Speech Pause Patterns in Collaborative Dialogs

Maria Koutsombogera and Carl Vogel

Abstract This chapter discusses the multimodal analysis of human behavioral data from the big data perspective. Though multimodal big data bring tremendous opportunities for related applications, we present current challenges in the domain of multimodal and behavioral analytics. We argue that in the case of analysing human behavior in interaction, we need to shift to the analysis of samples, smaller datasets, before scaling up to large data collections. We describe a dataset developed to study group collaborative interaction from measurable behavioral variables. As a case study, we investigate speech pauses and their patterns in the data, as well as their relationship to the topics of the dialog and the turn-taking mechanism, and we discuss their role in understanding the structure of collaborative interactions as well as in interpreting the behavior of the dialog participants.

1 Introduction

Human communication is rich and complex, in that it consists of an interplay between speech, body activity and cognition. It is not only the content of words, but also the way and the time in which they are uttered that contribute to the successful delivery of the message. Human behavior in interaction depends on an individual's communicative intent and is influenced by other people as well as by the context of the interaction. Therefore, interactions may largely vary depending on the context or the content of the interaction or the nature of the task that the interaction participants are involved in, resulting in heterogeneous multimodal signals. This heterogeneity and variability in human behavior signals make the understanding and automatic decoding of human behavior cues a challenging engineering problem (?, ?).

Maria Koutsombogera
Trinity College Dublin, Dublin 2, Ireland, e-mail: koutsomm@cs.tcd.ie

Carl Vogel
Trinity College Dublin, Dublin 2, Ireland e-mail: vogel@cs.tcd.ie

Multimodal interaction analysis provides information about the way different expressive modalities shape the structure of the interaction (i.e. turn management) and convey the speakers' cognitive and affective state in any given moment (including feedback responses and emotions), thus demonstrating the speakers' interactional and social behavior (?, ?, ?).

In recent years, new methodologies and tools have emerged to quantitatively understand and model human behavior in interaction, by exploiting the richness of multimodal signals that humans convey, and based mostly on affective and social behavior cues, as expressed through speech, acoustic and visual activity features. Aspects of behavior are highly represented in multimodal data, as aspects of distinct data sources that are nevertheless related to each other in a common context.

From this perspective, multimodality is important in the context of big data in the sense of developing models useful for analyzing unstructured behavioral data and responding to challenging optimization problems. Multimodal data analytics provides great opportunities to build better computational models to mine, learn, and analyze enormous amounts of social signals, i.e. data related to human behavior, because of the richness of the data in terms of content, context and speakers (?, ?). While multimodal analysis is usually focused on micro-interaction, the analysis may scale up when linked with broader application areas, such as learning analytics (?, ?) and healthcare analytics (?, ?). For example, multimodal learning analytics is about using advanced sensing and artificial intelligence technologies to measure, collect, analyse and report data about learners and their contexts, with the purpose to improve pedagogical support for students' learning and the optimisation of learning and learning environments. The latter can benefit from multimodal analyses due to the heterogeneity of the data sources available (?, ?), while new high-frequency data collection technologies and machine learning analysis techniques could offer new insights into learning, and students' learning trajectories (?, ?).

1.1 Challenges in Multimodal Analytics

Advances in multimodal and multimedia big data are expected to provide more opportunities to build better computational models to mine, learn and analyse enormous volumes of data. Multimodal analytics is related to behavioral analytics and brings new insights to the analysis of human behavior, focusing on the understanding of how humans behave and why, and enabling accurate predictions of how they are going to act in the future. The rationale behind multimodal and behavioral analytics is to utilize the massive volumes of data collections where users interact in various social or professional contexts, and have the ability to query data in a number of ways and create predictive models of behavior.

The collection of behavioral data from either real-world or laboratory settings offers novel processing and modelling opportunities to extract measurable behavioral cues and develop predictive models of behavior on a large scale. This is a challenge per se, in the sense of developing tools to acquiring multimodal data, controlled or in-the-wild, together with the necessary contextual information that will allow for robust processing and prediction (?, ?). However, an ongoing challenge remains the representation and modelling of multimodal data, as well as the development of computing methods that effectively analyse data. This challenge in the representation is due to the unstructured and heterogeneous nature of data (i.e. data come from multiple sources with different representations), and mainly because of the complexity in understanding the semantic gap between the low-level behavioral features and the high-level semantics they bear (?, ?). Because of their nature, another challenge in multimodal big data are the real-time requirements in the related applications and services that demand more efficient processing and large-scale computation, but also in the optimization of storage and networking resources.

1.2 From Big Data to Sample Data

Because of the multiple levels of information multimodal data conveys, big data analytics need to focus not only on large volumes of multimodal data, but also on the high quality of this data (?, ?). Thus, a limitation that has been acknowledged in big data is that, although enormous data volumes are exciting, data quality is not always guaranteed and that data quality matters more than quantity (?, ?). An important methodological issue in the analysis of multimodal data is sampling. It is essential to investigate samples of multimodal interaction, i.e. datasets of human interactions in a specific context, to better understand what data is about and to be in a position to make valid claims about aspects of human behavior, e.g. examine the frequencies of certain representative features or account for outliers. Sample data therefore enable the extrapolation of arguments and claim representativeness in a specific context (?, ?).

Also, big data introduces the possibility of analyzing whole datasets, but this is extremely complex to do in multimodal analytics, as it is impossible to have access to all possible behaviors in human interaction. Again, because of the variability in human behavior, an out-of-context investigation cannot guarantee whether some behaviors are over- or underrepresented or whether data collections have been created with the same methodology.

Thus, the advantages of working with samples is that we know where the data comes from, the conditions under which they were captured and their quality. We also understand the meaning of the representations they carry, because we are aware of the context in which they have been created. And most importantly, we know the purpose for which the data was collected and the questions we are aiming to answer, the parameters we want to measure and the methodology we follow.

Also important aspect in multimodal data analysis is understanding the context, i.e. the context of the interaction, the physical and discourse setting on which the interaction takes place. As Goffman notes, the awareness of social framings is critical, since speakers adapt their speech depending on who they are speaking with and the expectation that the context raises (?, ?); therefore, framing the research of interaction behavior in a specific environment enables its interpretation.

Thus, when it comes to multimodal human behavior, there are social and cognitive parameters complex to model, at least on a big scale. The investigation of samples of behavior, of smaller datasets, is important in understanding the underlying structure and intentionality, in developing and testing hypotheses and then considering the possibilities of scaling up the research questions to big data.

1.3 Multimodal Group Dialog: Turn-Taking and Pauses

While the most common setup of human multimodal interaction, the two-party dialog, is already a rich and informative setup, multiparty interaction is even more challenging because of the dynamics developed among group members (?, ?). The importance of analyzing collaborative dialog datasets lies in decoding communicative patterns involving verbal and non-verbal modalities. By definition, group dialogue is a canvas where different communicative intentions, personalities, lexical choices that may affect the outcome and the effectiveness of the interaction are manifested by the participants. In terms of behavior modelling, efforts focus on automatically analyzing various facets of group interactions and collecting this knowledge to improve the quality of the interaction either in human-human or in human-machine settings. Related work that exploits group dialogue and multiparty corpora studies prediction of the next speaker in multiparty meetings (?, ?), dominance and leadership (?, ?, ?), and personality traits (?, ?), among others.

The speech signal is a modality that offers important cues for the investigation of turn-taking, the process where there is a speaker change in the conversation. Understanding the turn-taking mechanism is especially important in group dialogue. To achieve smooth turn-taking, dialog participants need to allocate a turn or predict the person who will speak next, and need to consider a strategy for themselves to achieve accurate timing for their own turn. In this respect, the information coming from the speech signal may clarify the behavior that contributes to smooth turn-taking, but also inform about the degree of participation of the speakers in the discussion, revealing possibly unbalanced participation or signs of dominance from certain speakers. This information consists of the speakers' words, the number and duration of their turns, but also of the silence intervals, the pauses that occur during the dialog.

The investigation of pauses and their functions in dialog helps achieving better understanding of human communication and their role in turn organisation, including the examination of who takes a turn and when, how turn allocation is performed and the association of pauses to overlapping speech. Pauses are frequent in spoken language, and in addition to their use and functions, the local context where they occur, but also the linguistic and cultural contexts are also important. As mentioned in detail in the next section, there are numerous studies that address the importance of speech pauses in delivering the discourse message in a successful way, as well as aspects of language specificity. Furthermore, understanding speech pauses contributes to the design and implementation of dialog systems that can perceive and generate natural turn exchanges to interact with humans in a successful and cognitively natural way.

In the remainder of this chapter we will present a study that was carried out in what we called in the previous section *sample* human behavioral data, i.e. a corpus of multimodal group dialogs. In this study, we are looking into silent and breath pauses, i.e. non-filled pauses. The main goals of this study are (a) to discover possible relations among the pauses and the topic of the dialog, (b) to investigate the extent to which the context of the pauses contributes to the identification of next speakers, (c) quantitative aspects in the use and frequency of pauses and (d) the effects of the English native and non-native linguistic background of speakers on the duration of their pauses. In the next sections we discuss related literature, the dataset used for this study, the analysis of the speech pauses and the results of the study.

2 Speech Pauses: Background

Speech pauses can be categorized in general in two groups: silent pauses that may include breath, and filled pauses, which usually include disfluencies, i.e. non-lexicalised words such as *ahm*, *ehm*, etc. Speech pauses are considered as a mechanism related to internal cognitive processes that speakers employ in their messages. They have been described as signals of discourse planning, used as markers of discourse structure (?, ?, ?), and indicating the way speakers are planning their message and speech (?, ?, ?). Most importantly, pauses have been considered as interaction management markers through which speakers regulate the interaction (?, ?, ?). In this respect, pauses have also been associated to the communicative functions of feedback and turn management (?, ?). Other functions related to cognitive processing are those associated to lexical retrieval (?, ?) or difficult concepts that speakers need to think about (?, ?). The effects of filled pauses to memory for discourse have been also investigated and it has been shown that they facilitate recall (?, ?).

From the multimodal perspective, research literature has reported on the temporal, semantic and functional relations in the co-occurrence of pauses and gestures (?, ?, ?, ?, ?). Pauses have also been considered an important functional cue to be taken into account in the design and development of conversational agents, in terms of both perception of human speech pauses as well as in the generation of pauses that contribute to the naturaleness of the agents' output (?, ?, ?, ?).

As far as automatic prediction is concerned, it has been reported that words preceding pauses are reliable predictors of their function as clause boundary markers (?, ?, ?) and that discourse structure can to some extent be predicted from characteristics of filled pauses (?, ?). In tasks related to automatic detection of conversational dominance, the duration of silence intervals (pauses) is important when looking for instances of floor grabbing that occur right after these intervals (?, ?). Similarly, the events of taking a turn during silence or breaking a silence are variables that are associated with dominant speakers (?, ?). Also, approaches on the semi-automatic recognition of filled pauses have been proven to reduce the effort of manual transcription of filled pauses. (?, ?).

Silent and filled pauses have also been explored in relation to personality aspects, especially regarding the fluency in speech and the quantity of pauses produced with regard to personality traits such as extraversion or neuroticism, but it has been also stressed that these relations may be affected by other factors such as social skills (?, ?, ?).

Different cultural and language backgrounds have different effects on speech production. Since speech pauses are part of a speaker's linguistic production, these effects are also evident in pause patterns. Language-specificity and speaker-specificity in pause production have been investigated in the literature. Speakers adopt their own variants of filled pauses (?, ?), while cross-linguistic analyses have provided evidence about the language-specific patterns in the vocalic quality of filled pauses (?, ?). A study about the impact of different factors on pause length (region, gender, ethnicity, age) has shown that region and ethnicity have significant influences on pause duration (?, ?). Also, a comparative study of speech rate in nine languages has shown that the probability of pauses occurring before nouns is about twice as high than before verbs (?, ?). Different pause tolerance (i.e. perception of what is considered e.g. a long pause) in the conversations among speakers of different languages may result in difficulties in communication (?, ?) and, as pause tolerance can vary distinctly between different cultures, different pause patterns may cause problems in intercultural communication (?, ?). An investigation of turn transitions with regards to pauses and overlaps in ten different languages has given evidence about differences across the languages in the average gap between turns (?, ?).

3 Data Description

In this work we used the MULTISIMO corpus, a multimodal corpus consisting of 23 sessions of collaborative group interactions where two players need to provide answers to a quiz and are guided by a human facilitator. Players work together while the facilitator monitors their progress and provides feedback and hints when needed. In this setup, collaboration refers to the process where the two players coordinate their actions to achieve their shared goal, i.e. find the appropriate answers and rank them.

The scenario was designed in a way that would elicit the desired behavior from the participants, that is, encourage their collaboration towards a goal. We thus designed sessions, in which 3 members of a group, 2 players and 1 facilitator, collaborate with each other to solve a quiz. The sessions were carried out in English and the task of the players was to discuss with each other, provide the 3 most popular answers to each of 3 questions (based on survey questions posed to a sample of 100 people), and rank their answers from the most to the least popular. The players expressed and exchanged their personal opinions when discussing the answers, and they announced the facilitator the ranking once they reached a mutual decision. They were also assisted by the facilitator who coordinated this discussion, i.e. provided the instructions of the game and confirmed participants' answers, but also helped participants throughout the session and encouraged them to collaborate.

The corpus consists of a set of audio and video recordings that are fully synchronised. During the recording of the sessions the participants were seated around a table and were captured with three HD cameras, one 360 camera, three head-mounted microphones, one omnidirectional microphone and one Kinect 2 sensor. The head-mounted microphones were recording the individual audio signals (SR 44.1 kHz), while the omnidirectional microphone was used as a backup audio source (SR 44.1 kHz). Thus, the audio files that were used for the present study come from the individual head-mounted microphones.

The overall corpus duration is approximately 4 hours and the average session duration is 10 minutes. Overall, 49 participants were recruited and the pairing of players was randomly scheduled. 46 were assigned the role of players and were paired in 23 groups. The remaining 3 participants shared the role of the facilitator throughout the 23 sessions. In most of the sessions the participants don't know each other, although there are a few cases (i.e. in four groups) where the players are either friends or colleagues. The average age of the participants is 30 years old. Furthermore, gender is balanced, i.e. with 25 female and 24 male participants. Nevertheless, the gender distribution varies, depending on the pairing of the players. For example, there are groups where both of the players are female, or groups with male players, and groups with both genders. The participants come from different countries and span eighteen nationalities, one third of them being native English speakers. More information about the corpus is provided in (?, ?) and at a dedicated webpage.[1]

This dataset addresses multiparty collaborative interactions and aims at providing tools for measuring collaboration and task success based on the integration of the related multimodal information, including collaborative turn organization, i.e. the multimodal turn managing strategies that members of a group employ to discuss and collaborate with each other. The corpus will serve as the knowledge base for identifying measurable behavioral variables of group members with the goal of creating behavioral models. These models may be exploited in human-computer interfaces, and specifically in the design of embodied conversational agents, i.e. agents that need to be able to extract information about their interlocutors to increase the intuitiveness and naturalness of the interaction.

4 Data analysis

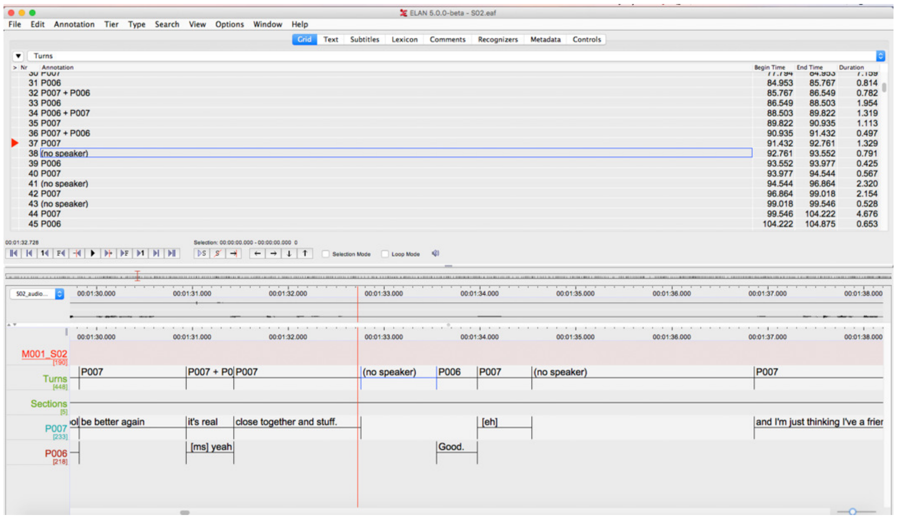
The pauses were manually annotated during the transcription process. The audio signal of the files was then transcribed by 2 annotators using the `Transcriber`

[1] https://www.scss.tcd.ie/clg/MULTISIMO/

tool. [2] The annotators listen to the audio files, segment the speech in turns and transcribe the speakers' speech. Apart from the speaking activity, pause intervals are also annotated by using a 'no speaker' value. The annotators also segmented each group dialogue in 5 topics, i.e. introduction, question 1, question 2, question 3 and closing. Transcripts were then imported into the ELAN annotation editor,[3] so that all the information coming from the transcript was visible and further editable (cf. Fig. 1).

4.1 Manual and Automatic Extraction of Pauses

Using the ELAN search functionality, the transcripts were exploited to create an index where pauses are searchable within their context. We thus exported the list of pauses (i.e. the segments that were annotated with the 'no speaker' value) and their context as a set of concordance lines. Each concordance line includes (a) the filename where the pause occurs, (b) the speaker turn id that precedes the pause, (c) the speaker turn id that follows the pause, (d) the topic or substructure in which the pause occurs, i.e. whether it is in the introduction, the closing or one of the 3 questions part, (e) the start time of the pause, (f) the end time of the pause and (g) its duration. Table 1 presents a sample.



Nr	Annotation	Begin Time	End Time	Duration
30	P007	77.794	84.953	7.159
31	P006	84.953	85.767	0.814
32	P007 + P006	85.767	86.549	0.782
33	P006	86.549	88.503	1.954
34	P006 + P007	88.503	89.822	1.319
35	P007	89.822	90.935	1.113
36	P007 + P006	90.935	91.432	0.497
37	P007	91.432	92.761	1.329
38	(no speaker)	92.761	93.552	0.791
39	P006	93.552	93.977	0.425
40	P007	93.977	94.544	0.567
41	(no speaker)	94.544	96.864	2.320
42	P007	96.864	99.018	2.154
43	(no speaker)	99.018	99.546	0.528
44	P007	99.546	104.222	4.676
45	P006	104.222	104.875	0.653

Fig. 1 Screenshot of a sample file in ELAN. At the top part of the editor the sequence of speaker turns may be viewed, including pause intervals (no speaker). The bottom part includes the timeline with the transcript annotations for words and turns for each speaker, as imported from Transcriber.

[2] http://trans.sourceforge.net/ last accessed 28.02.2018

[3] https://tla.mpi.nl/tools/tla-tools/elan/ last accessed 28.02.2018

Table 1 A sample of exported information about pauses, that includes the filename, the speaker turns before and after the pause, the topic in the discussion where the pause occurs, and its begin and and time and duration in milliseconds

Filename	Turn before	**Pause**	Turn after	Topic	Begin time	End time	Duration
S10	P20	**no speaker**	P22	Introduction	6433	6895	462
S10	P21	**no speaker**	P22	Question 1	93991	94299	308
S10	P21	**no speaker**	P20	Question 2	214733	215300	567
S10	P22	**no speaker**	P20	Question 3	236384	237770	1386

Since the speech pauses were extracted from the manual transcriptions, the transcript annotation serves as the gold standard. However, to test whether pauses can be also automatically detected in a way that is comparable to human annotations in terms of accuracy, we automatically extracted silent pauses using the `auditok` tool.[4] Voice activity detection was then extracted, with the energy threshold for the perceived voice loudness set to 55, where values below this threshold are considered silence, and values above are considered speech. This resulted in a list of time intervals where the voice activity occurs, and this list was used to measure the pauses by calculating the silence intervals that are located between two successive voice activity segments. The results were very similar to the manual annotations, indicating that the set threshold gives reliable results for the specific audio quality of the dataset, by generating silence intervals above 0.2 seconds. However, since the threshold depends on the audio quality and the sensitivity of the microphone used for the recordings, the threshold tuning proves to be an important factor for automatic silence detection.

4.2 Speech Pause Frequency

Overall, 1719 silent and breath pauses were extracted from the 23 dialogs of the corpus in a duration of approximately 4 hours. Because the duration of dialogs varies, the frequency of pauses was calculated with respect to each individual dialog duration. Specifically, we calculated the number of pauses occurring in one minute of a dialog, and the percentage of silence in a dialog, by exploiting the number of pauses and their duration respectively, cf. Table 2. The results confirm previous findings in that speech pauses are frequent in dialog, with a median value of 7 speech pause occurrences per minute, and pause intervals occupy (on median basis) a 14% of the overall dialog duration.

The duration of a silence interval in the data may vary from 0.2 seconds to 10 seconds. However, the median values for the whole corpus fall within the range of 0.5 seconds to 1.7 seconds. A detailed presentation of the silence interval durations

[4] https://github.com/amsehili/auditok last accessed 28.02.2018

is depicted in Fig. 2, which includes the distribution of Min, Max and Median values of the duration of pause segments in the 23 files. Therefore, speakers in the corpus have the tendency to produce short pauses, of less than 2 seconds.

Table 2 Frequency information for speech pauses, i.e. MIN, MAX and MEDIAN values for (a) the number of pauses occurring in 1 minute, and (b) percentage of the duration of pause intervals in the dialogs

Speech pauses	MIN	MAX	MEDIAN
Pauses per minute	2	13	7
Duration in the file (%)	7%	24%	14%

4.3 Silence Intervals and Dialog Topics

All corpus dialogs have a uniform structure, that is, they consist of the introduction to the game, the discussion of the 3 questions and the closing. We then examined the distribution of pauses in the dialog topics, to investigate whether there is an association of the speech pauses with a particular topic in the dialog. Fig. 3 shows the distribution of pauses in the four of the five topics for each of the dialogs, i.e. the percentage of pause interval occurrences in each topic with regard to the total

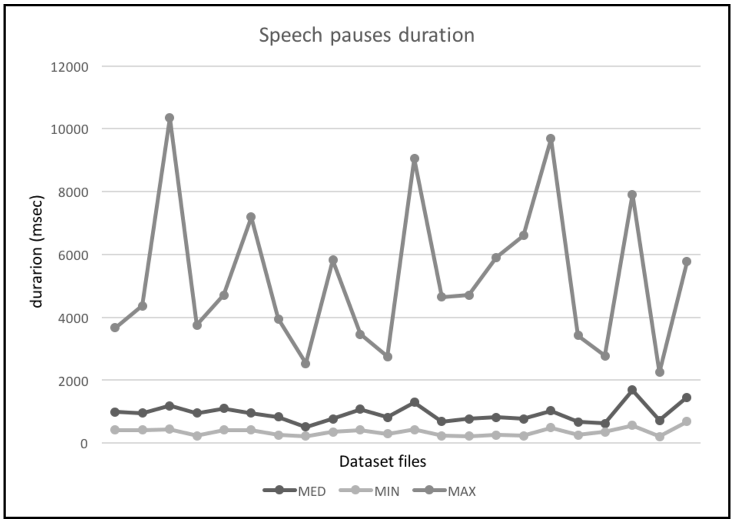


Fig. 2 Distribution of MIN, MAX and MEDIAN values (in msec) for pause duration in the 23 corpus dialogs.

number of pauses in each of the 23 dialogs. No pause segments were identified in the closing section, therefore this topic was left out of the plot. A possible reason for the absence of pauses in the closing section is that it is about a very brief section where the facilitator thanks the players for their participation, therefore there are almost no turn exchanges.

Table 3 lists the Min, Max and Median values of (a) the percentage of the duration of the pause intervals with regard to the duration of each section, and (b) the percentage of pause interval occurrences in each topic with regard to the total number of pauses in each dialog (i.e. a summary of Fig. 3). The shortest (in duration) and fewest (in number) pauses are included in the introduction section. This may be due to the fact that this section is very predictable in terms of turn-taking and has few turn exchanges: it is the section where the facilitator introduces the game and the instructions to the players. The players often acknowledge what the facilitator is explaining with brief verbal feedback, or they may ask brief clarification questions. Nevertheless, the majority of pauses in this topic are performed by the facilitators themselves, who pause briefly to elicit feedback from the players that they follow them, to mark the sequence of instructions, or to take a break as they holds the floor for a long time.

The three questions are the core topics of the dialog, where the participants need to discuss and agree upon the appropriate answers. As expected, pauses are frequent during the question topics; what is interesting though is that Question 2 presents the lowest number of pauses among all 3 questions, and that Question 3 includes both the highest number of pauses, as well as the longest pauses. This observation is interesting in that it is an indicator that Question 3 is more complex than the other two, either because it is more difficult to address or it requires an elaborate

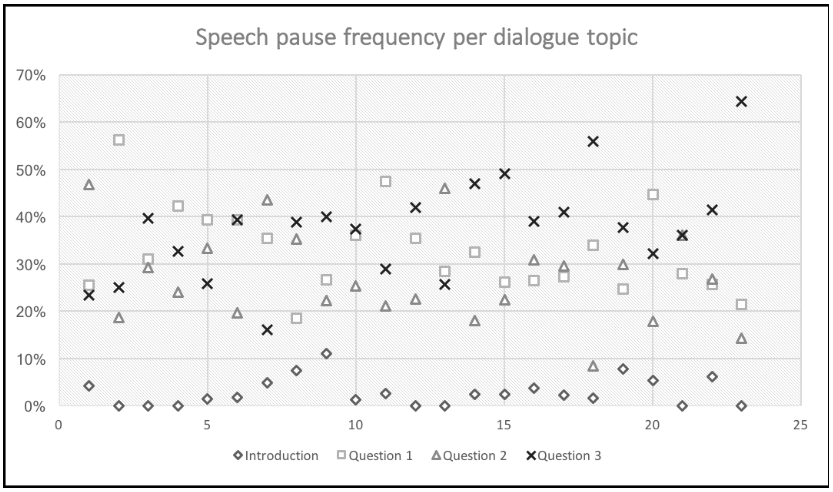


Fig. 3 Distribution of pause occurrences (percentage) in 4 dialog topics (Introduction, Questions 1, 2 and 3) in the 23 corpus dialogs.

Table 3 Distribution of MIN, MAX and MEDIAN values for (a) the percentage of duration of pauses occurring in each of the dialog topics, and (b) the percentage of the quantity of pauses in each of the dialog topics

Topic	Pause duration			Pause number		
	MIN	MAX	MEDIAN	MIN	MAX	MEDIAN
Introduction	1%	11%	5%	1%	11%	3%
Question 1	2%	28%	14%	19%	56%	31%
Question 2	4%	28%	15%	8%	47%	25%
Question 3	11%	34%	17%	16%	64%	39%
Closing	0	0	0	0	0	0

discussion. Both in terms of quantity and duration, the high percentages of pauses in Question 3 may indicate what the literature has often claimed, that pauses are related to cognitive discourse planning; in our case, the speakers may for example need more time to think about potential answers, or may have difficulty in identifying answers to the questions, and speech pauses are a mechanism they employ for this.

4.4 Local Context

As mentioned in Sect. 4.1, the list of pauses was extracted together with the speaker id of the turns that are located before and after the pause interval. This was done to examine the position of pauses with regard to speaker change, and to identify whether we can derive any conclusions about the speaker who takes a turn after a speech pause. Pauses may occur within the same speaker's turns, may be located between the turns of two different speakers, but may also occur before or after simultaneous talk. Fig. 4 presents the percentages of occurrences of the various combinations of the speaker ids that were found before and after a pause interval in the dataset, including cases of overlapping talk.

The majority of pauses (46%) are located among the turns of the same speaker. This case is most probably related to the fact that pauses are a mechanism that speakers employ to plan their discourse and mark its structure, but also to the cognitive processing aspects, i.e. the time speakers need to find appropriate words or think about difficult concepts. Pauses that occur between the two players of the game are also frequent (20%) and indicate that speakers with this role (i.e. player) exchange turns more frequently with each other than with the facilitator. An interesting aspect that cannot be clarified from the pause measurements alone, is that of the way the turn change occurs. It would be important to be able to infer whether the transition from one player to the other is done in a smooth way, i.e. in the case where a player offers the turn to his/her co-player, or whether a speaker takes advantage of a pause to grab the floor.

The next most frequent pattern is that of the pause happening after a facilitator's turn, followed by a turn from one of the players (13%). This is the most frequent case with the facilitator in the left context, and it highlights the role of the facilitator, i.e. to give the floor to the participants after addressing a question or providing helping cues. A 7% of the occurrences refers to cases where the player pauses and the facilitator takes the floor. These cases usually indicate either feedback elicitation, i.e. the players have provided the right answer and await for confirmation, or that the players need help to address the questions posed as they cannot keep on further guessing or discussing potential answers.

An interesting case for further investigation is that of overlapping talk, either before or after the pause. For example, when two players talk simultaneously and after they pause, one of them takes the floor (6% of the cases), it is very possible that an interruption has taken place and it is resolved after the pause. In this respect, pauses may be associated with conversational dominance detection, in the sense that they help identify the person that takes the floor after overlapping talk. However, the frequency of the rest of the patterns where overlapping talk is involved is similar, hence the context is not very informative in providing helpful cues, and information from language or other features is needed to interpret behaviors before or after the overlaps.

4.5 Language Effects on Pause Duration

Silence in conversations is a critical communication device and the beliefs expressed in talk and silence are culture dependent (?, ?). Language and cultural differences are important factors in the analysis and interpretation of pauses. Although the conversations in our dataset were carried out in English, the dataset consists of speakers that are both native and non-native English speakers. At the same time, since most of the cross-lingual and cross-cultural studies on pauses are focused on the pause

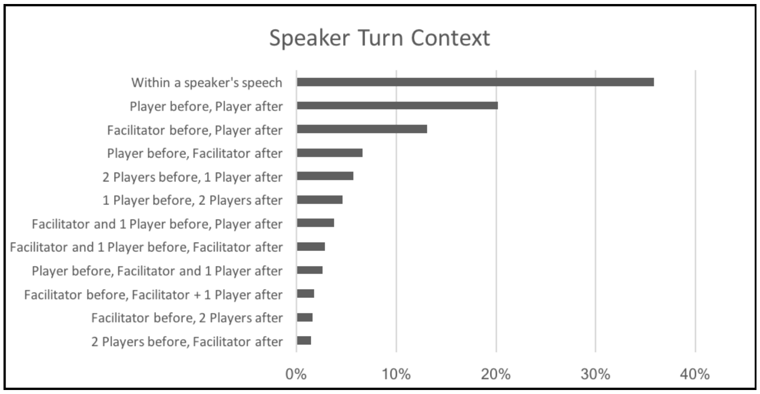


Fig. 4 Distribution of the speaker turn context (i.e. who speaks) that precedes and follows a speech pause. All possible combinations and the related percentage are listed.

length aspect (?, ?, ?, ?), we specifically tested the effects of the native and non-native linguistic background of the corpus speakers on the duration of their pauses.

The dataset includes 16 native English speakers and 33 non-native English speakers. The non-native speakers have a fluent command of English and they span 15 nationalities from Europe, Asia and South America. The majority of the native speakers are Irish[5] (13), while there are 2 speakers from the UK and one speaker from the US.

We took into account two features related to the speaker who holds the floor before and after a pause: the nationality of the speaker and whether that speaker is a native speaker of English or not. In the case of overlapping talk occurring before or after a pause, we considered the aforementioned features for both of the speakers who talk simultaneously.

The results show that there is a significant difference (p=0.01) in pause durations for pauses that are preceded and followed by speech from native speakers of the same language, whatever that language is, and those that are not. The cases where the speech before and after the pause comes from speakers of the same language may refer to pauses that occur within a speaker's speech, or in cases where there is a pause between two turns of speakers of the same linguistic background. In those cases, pauses turned out to be shorter than the pauses that are preceded and followed by speech from native speakers of different languages, with a difference of 267 milliseconds in the mean duration of pauses.

A more detailed analysis of the pause production of the English native speakers shows that there are differences in pause duration patterns among Irish, British and American speakers. Irish and British speakers have longer pause durations for their own pauses than the pauses that they yield, the latter e.g. in cases when the speaker changes after they pause (a difference of 177 and 199 msec. respectively in the pauses mean duration). The American speaker has shorter pause durations for pauses he owns than those he yields (a difference of 244 msec. in the pauses mean duration). Those differences suggest that any generalizations about pause durations appear to require dialect-level articulation rather than larger-language level articulation.

5 Conclusion

In this chapter, we presented a study about silent and breath pauses that occur in a multimodal dataset of collaborative group dialogs. We presented data related to the frequency and the duration of the speech pauses in the corpus as a whole and in the distinct sections of each dialogue where different topics are discussed. Also, we provided measurements related to the context that precedes and follows a speech pause in terms of speaker turns. Finally, we investigated the effects of the linguistic background of the speakers on the duration of pauses they produce. Our observations

[5] All Irish speakers in the corpus are English native speakers and use English in their conversations and elsewhere, and not Irish Gaelic, the first official language in the Republic of Ireland.

confirm that speech pauses are frequent in human dialogs, that they serve several functions and that they are employed by the speakers as a means to structure and emphasise their discourse and give time to reflect the conversation messages. Furthermore, we argue that pauses may provide important cues about the complexity of a given topic, in that this topic elicits more discussion time from the participants or requires more reflexion. We also suggest that, when investigating language and culture specificity in relation to duration aspects of pauses, one should consider both the language and the dialectal variation of the speakers. While exploring the context of the pauses, we focused on the structure of the turns (the sequence of speakers) and not their content. Although in some patterns the sequences are informative per se, in the majority of the cases additional information is needed to draw conclusions about who could be the speaker after a pause. Such information could be drawn from linguistic cues, i.e. the words or the syntactic boundaries of the utterances, but also from information related to other modalities, such as prosodic features or gestures co-occurring with speech. Additional information is therefore needed in cases where pause context patterns, pause frequency and duration do not provide sufficient information. Furthermore, psychological aspects are equally important to co-investigate, as personality and social skills features contribute to the interpretation of speakers' behavior.

Although more information apart from the pause intervals is needed to further investigate the speakers' actions and intentions, we believe that the investigation of pauses is important to achieve better understanding of the human communication and exploit this knowledge in the development of models used in dialog systems and conversational agents. We consider this as a challenging topic and addressing this scientific question in sample data will provide significant input to multimodal big data computing, escpecially in terms of dealing with cognition and understanding complexity. Since multimodal data analysis is focused on how to fuse the information from the different modalities and different features within modalities, to form a coherent decision, speech pause cues should be further explored and included in future investigations on human behavior model development.

Acknowledgements The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 701621 (MULTISIMO).

References

Allwood, J. (1999). The structure of dialog. In M. M. Taylor, F. Neel, & D. Bouwhuis (Eds.), *The structure of multimodal dialogue ii* (p. 3-24). John Benjamins.

Blikstein, P. (2013). Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 102–106). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2460296.2460316 doi: 10.1145/2460296.2460316

Boomer, D., & Dittmann, A. (1962). Hesitation pauses and juncture pauses in speech. *Language and Speech*, *5*, 215-220. doi: 10.1177/002383096200500404

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662-679. Retrieved from https://doi.org/10.1080/1369118X.2012.678878 doi: 10.1080/1369118X.2012.678878

Candea, M., Vasilescu, I., & Adda-Decker, M. (2005, September). Inter- and intra-language acoustic analysis of autonomous fillers. In *DISS 05, Disfluency in Spontaneous Speech Workshop* (p. 47-52). Aix-en-Provence, France. Retrieved from https://halshs.archives-ouvertes.fr/halshs-00321914

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., ... Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In (pp. 413–420).

Chafe, W. (1987). Cognitive constraint on information flow. In R. S. Tomlin (Ed.), *Coherence and grounding in discourse* (p. 21-51). John Benjamins.

Clark, H., & Fox Tree, J. E. (2002, May). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73-111. doi: 10.1016/S0010-0277(02)00017-3

Duncan, S. J., & Fiske, D. W. (1977). *Face-to-face interaction: Research, methods, and theory*. Lawrence Erlbaum Associates.

Egorow, O., Lotz, A., Siegert, I., Bock, R., Krger, J., & Wendemuth, A. (2017, Sept). Accelerating manual annotation of filled pauses by automatic pre-selection. In *2017 international conference on companion technology (icct)* (p. 1-6). doi: 10.1109/COMPANION.2017.8287079

Esposito, A., & Esposito, A. M. (2011). On speech and gestures synchrony. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, & A. Nijholt (Eds.), *Analysis of verbal and nonverbal communication and enactment. the processing issues* (pp. 252–272). Berlin, Heidelberg: Springer Berlin Heidelberg.

Esposito, A., Esposito, A. M., & Vogel, C. (2015). Needs and challenges in human computer interaction for processing social emotional information. *Pattern Recognition Letters*, *66*, 41–51. doi: http://dx.doi.org/10.1016/j.patrec.2015.02.013

Esposito, A., Stejskal, V., Smékal, Z., & Bourbakis, N. (2007). The significance of empty speech pauses: Cognitive and algorithmic issues. In F. Mele, G. Ramella, S. Santillo, & F. Ventriglia (Eds.), *Advances in brain, vision, and artificial intelligence* (pp. 542–554). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of memory and language*, *65 2*, 161-175.

Fujio, M. (2004). Silence during intercultural communication: a case study. *Corporate Communications: An International Journal*, *9*(4), 331-339. Retrieved from https://doi.org/10.1108/13563280410564066 doi:

10.1108/13563280410564066

Gatica-Perez, D., Aran, O., & Jayagopi, D. (2017). Analysis of small groups. In J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, & A. Vinciarelli (Eds.), *Social signal processing* (p. 349-367). Cambridge University Press. doi: 10.1017/9781316676202.025

Goffman, E. (1974). *Frame analysis : an essay on the organization of experience* [Book]. Harper and Row.

Goldman-Eisler, F. (1972, April). Pauses, clauses, sentences. *Language and Speech*, *15*(2), 103-113. doi: 10.1177/002383097201500201

Hirschberg, J., & Nakatani, C. (1998). Acoustic indicators of topic segmentation. In *Proceedings of the international conference on speech and language processing.*

Ishii, R., Otsuka, K., Kumano, S., & Yamato, J. (2016, May). Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, *6*(1), 4:1–4:31. Retrieved from http://doi.acm.org/10.1145/2757284 doi: 10.1145/2757284

Jayagopi, D., Hung, H., Yeo, C., & Gatica-Perez, D. (2009). Modeling dominance in group conversations from non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, *17*(3), 501-513.

Kendall, T. (2013). *Speech rate, pause, and sociolinguistic variation: Studies in corpus sociophonetics* [Book]. Palgrave Macmillan.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Koutsombogera, M., & Vogel, C. (in press). Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Paris, France: European Language Resources Association (ELRA).

Krauss, R. M., Chen, Y., Gottesman, R. F., Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In *In d. mcneill [ed.], language and gesture* (pp. 261–283). University Press.

Künzel, H. (2013). Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech Language and the Law*, *4*(1). Retrieved from https://journals.equinoxpub.com/index.php/IJSLL/article/view/17298

Maatman, R. M., Gratch, J., & Marsella, S. (2005). Natural behavior of a listening agent. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist (Eds.), *Intelligent virtual agents* (pp. 25–36). Berlin, Heidelberg: Springer.

Maclay, H., & Osgood, C. (1959). Hesitation phenomena in spontaneous english speech. *Word*, *15*, 19-44.

McNeill, D. (1992). *Hand and mind : what gestures reveal about thought / david mcneill* [Book]. University of Chicago Press Chicago.

Mohammadi, G., & Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, *3*(3), 273-284. doi: 10.1109/T-AFFC.2012.5

Muñoz-Cristóbal, J. A., Rodríguez-Triana, M. J., Bote-Lorenzo, M. L., Villagrá-Sobrino, S., Asensio-Pérez, J. I., & Martínez-Monés, A. (2017). Toward multimodal analytics in ubiquitous learning environments. In *Mmlacrosslak@lak* (Vol. 1828, pp. 60–67). CEUR-WS.org.

Nakano, Y., & Fukuhara, Y. (2012). Estimating conversational dominance in multiparty interaction. In *Proceedings of the 14th acm international conference on multimodal interaction* (pp. 77–84). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2388676.2388699 doi: 10.1145/2388676.2388699

Narayanan, S., & Georgiou, P. G. (2013, May). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, *101*(5), 1203-1233. doi: 10.1109/JPROC.2012.2236291

Navarretta, C. (2015). Pauses delimiting semantic boundaries. In *Proceedings of the 6th ieee international conference on cognitive infocommunications (coginfocom2015)* (pp. 533–538). IEEE Signal Processing Society.

Oviatt, S., & Cohen, P. R. (2015). *The paradigm shift to multimodality in contemporary computer interfaces*. Morgan & Claypool Publishers.

Raghupathi, W., & Raghupathi, V. (2014, Feb 07). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. Retrieved from https://doi.org/10.1186/2047-2501-2-3 doi: 10.1186/2047-2501-2-3

Rehm, M., Nakano, Y., André, E., & Nishida, T. (2008). Culture-specific first meeting encounters between virtual agents. In H. Prendinger, J. Lester, & M. Ishizuka (Eds.), *Intelligent virtual agents* (pp. 223–236). Berlin, Heidelberg: Springer Berlin Heidelberg.

Rienks, R., & Heylen, D. (2006). Dominance detection in meetings using easily obtainable features. In S. Renals & S. Bengio (Eds.), *Machine learning for multimodal interaction* (pp. 76–86). Berlin, Heidelberg: Springer Berlin Heidelberg.

Rochester, S. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, *2*(1), 51-81. doi: 10.1007/BF01067111

Scherer, K. R. (1979). Personality markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (p. 147-209). Cambridge University Press.

Scollon, R., & Scollon, S. B. K. (1981). *Narrative, literacy, and face in interethnic communication* [Book]. Ablex Pub. Corp Norwood, N.J.

Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., … Bickel, B. (2018). Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences*. Retrieved from http://www.pnas.org/content/early/2018/05/09/1800708115 doi: 10.1073/pnas.1800708115

Siegman, A. W., & Pope, B. (1965). Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. *Journal of Personality and Social Psychology*, *2*. doi: 10.1037/h0022491

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. Retrieved from http://www.pnas.org/content/106/26/10587 doi: 10.1073/pnas.0903616106

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485 - 496. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378216698000149 doi: https://doi.org/10.1016/S0378-2166(98)00014-9

Ting-Toomey, S. (1999). *Communicating across cultures* [Book]. The Guilford Press, New York; London.

Vinciarelli, A., Esposito, A., André, E., Bonin, F., Chetouani, M., Cohn, J. F., ... others (2015). Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, *7*(4), 397–413.

Zhu, W., Cui, P., Wang, Z., & Hua, G. (2015, July). Multimedia big data computing. *IEEE MultiMedia*, *22*(3), 96-c3. doi: 10.1109/MMUL.2015.66