INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speaker transition patterns in three-party conversation: Evidence from English, Estonian and Swedish

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

[1]Stockholm University, Sweden
[2]Trinity College Dublin, Ireland

wlodarczak@ling.su.se, gilmare@tcd.ie

Abstract

During conversation, speakers hold and relinquish the floor, resulting in turn yield and retention. We examine these phenomena in three-party conversations in English, Swedish, and Estonian. We define within- and between-speaker transitions in terms of shorter intervals of speech, silence and overlap bounded by stretches of one-party speech longer than 1 second by the same or different speakers. This method gives us insights into how turn change and retention proceed, revealing that the majority of speaker transitions are more complex and involve more intermediate activity than a single silence or overlap. We examine the composition of within and between transitions in terms of number of speakers involved, incidence and proportion of solo speech, silence and overlap. We derive the most common within- and between-speaker transitions in the three languages, finding evidence of striking commonalities in how the floor is managed. Our findings suggest that current models of turn-taking used in dialogue technology could be extended using these results to more accurately reflect the realities of human-human dialogue.

Index Terms: conversation, turn-taking, spoken dialog

1. Introduction

Human conversation generally proceeds one speaker at a time, with speaker turns managed locally by the participants[1], resulting in single-party speech bounded by stretches of silence and overlap. Existing studies of speaker transitions, whether conversation analytic descriptions of fragments of conversation or large scale corpus studies, have greatly added to our understanding of the underlying mechanisms. At the same time, both methods have serious (and largely complementary) limitations. While conversation analysis gives detailed descriptions of particular examples of interaction, e.g. [2], it fails to provide statistical evidence of which patterns are most common. Conversely, stochastic models of turn-taking [3], based on speech and silence timing, can be used to build predictive models of how conversations are progressing, e.g. [4], but they do not provide human level insights into what is happening in conversation.

Our work threads a middle path, in order to increase our understanding of conversational dynamics by using datasets rather than examples to classify common patterns in spoken interaction. To that end, we use basic conversational data - speech and silence labels - to explore in detail how interaction proceeds. We use annotations of changes in who is speaking, silent or in overlap between stretches of one-party speech in order to define between and within speaker transitions. In order to approximate turn change and retention more accurately, we only consider transitions between single-party speech intervals of at least one second in duration, to avoid treating backchannels or other short utterances as initiating or finishing floor changes. This method gives us insights into how talk proceeds, and also allows us to catalogue the most common transitions present in three-party spoken interaction in 24 conversations drawn from three datasets in the three languages: English, Estonian and Swedish.

We define the 'floor state' at any point of a conversation as the totality of participants speaking at the time, and represent interaction as a series of intervals of varying length where a particular floor state prevails. For example, an interval where A and B are speaking in overlap would be labelled AB, C speaking alone would be labelled C, and general silence would be labelled X. Sequences of these floor state labels describe the progression of speaker activity in conversation.

Two-party interaction can be described using 2^2 floor state labels, X, A, B, and AB. A stretch of single party speech from speaker A could theoretically be followed by silence (X), overlap (AB) or a smooth switch to speech by B. Given that smooth switches (speaker changes with no gap or overlap, e.g A_B) are very rare, earlier work on turn-taking in two-party interaction (e.g. [5]) focused on the four possible patterns of silence and overlap: between-speaker silence (A_X_B), between-speaker overlap (A_AB_B), within-speaker silence (A_X_A), within-speaker overlap (A_AB_A).

Multiparty talk is more complex. In particular, in three-party speech there are 2^3 possible floor states: X, A, B, C, AB, AC, BC, ABC. We define *transitions* as stretches of conversation beginning and ending with intervals of single party speech of at least one second. We chose this threshold with reference to the median value of talkspurts longer than 400 ms (to exclude backchannels or very short utterances). Each transition is further categorised as within- or between-speaker (WST and BST, respectively), depending on whether the first and last single-speaker intervals are produced by the same or different speakers.

Figure 1 shows an example of a between-speaker transition involving two instances of two-party overlap (AC), an instance a three-party overlap (ABC), an interval of solo speech (C) and a silence (X). Note that if the one-second threshold were not applied, the example would be classified as involving two transitions (from A to C and from C to B), even though the short stretch of solo speech by C is unlikely to comprise an actual claim for turn possession.

For convenience we term intervals of single party speech ***1Sp***, intervals of single party speech of at least one second in

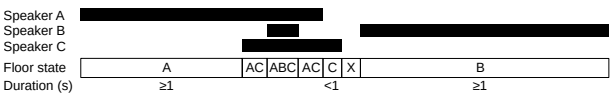

Figure 1: *An example of a between-speaker transition.*

http://dx.doi.org/10.21437/Interspeech.2021-199

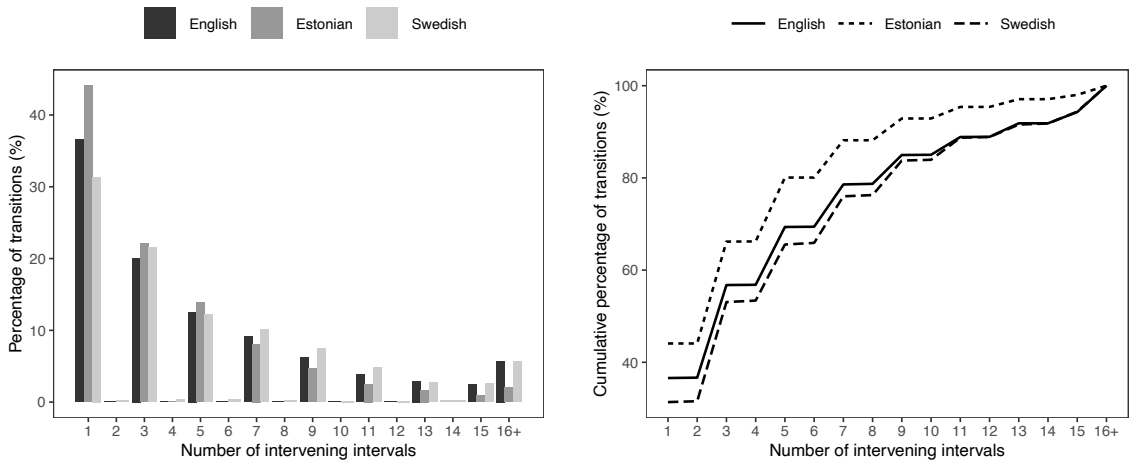

Figure 2: *Frequency distribution (left) and cumulative distribution (right) of speaker transitions for English, Estonian and Swedish depending on the number of intervening intervals.*

duration ***1Sp1***, and transitions between ***1Sp1*** intervals ***1Sp1 - 1Sp1***. We term the labels between the bounding 1Sp1 intervals *intervening* intervals; in the A_AC_ABC_AC_C_X_B example in Figure 1, there are five intervening intervals: AC, ABC, AC, C, and X.

In previous work, we have explored the nature and frequency of within and between speaker transitions in multiparty talk, finding that the majority of transitions involve more than one intervening interval to complete and that the vast bulk of transitions involve odd numbers of intervening intervals [6]. The scarcity of even numbers of intervening intervals follows from the rarity of smooth switches and instances of simultaneous onset or offset of speech. For the data used in this paper, we found that over 95% of transitions are completed in 15 or fewer intervening intervals, and over 65% involved more than one interval, with a higher likelihood of more complex transitions in the BST category. We also found that the duration of the right hand threshold affected the transition type label and number of intervening intervals assigned, with transition type label (WST or BST) changing for 28.29% of transitions depending on how the right hand interval is defined, while almost 60% would change the number of intervening intervals involved. We found one-interval transitions to be the largest class, with a higher proportion of one-interval transitions in WST, perhaps reflecting breath pauses or single backchannels during monologic stretches [7].

Here, we build on this work by exploring transitions in more depth. We first analyse the internal composition of transitions involving different numbers of intervening intervals, in terms of speech, silence and overlap, and transition type (BST or WST). We then isolate the most frequent between and within speaker transition sequences and compare them across the three languages. We discuss implications of these findings for artificial dialog technology.

2. Data and Annotation

The data used in this work are three-party spontaneous conversations in Estonian [8] and Swedish [9], and collaborative conversational games in English [10]. Eight interactions were drawn from each corpus. The Estonian data were collected from 24 unique speakers (13F/11M; mean age = 24.5, *SD* = 2.75). The Swedish data were collected from 24 unique speakers (12F/12M; mean age = 25.5, *SD* = 10). The English data were collected from 24 unique speakers (11F/13M; mean age = 24, *SD* = 12). All the data had been segmented manually.

The annotated data set contained 22106 talkspurts in 9 hours and 51 minutes hours of conversation. The average conversation length was 24.7 minutes (Estonian: 25.8 , Swedish: 23.0, English: 25.2). There were 44160 floor state intervals – 13774 silent intervals accounted for 29.1% of conversation time, 21136 single-party speech (***1Sp***) for 60%, 8312 two-party overlap for 10.1%, and 938 three-party overlap for less than 1%.

The segmented data were processed with a Python script using TextGridTools [11] to create floor state and transition labels as well as to extract the number and identity of participants speaking during the transition. The speech and silence annotations as well as the code for data prepocessing and analysis is available at `https://zenodo.org/record/4923246`.

3. Results

The dataset contains 21136 1Sp intervals (solo speech of any length), of which 35% were 1Sp1 intervals (solo speech of at least one second in duration). 7450 ***1Sp1 - 1Sp1*** transitions were identified in the 9 hours and 51 minutes of talk in the dataset, an average of one transition every 4.7 seconds. Figure 2 shows the frequency and cumulative distributions of all ***1Sp1-1Sp1*** transitions in the English, Estonian, and Swedish data. One-interval transitions are the largest class, particularly in Estonian, and the frequency of transitions decreases with increasing numbers of intervening intervals. The vast bulk of transitions (95.42%) are completed in 15 intervals or fewer. There are very few transitions involving even numbers of intervening intervals (<1%), as such transitions would involve smooth switches or simultaneous onset or offset of speech, which are very rare in conversation. Overall, 57% of transitions are WST. The ratio of WST to BST is particularly high for 1-interval transitions, where 65% are WST, probably reflecting breath pauses. This ratio evens out as the number of intervals per transition grows, with a higher proportion of BST for the 7 and 9-interval cases.

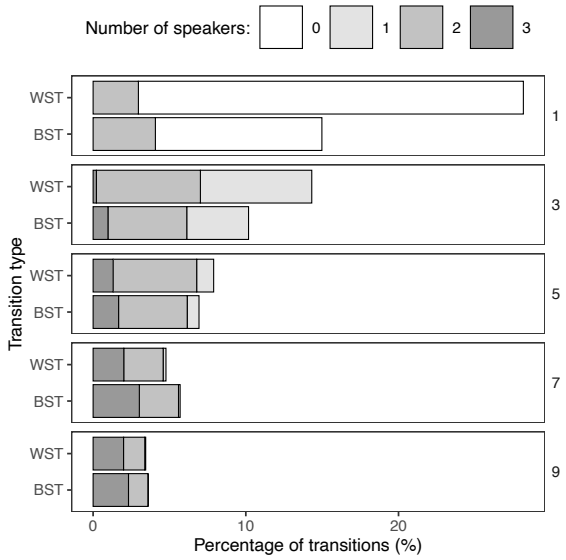

Figure 3: *Distribution of the number of participants speaking during between and within transitions depending on the number of intervening intervals.*

3.1. Speaker participation in transitions

Figure 3 shows the level of participation in 1, 3, 5 and 9 interval transitions, in terms of the total number of speakers appearing in the intervening intervals. For one-interval transitions there are two possibilities - the intervals can contain only silence or only overlap, and thus have 0 or 2 speakers involved. Transitions can involve more participants with increasing numbers of intervening intervals, with participation by all three speakers more likely in BST than WST.

3.2. Composition of transitions - incidence

Across the entire dataset, at least one silent interval is present in 91% of transitions, with overlap appearing in 53% of all transitions. Details of the incidence of silence, solo speech and overlap in 1, 3, 5, 7 and 9-interval transitions can be seen in Table 1.The bulk of one interval transitions are silences for BST (73%) and even more so for WST (90%).

As the number of intervals increases, transitions contain more complex combinations of speech and silence, and the occurrence of all of these features becomes more likely. Silence occurs in the vast bulk of transitions of 3 or more intervals, as does solo speech. The incidence of overlap increases with increasing number of intervals, and is more common throughout in BST than WST.

3.3. Composition of transitions - duration

Figure 4 shows the percentage of transition duration occupied by silence, solo speech and overlap for BST and WST of 3, 5. 7 and 9 intervals. The 1-interval cases are not shown as they are either completely silent or completely in overlap. Silence occupies a large share of the duration of both within and between speaker 3-interval transitions, and the duration occupied by silence remains the leader but decreases as the number of intervening intervals increases, while the duration accounted for by overlap increases. The percentage of duration taken up by solo

Table 1: *Incidence of silence, solo speech, and overlap in within and between speaker transitions in 3 to 9 interval transitions. Incidence expressed as percentage of transitions for each class*

WST	1 (%)	3 (%)	5 (%)	7 (%)	9 (%)
Silence	89	94	95	99	98
Solo	-	99	100	100	100
Overlap	11	49	69	83	90
BST	**1 (%)**	**3 (%)**	**5 (%)**	**7 (%)**	**9 (%)**
Silence	73	86	94	96	97
Solo	-	96	99	100	100
Overlap	27	60	74	88	91

speech increases between the 3- and 5- interval cases in both WST and BST, and then appears to level off or even decrease.

3.4. Most Common Transition Sequences

In this section we further investigate the most common BST and WST sequences occurring in the data.

The most common sequences overall were A_X_A (within speaker silence) and A_X_B (between speaker silence). Interestingly, for WST, both A_X_A_X_A and A_X_B_X_A were more common than 1-interval overlap (A_A:B_A), while the second most common BST was 1-interval overlap (A_A:B_B).

The 3-interval transition class accounts for 21% of all transitions, 22% of WST, 21% of BST, and, with the 1-interval class (38%) accounts for 59% of all transitions. For all languages, the frequency of the most common 5-interval transitions were lower than the fifth most frequent 3-intervals, except for English WSTs, where the most common 5-interval transition was marginally more frequent than the fifth most frequent 3-interval WST. We therefore further explored the 3-interval transitions to better understand the most common transitions in the data.

Table 2 shows the five most common 3-interval WST and BST sequences for English, Estonian and Swedish. The first and second most common within and between transitions are the same in all three languages. All of the top five 3-interval within speaker transitions across the three languages are accounted for by six transition labels, while the top five between speaker transitions are covered by seven transition labels. The categories also show striking similarity in their percentage frequencies across the three datasets.

4. Discussion

Our explorations of multiparty talk in English, Estonian and Swedish have added detail on within and between speaker transitions, and, by extension, give us new insights into how speakers locally manage turn-taking, without requiring very detailed human annotation of turns.

Our previous findings that most transitions involve more than one intervening interval of speech, silence or overlap prompted us to describe and investigate what happens in the more common transitions in three-party English, Estonian and Swedish conversational data. The analysis has shown that transitions are quite complex. In BSTs, contributions from the opening and closing speakers speaking solo and in overlap interspersed with silent intervals appear regularly. Participation by the third speaker occurs less frequently in 3-interval transitions but becomes more likely in transitions involving more in-

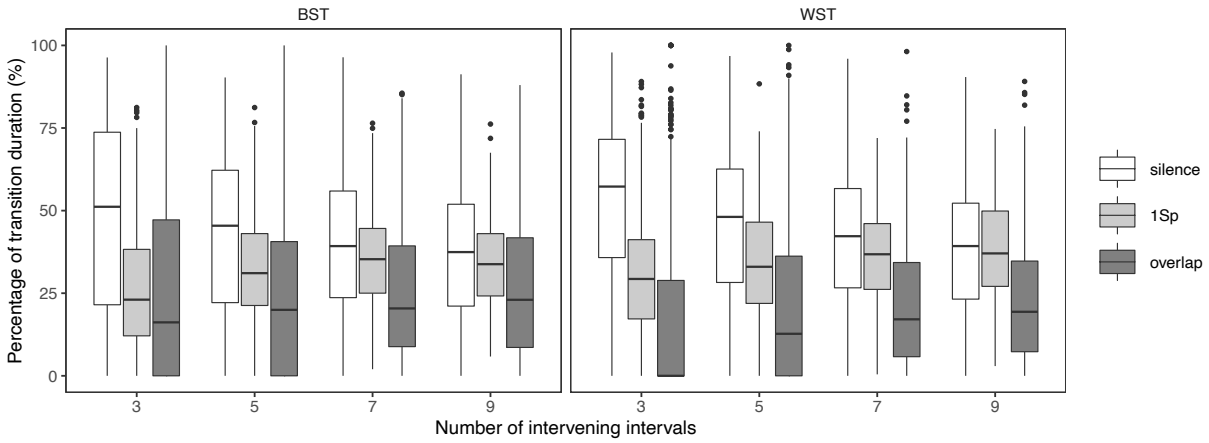

Figure 4: *Distribution of silence, solo speech and overlap as percentage of transition duration for 3, 5, 7 and 9-interval BST and WST transitions.*

Table 2: *Five most common between- and within-speaker transition types with three intervening intervals for the three languages, English (Eng), Estonian (Est) and Swedish (Swe), with ranking. Ties are denoted with* t*. The percentage frequency of each transition in each language's total 3-interval WSTs and BSTs respectively is shown in brackets.*

WST Label	Eng (%)	Est (%)	Swe (%)
A_X_A_X_A	1 (28)	1 (27)	1 (29)
A_X_B_X_A	2 (19)	2 (24)	2 (25)
A_X_B_AB_A	3*t* (17)	3 (15)	3 (14)
A_AB_B_X_A	3*t* (17)	4 (13)	-
A_AB_A_X_A	-	-	4 (13)
A_X_A_AB_A	5 (7)	5 (10)	5 (8)
BST Label	**Eng (%)**	**Est (%)**	**Swe (%)**
A_X_B_X_B	1 (20)	1 (19)	1 (20)
A_X_A_X_B	2*t* (13)	2*t* (11)	2 (15)
A_X_C_X_B	-	2*t* (11)	-
A_X_A_AB_B	2*t* (13)	5 (8)	3 (11)
A_AB_B_X_B	4 (7)	-	4*t* (8)
A_X_B_AB_B	5*t* (6)	4 (7)	-
A_X_C_BC_B	5*t* (6)	-	4*t* (8)

tervals. In WSTs, speakers other than the opening and closing speaker regularly appear in the intervening intervals, with all three speakers appearing in increasing proportion with increasing interval numbers. Silent intervals account for a significant part of the duration of both WST and BST transitions, particularly for 1- and 3-interval transitions. Even though there is a very large range of possible transition labels possible and indeed present in the data, we have seen that a small subset of 17 labels (4 one-interval and 13 three-interval) will account for many of floor state transitions occurring in the data considered.

Notably, these trends were observed across the three corpora, which differ not only in terms of the language used but also conversation type (casual conversation in Estonian and Swedish, collaborative games in English). It remains to be seen how these results generalise to other conversational scenarios and how they are affected by variables such as familiarity. We are planning to pursue this line of research in future studies.

These results have implications for spoken dialog system research and design. While task based systems, e.g. airline reservations or pizza ordering, may be amply served by simple turn-taking mechanisms based on elapsed silence, applications which seek to create the illusion of natural conversation, and particularly those which attempt to engage users over multiple sessions, will need to more accurately model what really happens in human spoken interaction. In recent years, data driven approaches to turntaking management have become popular. While machine learning approaches are attractive, there is a lack of suitable data to train fully data driven models on, and of understanding of how turn change and retention proceeds in practice to inform data selection and feature engineering. It is also well known that several prosodic, linguistic, and non-vocal features (e.g. gaze) can contribute to turn change prediction, and more sophisticated turn change mechanisms have been devised (e.g. [12]), with much recent work machine learning predictive models incorporating multimodal features (e.g. [13]). All of the analysis we perform is on data for which video recordings exist, allowing future analysis of features beyond the mere presence of speech and silence. Knowledge of the most likely transitions in natural spoken interaction could help in the design of more realistic artificial dialog. We thus hope that the insights gained in exploring the details of what happens in the most common sequences of floor state intervals between stretches of single party speech will help in better understanding of spoken dialog and also in the design of more human-like, and thus more efficient artificial dialog.

5. Acknowledgements

This work was conducted with the support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is 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.

6. References

[1] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, pp. 696–735, 1974.

[2] E. Schegloff and H. Sacks, "Opening up closings," *Semiotica*, vol. 8, no. 4, pp. 289–327, 1973.

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

[4] K. Laskowski, "A framework for the automatic inference of stochastic turn-taking styles," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 202–211.

[5] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[6] 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.

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

[8] 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

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

[10] 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.

[11] H. Buschmeier and M. Włodarczak, "TextGridTools: A TextGrid processing and analysis toolkit for Python," in *Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013)*, ser. Studientexte zur Sprachkommunikation, P. Wagner, Ed., vol. 65. Dresden: TUDpress, 2013, pp. 152–157.

[12] A. Raux and M. Eskenazi, "Optimizing end-of-turn detection for spoken dialog systems," in *Workshop on Modeling Human Communication Dynamics at NIPS*, 2010, pp. 14–17.

[13] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale RNNs," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2018, p. 186–190.