10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Between and Within Speaker Transitions in Multiparty Conversation-cameraready

Emer Gilmartin[1], *Kätlin Aare*[2,3], *Maria O'Reilly*[1], *Marcin Włodarczak*[2]

[1]ADAPT Centre, Trinity College Dublin, Ireland
[2]Stockholm University, Sweden
[3]University of Tartu, Estonia

{gilmare,oreil12}@tcd.ie, {katlin.aare, marcin.wlodarczak}@ling.su.se

Abstract

Casual conversation proceeds as a series of contributions from participants, either speaking in the clear or in overlap. The pattern of who is speaking or not (the conversational floor state) changes constantly throughout a conversation. We examine the nature and frequency of these state changes or transitions in multiparty talk, which may involve more complicated floor state transitions than dyadic interactions. We contrast between and within speaker transitions, analyzing the evolution of the conversational floor state from a stretch of single party speech in the clear to the next stretch of single party speech in the clear by the original or a different speaker. We investigate the effect of applying a minimum duration of single party speech in the clear to the incoming speaker's production, finding substantial differences in how transitions are categorized. Over 40% of the transitions categorized as between or within speaker change category depending on whether a minimum duration is applied to the following stretch of single party speech.
Index Terms: dialogue, turn taking, interaction

1. Introduction

The *floor state* of a conversation provides an overview of who is speaking and who is silent, and *floor state changes* reflect conversational dynamics as different speakers contribute to the conversation. It is likely that multiparty conversational dynamics are more complex than well-studied dyadic cases. Thus, we explore three-party spoken interaction in three languages to better understand the floor state dynamics involved as participants speak and fall silent. We use speech and silence timing data to investigate the possible floor state progressions between a stretch of clear (non-overlapped) speech by one speaker and the onset of the next stretch of clear speech - from the same speaker (a *within speaker transition*) or another (a *between speaker transition*), transitions loosely corresponding to turn retention or turn change.

Research into speech and silence patterns in conversation has generally followed two paths - large scale statistical modelling, and minute examination of examples of interaction using Conversation Analysis. The former has been used to generate predictions on several aspects of spoken interaction, including conversations and psychiatric interviews [1, 2], spoken interaction in stressed and unstressed conditions and interaction with infants [3, 4, 5], interspeaker effects in interviews [6, 7], and turntaking and overlap dynamics in multiparty meetings [8]. In Conversation Analysis, where examination of meticulously transcribed examples from real conversations are used to highlight broader patterns of interaction, timing data have been used in conjunction with transcripts in Sacks, Schegloff, and Jefferson's seminal work on conversational phenomena [9, 10, 11] and their turntaking algorithm [12], and in a large body of later related work.

Our ultimate goal is to thread a middle path between large scale prediction and fine grained explicative analysis in order to explore how multiparty casual conversation evolves around silence or overlap, and in particular to gain insight into patterns of speaker activity around these phenomena. In this paper, we analyse multiparty casual conversations in Estonian, Swedish, and English to investigate the nature of the floor state in these interactions and of transitions from floor state to floor state. Below we describe the methods and data used, present our investigations, performed using R statistical software, [13], and report and discuss our findings.

2. Floor State

Speech and silence timing data have been used to model several aspects of spoken interaction. Such *chronemic* analysis is based on observance of the *presence* of speech and silence, and thus does not depend on human judgment of participant intent [14] or theory dependent definitions of turns. A basis for analysis of spoken interaction would seem to be a 'complete' stretch of speech from one participant. However, the definition of such a stretch has proven complicated, often relying on theory-dependent concepts such as 'turn' or 'utterance'[15], or more theory neutral but often speaker-intent dependent concepts such as talkspurts [16, 17]. We base our analysis on spoken interaction data manually segmented into interpausal units (IPUs), defined as a stretch of speech from a particular speaker bounded by silence from that speaker. We then define the 'floor state' of a conversation at any time 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. Relevant concepts discussed below are illustrated in Figure 1.

An *n*-party conversation, where each participant may be speaking or silent at any moment, has 2^n possible floor states, including global silence. In Figure 1, 13 such states are shown. The entire floor state sequence in Figure 1 would be represented as **A_AC_C_BC_ABC_AB_B_BC_C_GX_B_GX_A** where A, B, and C are participants speaking and GX represents global silence. For the purposes of this study, we use a shorthand of the sequence of the number of speakers in each floor state interval over longer stretches of conversation. As an example, **1_2_1_0_1** defines a stretch where single party speech is followed by two-party speech, then single-party speech, silence, and single-party speech. Note that this shorthand does not identify which speakers are involved in each interval.

We adopt the term ***1Sp*** to mean single-party speech in the

10.21437/SpeechProsody.2020-163

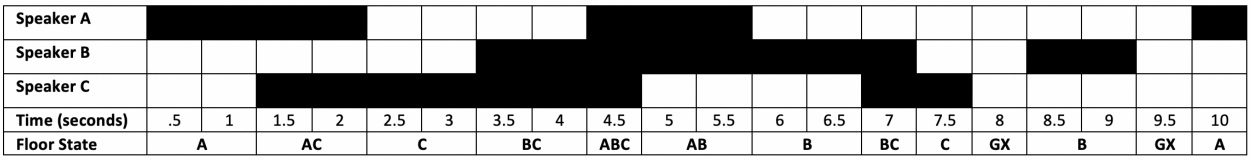

Figure 1: *Three participants in a 10-second stretch of conversation, where black represents speech and white represents silence. Speaker A produces 3 inter-pausal units (IPUs) totalling 4 seconds (2+1.5+.5) of speech, Speaker B produces 2 IPUs totalling 5 seconds (4+1), and Speaker C speaks for a total of 4.5 seconds(3.5+1) across 2 IPUs. There is global silence for 1 second (.5+.5). There are a total of 13 floor state labels. The entire floor state sequence would be represented as* ***A_AC_C_BC_ABC_AB_B_BC_C_GX_B_GX_A****, while the simpler numerical shorthand version is* ***1_2_1_2_3_2_1_2_1_0_1_0_1****. Note that interval durations are for illustration purposes and are somewhat unrealistic as overlap is generally shorter than in the example.*

clear – an interval of speech by a single speaker which is not overlapped by any other speaker. Note that this interval is a floor state interval and may not comprise a complete IPU. In dyadic interaction, an interval of ***1Sp*** can theoretically end in silence, two-party overlap, or a *smooth switch* or transition to the other speaker. Here, we use the term smooth switch to mean a change from speaker A to B where the endpoint of A's utterance coincides with the start of B's. In manually segmented data, the likelihood of such a smooth switch is low, given the precision of annotation software based on graphical interfaces (as an example, using Praat software results in possible error of 50ms when segmenting 10 seconds of audio on fullscreen display on a 23" monitor). As a result, smooth switches (e.g. **1_1** in our shorthand) and cases where two speakers start or stop speaking 'simultaneously' (e.g. **0_2** or **2_0**) should be very rare in the data. The type of transition is further determined by the right hand context. To describe transition types we adapt the terminology used in [18] for dyadic interaction. In two-party interaction, silence can transition to ***1Sp*** resulting in a within or between speaker silence, while overlap can transition back to ***1Sp***, resulting in within or between speaker overlap. For speakers A and B, within speaker silence (WSS) is defined as **A_GX_A** and between speaker silence (BSS) is defined as **A_GX_B**, while within and between speaker overlap are **A_AB_A** and **A_AB_B**. Thus, ***1Sp*** can transition back to ***1Sp*** with one intervening interval of silence or overlap, e.g. **1_0_1** or **1_2_1**.

Multi-party speech can transition back to ***1Sp*** with one intervening interval of silence or overlap, as in the two-party condition, but the number of intervening intervals can increase once 3- or more party overlap occurs. Transitions from one substantial interval of single party speech to another, a situation loosely analogous to turn change or retention, can be operationalized by placing a lower bound on the first and final single speaker interval durations. We thus define ***1Sp1*** as an interval of duration one second or more. This one-second minimum duration was chosen with reference to the distributions of IPU duration in the data used in these explorations - where the median length of single speaker IPUs longer than 400ms (to exclude backchannels) was 1.22s. For convenience, we also define ***1SpAny*** as such an interval of *any* duration. Thus a ***1Sp1-1Sp1*** transition denotes a transition between two intervals of single party speech in the clear of at least one second, while ***1Sp1-1SpAny*** denotes a transition from an interval of single party speech of at least one second to an interval of single party speech of any duration. For ***1Sp1-1Sp1*** transitions, the possibilities multiply, as the intervening intervals can include ***1Sp*** segments shorter than the threshold duration, and thus sequences such as **1_2_1_0_1** are possible. For this work, we retain Heldner et al's notion of within and between speaker phenomena, but, as multiparty transitions can involve a combination of overlap and silence, we define only two transition types – within speaker transitions (WST) where we examine transitions beginning and ending with the same speaker, and between speaker transitions (BST), which start with one single speaker and transition to another single speaker. Below we investigate ***Sp1-1Sp1*** and ***1Sp1-1SpAny*** transitions in Estonian, Swedish and English 3-party spoken interaction.

3. Data and Annotation

The data used in this work are three-party spontaneous conversations in Estonian [19] and Swedish [20], and collaborative conversational games in English [21]. 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 segmented data were processed using a custom Praat script[1] [22] to create 'floor state' annotations – labelled intervals, as shown in Figure 1, where an alphanumeric code for each interval recorded who was speaking during the interval timespan, or marked the intervals of global silence (where nobody was speaking). These labels were further processed to form shorthand labels showing the number of speakers per interval.

The annotated data set contained 22106 IPUs 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%. ***1Sp1*** and ***1SpAny*** intervals were identified, and used to extract ***1Sp1-1Sp1*** and ***1Sp1-1SpAny*** intervals. For each ***1Sp1***, we searched forward to locate the next ***1Sp1*** and extracted the sequence of intervals (in terms of speaker numbers) from the initial ***1Sp1*** to the next ***1Sp1***. As an example, **1_2_3_2_1_0_1**

[1]This script was adapted from a script by Jose Joaquin Atria.

Bigram	All (%)	Est (%)	Swe (%)	Eng(%)
0_1	31.11	33.62	31.60	28.25
1_0	31.15	33.62	31.72	28.26
1_2	16.65	15.07	16.13	18.68
2_1	16.72	15.07	16.27	18.71
2_3	2.10	1.31	1.96	2.98
3_2	2.12	1.31	1.98	3.00

Table 1: *Six most frequent bigram transitions (by % frequency) overall, and by language. These bigrams account for 99.85% of all transitions in dataset.*

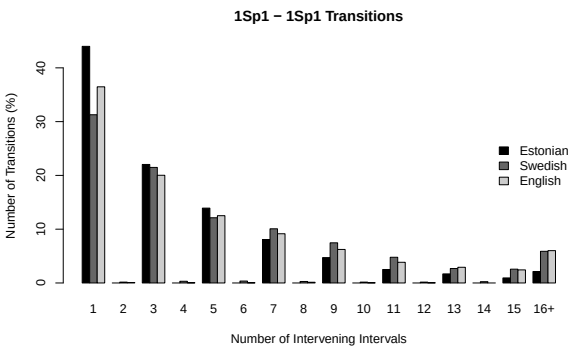

Figure 2: *Number of floor state intervals in 1Sp1-1Sp1 transitions.*

contains 5 intervening intervals between the two stretches of ***1Sp1***. The transitions were then labelled as Between (BST) or Within (WST) Speaker. A similar procedure was followed for ***1Sp1-1SpAny*** intervals.

4. Floor State Change Bigrams

In this section, we analyse how the floor state changes in the corpora. We use bigram representations to explore how states transition in general and from intervals of single party speech.

4.1. General Floor State Bigrams

Bigrams of the number of speakers in every two contiguous intervals in the dataset were created. As an example, **1_0** represented a transition from ***1Sp*** to silence while **3_2** represents a transition from three speakers to two speakers. There are 14 bigram transitions possible in theory, if simultaneous starts and stops and smooth switches are permitted (**0_0** and **3_3** are impossible). Of these, 12 appeared in the data (**0_1**, **0_2**, **1_0**, **1_1**, **1_2**, **1_3**, **2_0**, **2_1**, **2_2**, **2_3**, **3_1**, **3_2**), while the two transitions reflecting simultaneous starts/stops by three participants did not (**0_3**, **3_0**). Six bigrams reflecting transitions adding or subtracting one speaker accounted for 99.85% of the data, as shown in Table 1. The other 6 which also appeared, (**0_2**, **1_1**, **1_3**, **2_0**, **2_2**, **3_1**), reflecting two simultaneous starts/stops by two participants or smooth switches, accounted for between 0.006% and 0.06% of transitions each.

4.2. *1Sp-* and *1Sp1-* Bigrams

Across the dataset, there were 21136 intervals of ***1Sp***. Cases of ***1Sp*** followed by silence (**1_0**) accounted for 65.06% of all transitions from ***1Sp***, while ***1Sp*** followed by two-party overlap accounted for 34.77% of cases.

The proportions above apply to all intervals of ***1Sp***, regardless of duration. To more closely approximate conditions at the end of a turn, we examine transitions starting with a ***1Sp1*** (minimum one second). There were 7467 single-speaker intervals of 1 second or more (***1Sp1***). Cases of ***1Sp1*** followed by silence (**1_0**) accounted for 75.35% of bigrams, while ***1Sp1*** followed by two-party overlap accounted for 24.56%. In all bigram cases, transition from one-party speech to silence (**1_0**) is more common than transition from one-party speech to overlap (**1_2**). The proportion of transition to silence is higher for ***1Sp1*** (75.35% overall) than for ***1Sp*** (65.06% overall).

5. Transitions between Single Speakers

To investigate transitions between intervals of single party speech, a situation analogous to a turn change or retention, we analyse transitions from intervals of single party speech of at least one second's duration (***1Sp1***) to the next ***1Sp1***, and finally we investigate transitions from ***1Sp1*** to ***1SpAny***.

5.1. *1Sp1-1Sp1* Transitions

For each ***1Sp1*** interval, we searched forward in the dataset to locate the next ***1Sp1***. We then extracted the sequence of intervals (in terms of speaker numbers) from the initial ***1Sp1*** to the next ***1Sp1***. These sequences could be of any length, and thus we term them transition interval n-grams. As an example, **1_2_3_2_1_0_1** represents a sequence where ***1Sp1*** transitions to 2-, 3-, and back to 2-party overlap before returning to ***1Sp*** (<1sec) by the original speaker or another, and then transitions to silence before finally transitioning to ***1Sp1*** by the original speaker. This example contains 5 intervening intervals between the two stretches of ***1Sp1***.

Distributions of transitions between single-speaker intervals of 1 second or more in duration in the three languages are shown in Figure 2, where it can be seen that the vast majority of intervening intervals are in stretches of odd numbers of intervals, with the number of cases dropping with increasing intervals. The overwhelming prevalence of odd-numbered interval transitions can be explained by the low likelihood of smooth switches and simultaneous onset or offset of speech, which would be necessary for an even-number interval transition to occur.

Overall, 95.42% of all ***1Sp1-*** intervals are closed by a later ***1Sp1*** in fewer than 16 intervening intervals. The most frequent class of transitions are those with one intervening interval which account for 37.40% of cases. Even-number cases accounted for only 51 (0.71%) of the 7176 transitions between 1 and 15 intervals long.

Disregarding the even-number cases, the number of transitions declines monotonically with the number of intervening intervals between ***1Sp1*** intervals. It is likely that numbers continue to decline with increasing intervals in a long tail.

5.2. Between and Within Speaker Transitions

The even numbered cases and the 16+ interval bucket class were excluded from the data, leaving 7125 transitions, comprising 57.47% WST and 42.53% BST with intervening intervals ranging from 1 to 15. Figure 3 shows these BST and WST transitions by language, while Figure 5 shows interval types in the three languages, and the proportion of transitions per interval total. One-interval transitions were the largest group for BST

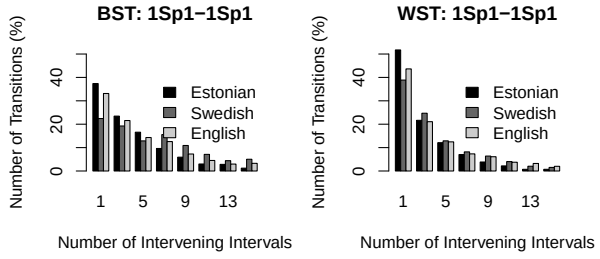

Figure 3: *Floor state intervals in (***1Sp1-1Sp1***) in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right).*

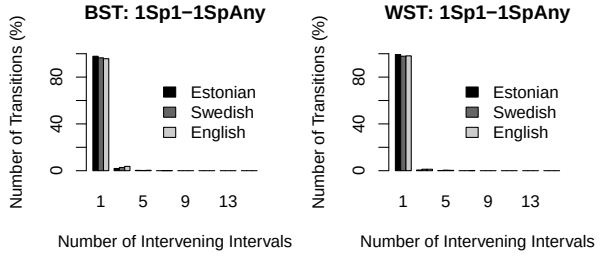

Figure 4: *Floor state intervals in (***1Sp1-1SpAny***) in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right).*

and WST for all languages, with the proportion of 1-interval transitions particularly high for WST. One-interval transitions only account for 35.05% of transitions overall.

5.3. *1Sp1-1Sp1* vs *1Sp1-1SpAny* Transitions

We generated the transition n-grams for each initial single speaker interval (***1Sp1***) to the next single speaker interval *of any duration* (***1SpAny***), as shown in Figure 4.

It can be seen that the vast bulk of the ***1Sp1-1SpAny*** transitions occur with one intervening interval (97.58%), while 3 and 5 intervals account for 1.89% and 0.28% overall respectively.

Comparing ***1Sp1-1SpAny*** transitions sharing their left-hand ***1Sp1*** intervals with the 7125 ***1Sp1-1Sp1*** transitions treated in Section 5.2 results in the confusion matrix in Table 2.

Transition type label (WST or BST) changes 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.

Table 2: *Confusion Matrix (%) for* 1Sp1 *and* 1SpAny *in Between and Within Speaker Transitions*

		1Sp1-1Sp1		
		BST	**WST**	**Sum**
	BST	32.48	18.25	50.72
1Sp1-1SpAny	**WST**	10.05	39.23	49.28
	Sum	42.53	57.47	100

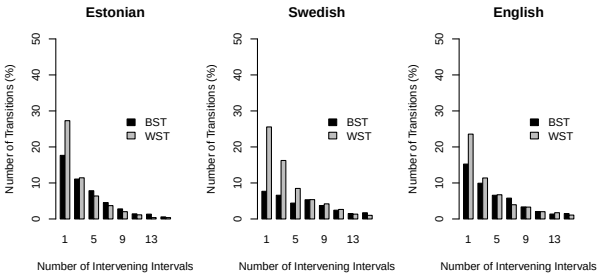

Figure 5: *Percentage of Between and Within Speaker Transitions per number floor state intervals in (***1Sp1**-**1Sp1***) by language.*

6. Discussion and Conclusions

The vast majority of bigram transitions involved the entry or exit of one speaker from the floor. The most common transitions occurred between one-party speech and silence, followed by transitions between m-party speech and m+1-party overlap, with decreasing frequency with increasing m, so, for example, transitions from one-party speech to 2 party overlap were more common than transitions from 2 to 3 party overlap. The bigram results also showed that silence is more likely than overlap to follow single party speech, while the higher likelihood of transition to silence from longer stretches of single party speech may point to the presence of more overlap around shorter utterances, possibly during transitions.

The results on ***1Sp1-1Sp1*** transitions show that BST are distributed more evenly over intervening intervals than WST, thus increasing the frequency of more complex transitions in BST. This could reflect more activity around turn change than around retention, or indeed more backchannels and acknowledgement tokens being contributed by more participants around speaker changes. One-interval transitions are the largest class, with a higher proportion of one-interval transitions in WST, perhaps reflecting breath pauses or single backchannels during monologic stretches. However, one-interval transitions only account for 37% of transitions overall, reflecting the need to consider more complex transitions around turn change and retention.

The high confusion levels between labels depending on how the right hand thresholds are defined also point to the need for clear and standard definitions and specifications for interspeaker activity, particularly in light of the increase in engineering applications of dialogue science.

It would be very interesting to separate within speaker breathing pauses from other transitions in order to better understand transitions around silence. Other future work involves further classification of transitions depending on the number of distinct speakers involved, and investigation of the duration of transitions. It is hoped that this study, and similar studies of other corpora, will allow us to inventory transition types in multiparty spoken interaction, and then analyse examples of the statistically more likely transitions in detail to better understand conversation dynamics.

7. 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 supported by the National Program for the Estonian Language Technology project EKTB3, and funded by Swedish Research Council project 2014-1072 *Andning i samtal* (*Breathing in conversation*) to Marcin Włodarczak and Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034 *Hidden events in turn-taking* to Mattias Heldner.

8. 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. [Online]. Available: http://www.sciencemag.org/content/144/3620/884.short

[2] J. Jaffe, S. Feldstein, and L. Cassotta, "Markovian models of dialogic time patterns." *Nature*, 1967. [Online]. Available: http://psycnet.apa.org/psycinfo/1968-01358-001

[3] L. Cassotta, S. Feldstein, and J. Jaffe, "The stability and modifiability of individual vocal characteristics in stress and nonstress interviews," *Research Bulletin*, no. 2, 1967.

[4] B. Beebe, D. Alson, J. Jaffe, S. Feldstein, and C. Crown, "Vocal congruence in mother-infant play," *Journal of psycholinguistic research*, vol. 17, no. 3, pp. 245–259, 1988.

[5] J. Jaffe, B. Beebe, S. Feldstein, C. L. Crown, M. D. Jasnow, P. Rochat, and D. N. Stern, "Rhythms of dialogue in infancy: Coordinated timing in development," *Monographs of the society for research in child development*, pp. i–149, 2001.

[6] J. D. Matarazzo, A. N. Wiens, G. Saslow, B. V. Allen, and M. Weitman, "Interviewer Mm-Hmm and interviewee speech durations." *Psychotherapy: Theory, Research & Practice*, vol. 1, no. 3, p. 109, 1964. [Online]. Available: http://psycnet.apa.org/journals/pst/1/3/109/

[7] J. D. Matarazzo and A. N. Wiens, "Interviewer Influence on Durations of Interviewee Silence." *Journal of Experimental Research in Personality*, 1967. [Online]. Available: http://psycnet.apa.org/psycinfo/1967-10402-001

[8] K. Laskowski and E. Shriberg, "Modeling other talkers for improved dialog act recognition in meetings." in *INTERSPEECH*. Citeseer, 2009, pp. 2783–2786.

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

[10] G. Jefferson, "Preliminary notes on a possible metric which provides for a'standard maximum'silence of approximately one second in conversation." 1989.

[11] ——, "A case of precision timing in ordinary conversation: overlapped tag-positioned address terms in closing sequences," *Semiotica*, vol. 9, no. 1, pp. 47–96, 2009.

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

[13] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[14] S. Feldstein and J. Welkowitz, "A chronography of conversation: In defense of an objective approach," *Nonverbal behavior and communication*, pp. 329–378, 1978.

[15] J. Edlund, "In search for the conversational homunculus: serving to understand spoken human face-to-face interaction," Ph.D. dissertation, KTH Royal Institute of Technology, 2011.

[16] A. C. Norwine and O. J. Murphy, "Characteristic time intervals in telephonic conversation," *Bell System Technical Journal*, vol. 17, no. 2, pp. 281–291, 1938.

[17] P. T. Brady, "A technique for investigating on-off patterns of speech," *Bell System Technical Journal*, vol. 44, no. 1, pp. 1–22, 1965.

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

[19] P. Lippus, T. Tuisk, N. Salvestre, and P. Tiras, "Phonetic corpus of estonian spontaneous speech," 2013.

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

[21] 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.

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