven approaches to turn-taking could treat this stretch as a series of four transitions: **A_AC_A** from A to A, **A_AB_A** from A to A, **A_X_B** from A to B, and **B_X_C** from B to C. However, looking at the transcript and the speech patterns, it seems more likely that the *longer* stretches of solo speech (**A**, **C**) delimit a single more complex transfer of floor possession from speaker A to C. In the analysis described below, we use floor state transitions to discover these larger conversational structures which can be otherwise overlooked by large-scale corpus studies.

2. BACKGROUND

Previous work on floor state transitions in several collections of multiparty dialogue has identified interesting similarities – for all multiparty corpora studied less than half of between and within speaker transitions are accomplished with a single interval of silence or overlap, indicating that turn change and retention usually involve more complex sequences of speech (either in the clear or involving overlap) and silence. It has also found high levels of uniformity in the most common WSTs and BSTs found in different languages and settings [9, 10, 11, 12, 13]. In a study of speaker transitions in multiparty three-, four- and five-party casual conversation in English, one-interval transitions were the largest category, with a higher proportion of one-interval transitions in WST, perhaps reflecting breath pauses or single backchannels during monologic stretches [10]. Nevertheless, the majority of transitions involved more than one intervening interval to complete. In work on English, Estonian and Swedish three-party casual face-to-face conversations [14], similar patterns were observed, with over 95% of transitions completed in 15 or fewer intervening intervals, and over 65% involving more than one interval, with a higher likelihood of more complex transitions in the BST category. Additionally, the minimum duration threshold imposed on single-party speech affected the transition type label and number of intervening intervals assigned, with transition type label (WST or BST) changing for 28.29% of transitions if the right duration hand threshold of one second was removed. These results imply that turn change and retention cannot be comprehensively modelled without consideration of both left and right hand context - if the right hand minimum duration threshold is relaxed, most transitions revert to one-interval and the complexity of what actually happens in real conversation is lost. Subsequent work investigated the more common transitions in three-party casual and task-oriented data [12], finding considerable complexity and growing incidence of participation by more speakers with transition length. Even though very many different transition labels were present in the data, a small subset of 17 labels (4 one-interval and 13 three-interval) accounted for most of floor state transitions occurring.

Having replicated these results across a number of multiparty corpora, below we (1) analyse the Switchboard corpus of dyadic phone conversations to investigate whether these findings on multiparty talk also hold for the Switchboard conversations, and (2) compare the characteristics of within and between speaker transitions in Switchboard with those of other corpora analysed previously.

3. DATA AND ANNOTATION

The analysis was based on the Switchboard-1 Telephone Speech Corpus (Release 2) [15], comprising 2438 dyadic phone conversations between 542 speakers of American English (302M/241F). Speech and silence labels for each participant were derived from word level transcription [16], with all non-speech sounds suppressed to silence, resulting in a set of 520135 talkspurts (interpausal units comprising speech from a participant bounded by silence from the same participant) from 259 hours of conversation. The speech/silence labels were used to generate labels for each floor state interval in the corpus, which were combined to form transition labels, stretches of conversation starting with an interval of single party speech in the clear of at least one second in duration and continuing through the next one-second interval of single party speech in the clear encountered in the data. Each transition was classified as either between-speaker (BST) or within-speaker (WST), depending on whether it was bounded by speech from the same or two different speakers.

We used three-party dialogue data from

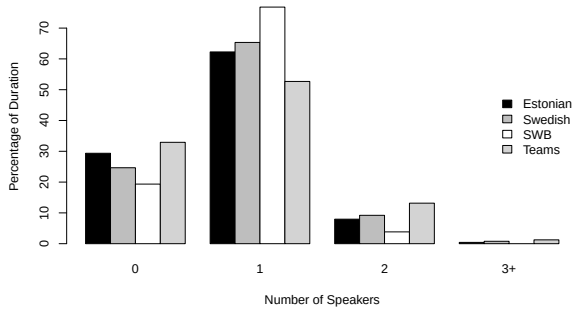


Figure 2: Distribution of silence (0 speakers), single-party speech (1 speaker) and overlap (2 or more speakers) as percentage of total conversation time for Switchboard (SWB) and for the English, Estonian and Swedish three-party corpora.

previous studies to compare with the Switchboard results. These data are three-party spontaneous conversations in Estonian [17] and Swedish [18], and collaborative conversational games in English [19]. All the data had been segmented manually. The three-party data set contained 22106 talkspurts in 9 hours and 51 minutes hours of conversation.

4. RESULTS

Results are first presented for the Switchboard corpus, and then contrasted with results from previous work on multiparty corpora.

Distribution of Speech, Silence and Overlap There were 1,042,516 floor state intervals identified in Switchboard – 415,552 silent (0 speaker) intervals accounted for 19.35% of conversation time, 520161 single-party speech for 76.82%, and 106,993 two-party overlap for 3.83%. The proportions of single-party speech, silence and overlap in Switchboard are shown along with the proportions in the three-party data for comparison in Figure 2. It can be seen that Switchboard has lower incidence of silence and overlap than the other data sets, and higher incidence of single-party speech in the clear.

Distribution of speaker transitions The Switchboard data set contained 256,655 speaker transitions in 259 hours of talk, an average of one every 3.7 seconds. In the three-party data, there was an average of one transition every 4.7 seconds.

Figure 3 shows relative frequencies for odd number interval transitions, further split between BST and WST classes, in the four data sets. Similar to the three-party corpora, the vast bulk (over 99%) of transitions in Switchboard comprised less than 16 intervening intervals. There are vanishingly few transitions involving even numbers of intervening intervals (47 out of 256,655). This low number is unsurprising as such transitions would involve smooth switches or simultaneous onset or offset of speech, which are very rare in conversation.

One-interval transitions are the largest class in Switchboard, and the frequency of transitions decreases with increasing numbers of intervening intervals: 47.72% of all transitions (41.65% of BSTs and 50.03% of WSTs) were accomplished with one intervening interval, 27.14% (24.77% of BSTs and 28.15% of WSTs) with two intervening intervals, and 12.86% (15.98% of BSTs and 12.16% of WSTs) with 3 intervening intervals.

Overall, 78.28% of transitions are WST, greatly outnumbering BST transitions. WST are particularly frequent in short transitions, accounting for 81% of one-interval transitions, 80% of three-interval, and falling with increasing numbers of intervals to 60% of 15 interval transitions.

Most Common Transition Sequences In the Switchboard data, we found all 4 possible one-interval transition sequences, two BSTs (A_X_B,A_AB_B) and 2 WSTs (A_X_A,A_AB_A). There were 16 three-interval transition sequences (8 BSTs, 8 WSTs), and 64 five-interval transition sequences (32 BSTs, 32 WSTs).

The most frequent transition types in Switchboard overall was A_X_A at 36.69% of all transitions, followed by A_X_A_X_A (13.1%) and A_X_B (6.55%). The most common WST was A_X_A, followed by A_X_A_X_A and A_X_B_X_A - all of which involve only single party speech and silence. The most frequent WST involving overlap was a three-interval transition, A_X_B_AB_A, with the one-interval overlap WST A_AB_A appearing as the 6th most common WST. For BSTs, the most common was A_X_B followed by A_AB_B. Transitions involving overlap were more frequent in BSTs than in WSTs.

5. DISCUSSION

The distribution of speaker transitions in Switchboard largely reflects the patterns found in the three-party data used for comparison (and also in four- and five- party data analysed in [11]. The largest category are one-interval transitions, even-number interval transitions are extremely rare, and the number of transitions drops off with increasing numbers of intervals. The proportion of one-interval transitions in Switchboard is greater than that seen in the three-party corpora, but still only accounts for 47.7% of all transitions, highlighting the fact that most transitions involve

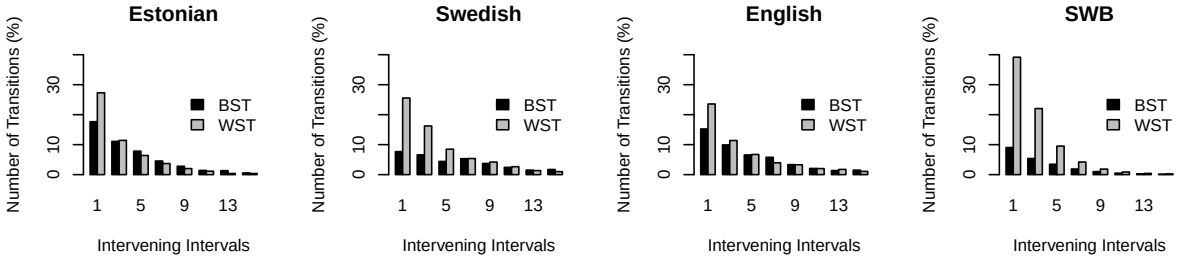


Figure 3: Distribution of between- and within-speaker transitions in Switchboard (SWB) and the English, Estonian and Swedish three-party corpora.

more than a simple single silence or overlap, even in dyadic phone conversations.

The split between within- and between- speaker transitions in Switchboard reflects that found in the three-party data in that there are more WSTs present. WSTs in Switchboard dramatically outnumber BSTs, more than in the three-party data. This could reflect long turns being taken in Switchboard, perhaps as a consequence of the fact that participants were strangers, or indeed, be a feature of telephone conversation.

The results on the distribution of specific transitions in Switchboard are very interesting to compare to those found on the three-party data used in this paper. The most frequent transition types in Switchboard was A_X_A at 36.69% of all transitions, followed by A_X_A_X_A (13.1%) and A_X_B (6.55%). In the 3 party data, the most common sequences overall were A_X_A (within-speaker silence) and A_X_B (between-speaker silence). The prominence of A_X_A_X_A in Switchboard reflects the high proportion of within-speaker transitions seen in Switchboard compared to the three-party data. For WSTs, the ranking in three-party data was similar to that in Switchboard, with both A_X_A_X_A and A_X_B_X_A (three-interval WSTs involving only single party speech and silence) appearing more frequently than one-interval WSTs involving overlap (A_AB_A). As with Switchboard, the top two three-party BSTs were one-interval silence and overlap (A_X_B and A_AB_B). Transitions involving overlap were more frequent in BSTs than in WSTs in both Switchboard and the three-party data.

Previous work on multiparty talk [12] has found that participation by all speakers becomes more likely in transitions involving more intervals, leading to quite complex and varied possibilities. However, in the dyadic Switchboard data, the number of transitions drops more rapidly as the number of intervals increases (see Figure 2).

Switchboard is also characterized by less silence and overlap and more speech in the clear than the three-party conversations. This may be due to the modality [20, 21] – on the phone, speakers may wait for their interlocutor to finish before commencing to speak, and may not give as much verbal feedback in overlap. It could also reflect differences between dyadic and multi-party talk.

In summary, our analysis of the 2400 Switchboard conversations has shown that more than half of all between and within speaker floor state transitions in these dyadic conversations involve more than one intervening interval of speech, silence or overlap between longer stretches of single party speech. This reflects previous results on multiparty spoken interaction, implying that turn change and retention even in dyadic phone conversations exhibit a level of complexity that is not covered by modelling them as a simple gap or overlap. This has implications for design of spoken dialog technology - while 'ping-pong' style turntaking may suffice for simple question-answer transactions, creating more convincing human-like interaction will require more nuanced modelling of turn taking.

6. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant Number 13/RC/2106. The work was also funded by Swedish Research Council project 2019-02932 *Prosodic functions of voice quality dynamics* to Marcin Włodarczak.

7. REFERENCES

[1] J. Jaffe, L. Cassotta, and S. Feldstein, "Markovian model of time patterns of speech," *Science*, vol. 144, no. 3620, pp. 884–886, 1964.

[2] B. Beebe, D. Alson, J. Jaffe, S. Feldstein, and C. Crown, "Vocal congruence in mother-infant play," *Journal of Psycholinguistic Research*, vol. 17, pp. 245–259, 1988.

[3] B. Beebe, J. Jaffe, F. Lachmann, S. Feldstein, C. Crown, and M. Jasnow, "Systems models in development and psychoanalysis: The case of vocal rhythm coordination and attachment," *Infant Mental Health Journal*, vol. 21, no. 1-2, pp. 99–122, 2000.

[4] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," in *Proceedings of SIGdial 2008*, Columbus, OH, 2008, pp. 148–155.

[5] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[6] L. H. Gilpin, D. M. Olson, and T. Alrashed, "Perception of speaker personality traits using speech signals," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. LBW514.

[7] M. Yu, E. Gilmartin, and D. Litman, "Identifying personality traits using overlap dynamics in multiparty dialogue," in *Proceedings of Interspeech 2019*, 2019, pp. 1921–1925.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1992, pp. 517–520.

[9] E. Gilmartin, K. Aare, M. O'Reilly, and M. Wlodarczak, "Between and within speaker transitions in multiparty conversation," in *Proceedings of Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 799–803.

[10] E. Gilmartin, M. Yu, and D. Litman, "Comparing speech, silence and overlap dynamics in a task-based game and casual conversation," in *Proceedings of ICPhS 2019*, 2019, pp. 3408–3412.

[11] E. Gilmartin, "Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis," Ph.D. dissertation, Trinity College Dublin, 2021.

[12] M. Włodarczak and E. Gilmartin, "Speaker transition patterns in three-party conversation: Evidence from English, Estonian and Swedish," in *Proceedings of Interspeech 2021*, 2021, pp. 801–805.

[13] E. Gilmartin, C. Saam, C. Vogel, N. Campbell, and V. Wade, "Just talking - modelling casual conversation," in *Proceedings SIGdial 2018*, Melbourne, Australia, 2018, pp. 51–59.

[14] E. Gilmartin, K. Aare, M. O'Reilly, and M. Wlodarczak, "Between and within speaker transitions in multiparty conversation," in *Speech Prosody 2020*, 2020, pp. 799–803.

[15] J. J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, CA, 1992, pp. 517–520.

[16] N. Deshmukh, A. Ganafileiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *Proceedings of ICSLP*, Sidney, Australia, 1998, pp. 1543–1546.

[17] P. Lippus, T. Tuisk, N. Salvestre, and P. Tiras, "Phonetic corpus of Estonian spontaneous speech." [Online]. Available: https://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech

[18] M. Włodarczak and M. Heldner, "Respiratory constraints in verbal and non-verbal communication," *Frontiers in Psychology*, vol. 8, p. 708, 2017.

[19] D. Litman, S. Paletz, Z. Rahimi, S. Allegretti, and C. Rice, "The teams corpus and entrainment in multi-party spoken dialogues," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1421–1431.

[20] L. ten Bosch, N. Oostdijk, and J. P. de Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *Proceedings of 7th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, 2004.

[21] F. Hammer, P. Reichl, and A. Raake, "Elements of interactivity in telephone conversations," in *Eighth International Conference on Spoken Language Processing*, 2004.