COMPARING SPEECH, SILENCE, AND OVERLAP DYNAMICS IN A TASK-BASED GAME AND CASUAL CONVERSATION

Emer Gilmartin [1], Mingzhi Yu [2], Diane Litman [2]

[1]ADAPT Centre Trinity College Dublin, [2]University of Pittsburgh
gilmare@tcd.ie

ABSTRACT

Spoken interaction occurs for many purposes and in many forms, and it is possible that there are differences in the distribution of features of different 'speech-exchange systems'. Greater understanding of the characteristics of different speech exchange systems is vital to the design of human-like artificial dialogue, and the need for clearer modelling has prompted our explorations into silence and overlap. We investigate these features in two different multiparty speech exchange systems – a collaborative board game and casual conversations of around one hour duration. We analyse speech activity at the end of intervals where one participant speaks in the clear for a second or more, and categorise patterns of overlap and turn change or retention. We compare and contrast these patterns between the game and phases of casual speech and report our results.

Keywords: multiparty dialog, pauses and gaps, genre

1. INTRODUCTION

Spoken interaction ranges from formal rituals through task-oriented practical exchanges, to casual or social conversation and is often divided into instrumental (task-based) or interactional (social) talk [4, 16, 12]. Instrumental dialog mediates practical activities such as service encounters (shops, doctor's appointments), information transfer (lectures), planning and execution of business (meetings), and collaborative or competitive games. Interactional or social talk does not seem to contribute to a clear short-term task, but rather to build and maintain social bonds, in interactions ranging from strangers briefly chatting at a bus-stop to friends 'hanging out' in 'a continuing state of incipient talk', [15], to stretches of smalltalk during business interactions.

Much early work on the dynamics of dialogue relied on natural or constructed corpora of two-party task-based exchanges, and results on such data may not transfer to other domains or multiparty spoken interaction. Early dialogue system researchers saw the complexity of dealing with social talk [1], and focused on task-based dialogue interactions, where lexical content drives task completion, conversation length is governed by task completion, and participants are aware of the goals of the interaction. Spoken dialog systems are moving beyond simple tasks into the realm of social interaction, strengthening the need for accurate modelling.

While instrumental interaction is delimited by task length, conversation is more open-ended and can extend for hours. Corpora of casual conversation have tended towards exchanges of 5-20 minutes, but research on casual talk has shown that after the first 8-10 minutes of interactive talk, conversation often settles into sequences of 'chat' and 'chunk' elements [6]. Chunks are segments where 'one speaker takes the floor and is allowed to dominate the conversation for an extended period' [6]. Chat phases are highly interactive with frequent turn changes, many questions and short comments, often occurring at the start of an interaction. Chat is often used to 'break the ice' among strangers involved in casual talk, [10]. As the conversation progresses, chat phases are interspersed with chunk phases. The 'ownership' of chunks seems to pass around the participants in the talk, with chat linking one chunk to the next [6]. In a study of workplace coffee breaks [18], fifty percent of all talk was classified as chat, while the rest comprised longer form chunks of storytelling, observation/comment, opinion, gossip, joke-telling and ridicule.

To improve understanding of the form and dynamics of multiparty social as well as task-based dialog, we contrast the distribution of speech and silence a multiparty collaborative (task-based) game mediated through spoken interaction, and in chat and chunk phases of multiparty casual conversation. We also investigate how turns end for a single speaker in the three conditions to explore turn change and retention. Below we de-

scribe our data and annotations, present our analyses, discuss our results, and conclude with an outline of future work.

2. DATA AND ANNOTATION

We use two datasets – the Teams Corpus of multiparty collaborative games, and CasualTalk, a dataset of multiparty casual conversations. The data used and the pre-processing steps taken for the current work are outlined below.

2.1. The Teams Corpus

The Teams Corpus [11] comprises 47 hours of multiparty interaction from 62 teams (35 three-person and 27 four-person), playing a collaborative board game – Forbidden Island. This game requires cooperation and communication among the players to win as a group. 213 native speakers of American English (79 males and 134 females) aged 18 years or older participated in the study for one session, playing two rounds (Game 1 and Game 2) of the game. Before each session, participants took a pre-game survey collecting personal information. After each game, the participant took a post-game survey to evaluate the team process. The audio recordings were segmented into interpausal units (IPUs) and transcribed manually using the Higgins Annotation Tools [17]. The speech label was applied to verbal and non-verbal vocal sounds such as laughter and sighs. The work presented in this paper is based on the 62 recordings of Game 1.

2.2. The CasualTalk dataset

The CasualTalk dataset is a collection of six 3 to 5 party conversations of around one hour each, drawn from the d64, DANS, and TableTalk corpora [13, 8, 5], all recorded in living room conditions or around a table. In each conversation participants were free to talk or not as the mood took them.

The data were segmented and transcribed manually at the intonational phrase (IP) level using Praat [3] and Elan [19]. The speech label was applied to verbal and non-verbal vocal sounds (except laughter) including contributions such as filled pauses, and short utterances such as 'oh' or 'mmhmm'. Laughter was annotated inline with speech. For this study, IPs were concatenated to IPUs, and annotated coughs, breaths, and laughter intervals were converted to silence. A total of 213 chat and 358 chunk phases were identified and annotated.

2.3. Working Dataset and Pre-processing

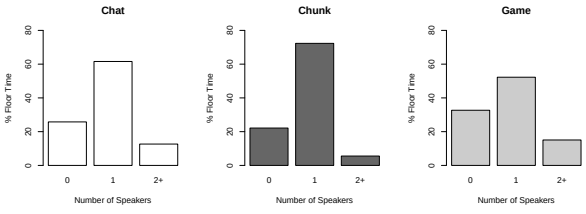
To explore the dynamics of the different speech-exchange systems, the recordings and transcripts for each game in the Teams corpus, and each chat and chunk phase in the CasualTalk dataset were processed using Praat to create 'floor state' annotations. These annotations divided each interaction into labelled intervals, where an alphanumeric code for each interval recorded who was speaking during the interval timespan, or labelled intervals of global silence (where nobody was speaking). For example, the label **aSbS** denotes that speakers **a** and **b** are speaking in overlap, while **cS** indicates that **c** is speaking alone, and **GX** denotes global silence. Any speakers not mentioned in labels are silent for the interval described. We chose to compare the games against chat and chunk phases of larger casual conversations as these separate phases constitute speech exchange systems or genres in their own right [2].

Below we present our investigations, performed using R statistical software [14], and compare and contrast the results.

3. CONVERSATIONAL FLOOR DISTRIBUTION

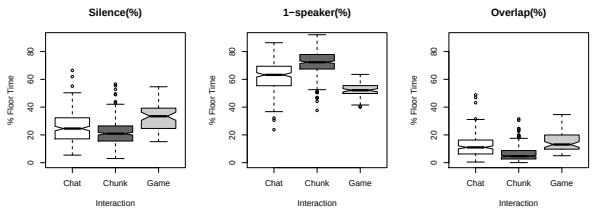
Our first investigation is a comparison of the proportions of silence, one-party speech, and overlapped speech (2 or more speakers) in casual conversation (subdivided into chat and chunk segments), and game interaction. Figure 1 shows the proportion of the interaction occupied by 0,1, or 2+ speakers in chat and chunk phases and in game. Overlap involving three or more speakers is much less frequent than two party overlap in all three conditions, and thus we amalgamate 2 and 3+ party overlap.

Figure 1: Distribution of the floor in chat (white, chunk (dark grey), and games (light grey). X-axis shows number of speakers (0,1,2+) speaking concurrently.



The boxplots in Figure 2 show the distributions of the proportions of silence, one-party speech, and overlap per individual chat, chunk, or game.

Figure 2: Distribution of the floor in Silence, 1-Speaker and Overlap in chat (white), chunk (dark grey), and games (light grey).



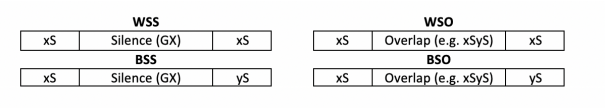
In all conditions the most common conversational situation in terms of time taken was single participant speaking in the clear, the second most common was global silence, with overlap accounting for much less of the conversational time. As some of the distributions skewed right while other skewed left (so that log values did not much improve the normality of the distributions), non-parametric Wilcoxon Rank Sum tests were applied to examine whether proportions of silence, single speaker, and overlap differed between the Chat, Chunk, and Game conditions. For silence, Game interactions had significantly more silence (median 33.4%) than Chat (median 24.6%), which in turn had significantly more silence than Chunk segments (median 20.9%). There was significantly more 1-party speech in Chunk (median 72.3%) than in Chat (median 63.2%), and in Chat than in Game conditions (median 33.4%). Overlap was significantly higher in Game than in Chat, and in Chat than in Chunk. All differences were significant at $p < .0001$

4. SPEAKER CHANGE ACTIVITY

The Teams dataset contains 132380 changes in speech/silence configuration, an average of 1.3 per second. The CasualTalk dataset contains 30688 changes, an average of 1.3 per second. Chunks account for 18081 of these changes, an average 1.2 per second, while there are 12607 changes in the chat data, for an average 1.5 per second. It should be noted that these averages are largely due to large numbers of short utterances in the data.

In an n-party conversation where each participant may be speaking or silent at any moment, there are n^2 possible states for the dialogue at any time. For an overview of state changes we look at cases where a speaker, SpX, speaks alone, followed by a change in floor configuration. SpX's single-party speech interval can be followed by silence, overlap, or a smooth switch to one or more other speakers. The likelihood of two or more speakers starting at the same instant or one speaker starting immediately as another finishes is very small. In the combined dataset, there are 76091 intervals of single party speech in the clear; global silence follows 62.3% of these, while overlap follows 37%. Simultaneous onset of speech by speakers other than the preceding speaker and smooth switching account for less than 1% of the data, and these cases were omitted. This results in four 'transition types' of interest, two around overlap and two around silence. We do not make any attempt to distinguish between backchannels or longer utterances from incoming speakers after the silence or overlap at this stage, although we plan a finer analysis in future work.

Figure 3: Silence (left) and Overlap (right) transitions



Using terminology from [7] based on dyadic interaction as shown in Figure 3, the transition types for SpX are :

WSS Within Speaker Silence - SpX speaks before and after a silence

BSS Between Speaker Silence - SpX speaks before silence, SpX is not speaking after silence

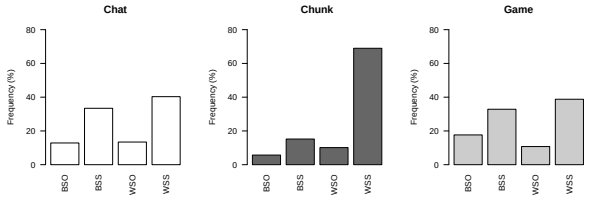
WSO Within Speaker Overlap - SpX speaks before, during and after overlap

BSO Between Speaker Overlap - SpX speaks before and during overlap, SpX is not speaking after overlap

Here WSO is used when the first speaker 'survives' the first overlap, but in multiparty interaction there can be sequential overlap states, when a second overlapper joins, so WSO can end in the original speaker speaking alone or in a different overlap configuration. A more accurate analysis of this case is planned

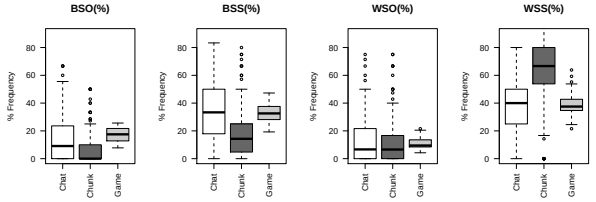
To more closely explore dynamics after a single speaker makes a contribution other than a backchannel or short utterance, we focus on what happens after a single speaker (SpX) speaks alone for at least one second, and impose a minimum silence threshold of 60ms to reduce the chance

Figure 4: Distribution (%) of the four transition types of interest in chat (white), chunk (dark grey), and games (light grey).



of counting stop occlusions as within speaker silences. In this reduced dataset , looking at the situation after a single speaker stretch of at least a second, there are a total of 23777 change points, comprising 10515 WSS, 7080 BSS, 2576 WSO, and 3606 BSO. Figure 4 gives an overview of how transition occurs over the different interaction conditions.

Figure 5: Distribution (%) of the 4 transition types (BSO,BSS,WSO,WSS) in chat (white), chunk (dark grey), and games (light grey).



The boxplots in Figure 5 show the distributions of the proportions of WSS, WSO, BSS, and BSO per individual chat, chunk, or game. Within speaker silence (WSS) was the most common transition in all speech exchange systems, and particularly so in Chunk. Between speaker silence (BSS) and between speaker overlap(BSO) was less common in Chunk than in Chat or Game, while overlap in general and particularly between speaker overlap (BSO) was more common in Game than in the other conditions. Wilcoxon Rank Sum tests were applied. For between speaker overlap, all three conditions were significantly different from one another ($p<.0001$). The proportion of between speaker silence did not differ significantly between Chat and Game, but was significantly lower than either of these in Chunk ($p<.0001$). Within speaker overlap did not differ significantly between Chat and Chunk, but was significantly higher in the Game condition than in Chunk ($p<.01$). Within speaker silence is significantly higher in Chunk than either of the other conditions ($p<.0001$).

5. DISCUSSION AND CONCLUSIONS

Our explorations have shown significant differences between the proportional distribution of silence, single-party speech and overlap in three types of multiparty spoken interaction - a collaborative game, interactive chat and more generic chunk phases of casual conversation. We have also seen that there were differences and similarities in the distribution of how single-speaker stretches end in the three conditions, giving insight into transitions from one floor state to another. In general, the Game and Chat conditions are most similar in transition terms, although they can be distinguished by the significantly higher silence quotient in Game. This may be due to time spent in thought during games while chat is more consistently interactive, and the higher within speaker overlap in games may signal periods of high interactivity alternating with long but infrequent silences. Chunk segments of casual talk contain less silence and overlap (particularly between speaker) and more single party speech than either of the other conditions, reflecting the dominance of one speaker.

While there have been major advances in the use of machine learning to model turntaking dynamics of spoken interaction, including multiparty dialog [9], such models are dependent on the availability of suitable data. Data is very scarce for some very common speech exchange systems - in the current work, there is an almost eightfold difference in the hours of data for casual talk (6 hours) and the games corpus (47). The differences shown between speech exchange systems here demonstrate the importance of understanding the differences in structure and dynamics of types of interaction, as artificial systems trained on game data will not necessarily be accurate in predicting the dynamics of conversational speech. Such knowledge has two main applications. Firstly, it signals the need for data collection in areas of spoken conversation beyond short chats or tasks. Secondly, by identifying similarities in aspects of different interaction types, cataloguing of the dynamics of different interaction types would in help in choosing suitable data to partially model speech exchange systems where sufficient exact data is not available.

We are currently exploring more complex transitions, such as sequences of overlapping talk where more speakers join or current speakers leave an overlap phase in order to more accurately model multiparty spoken interaction.

6. ACKNOWLEDGEMENTS

This work is supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

7. REFERENCES

[1] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering* 6(3&4), 213–228.

[2] Bakhtin, M. 1986. The problem of speech genres. In: *Speech genres and other late essays*. University of Texas Press 60–102.

[3] Boersma, P., Weenink, D. 2010. *Praat: doing phonetics by computer [Computer program], Version 5.1. 44*.

[4] Brown, G., Yule, G. 1983. *Teaching the spoken language* volume 2. Cambridge University Press.

[5] Campbell, N. 2008. Multimodal processing of discourse information; the effect of synchrony. *Universal Communication, 2008. ISUC'08. Second International Symposium on* 12–15.

[6] Eggins, S., Slade, D. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

[7] Heldner, M., Edlund, J. Oct. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38(4), 555–568.

[8] Hennig, S., Chellali, R., Campbell, N. 2014. The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. Reykjavik, Iceland.

[9] Laskowski, K. 2011. *Predicting, detecting and explaining the occurrence of vocal activity in multiparty conversation*. PhD thesis Carnegie Mellon University.

[10] Laver, J. 1975. Communicative functions of phatic communion. *Organization of behavior in face-to-face interaction* 215–238.

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

[12] Malinowski, B. 1923. The problem of meaning in primitive languages. *Supplementary in the Meaning of Meaning* 1–84.

[13] Oertel, C., Cummins, F., Edlund, J., Wagner, P., Campbell, N. 2010. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces* 1–10.

[14] R Core Team, 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[15] Schegloff, E., Sacks, H. 1973. Opening up closings. *Semiotica* 8(4), 289–327.

[16] Schneider, K. P. 1988. *Small talk: Analysing phatic discourse* volume 1. Hitzeroth Marburg.

[17] Skantze, G. 2009. TMH KTH :: Higgins Annotation Tool.

[18] Slade, D. 2007. *The texture of casual conversation: A multidimensional interpretation*. Equinox.

[19] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. Elan: a professional framework for multimodality research. *Proceedings of LREC* volume 2006.