Pause and gap length in face-to-face interaction

Jens Edlund[1], *Mattias Heldner*[1] & *Julia Hirschberg*[12]

[1] KTH Speech Music & Hearing, Stockholm, Sweden
[2] Columbia University, New York, USA

[edlund,mattias]@speech.kth.se, julia@cs.columbia.edu

Abstract

It has long been noted that conversational partners tend to exhibit increasingly similar pitch, intensity, and timing behavior over the course of a conversation. However, the metrics developed to measure this similarity to date have generally failed to capture the dynamic temporal aspects of this process. In this paper, we propose new approaches to measuring interlocutor similarity in spoken dialogue. We define similarity in terms of convergence and synchrony and propose approaches to capture these, illustrating our techniques on gap and pause production in Swedish spontaneous dialogues.

1. Introduction

People engaging in spoken interaction are often observed to grow increasingly similar to their interlocutors as the conversation proceeds. A range of terms have been employed to describe this phenomenon, including entrainment [1], alignment [2], coordination [3], priming [4], accommodation [5], convergence [6], inter-speaker influence [7], and interactional synchrony [8]. In many cases, a particular term has been associated with a specific theory, at least in some instances of its use. In this paper, however, we make no assumptions about the processes underlying this phenomenon or the theories surrounding it. Our aim is instead to explore ways to measure and model the observations of the general phenomenon in a manner that captures something that is lost in many existing studies: the dynamics and temporal aspects.

Most studies of similarity between people interacting apply variants of methods which can be classified into 1) *effects of different conversational partners* and 2) *effects over time of same conversational partner.* Examples of the first approach include correlating two-dimensional coordinates describing the average values for both participants in some interaction across different speaker pairs [9]. Unless there is reason to expect that some speaker pairs are more similar to each other ***before*** the dialogue being analyzed, tendencies of a linear correlation with this method are taken as an indication that speakers have become more similar to each other during the course of the dialogue. Assessment of whether interlocutors become more similar over time in a single conversation has been done, for example, by comparing speaker averages for the first and second halves of each dialogue [7, 10]. If the difference in features examined is smaller in the second half of the dialogue than in the first, this is taken as evidence that speakers have become more similar over the course of the conversation. Both methods and their variants have been used to demonstrate that interlocutors do become more similar to each other in a number of ways, including pitch, intensity, and response latency, inter alia. However, the methods both fail to capture the dynamics and temporal aspects of this similarity. The first method reduces the data to one two-dimensional point per dialogue, losing all temporal information about ***how*** and ***when*** speakers become similar; the second uses two points per dialogue, reducing the temporal information to "early" and "late". A noted exception to these approaches is the time-series modeling approach of [11], which has however not been widely adopted by other researchers, perhaps due to its relative complexity.

The present investigation proposes a method designed to capture the temporal aspects of speaker similarity in an intuitive yet objectively measurable way. As a proof-of-concept, we apply our method to a parameter that is by its nature discontinuous in that it is updated at irregular intervals and never at the same time for both speakers: the length of *pauses* (within-speaker silences) and *gaps* (between-speaker silences) in the terminology of [12]. Our immediate goal is to transform pause and gap length into a continuous parameter and to investigate whether this parameter co-varies dynamically between speakers in dialogue. Our long-term goal is to arrive at a general model which can be used to measure, dynamically, on-line and in real time, variations in similarity between interlocutors on an arbitrary parameter.

In developing this model, we will avoid labels for the general phenomenon and limit ourselves to using three terms in their most general usage, as cited in a standard dictionary. We will talk about *similarity*, *convergence*, and *synchrony*: two phenomena are similar when they are "almost the same", they converge when they "come from different directions and meet", and they are synchronous when they "happen at the same time or work at the same speed" (cf. Longman Dictionary of Contemporary English). We use similarity to refer to the phenomenon in general, and synchrony and convergence to refer to two *ways* of being similar: in converging, two parameters *become more similar* as shown in the left pane of Figure 1. Convergence in our definition captures both the *different conversational partner* and the *over time* measures previously discussed. Synchrony, as seen in the right pane of Figure 1, on the other hand, can occur entirely without convergence, as we will demonstrate below, and captures similarity in relative values rather than in convergence to the same values. We propose that a model of inter-speaker similarity will benefit from combining measures of both convergence and synchrony.

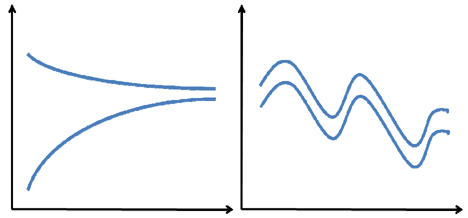

Figure 1: *Schematic illustrations of convergence (left pane) and synchrony (right pane) as they are used in this paper.*

2. Method

2.1. Data

We used data from the Spontal database[1] – audio, video and motion capture recordings of pairs of native speakers of Swedish engaging in 30 minutes of free conversation. It is worth noting that the speakers can see each other and may talk for at least several minutes before the recordings start. The recordings are done with close talking microphones with one speaker in each channel, and the data is labeled with automatic speech/non-speech decisions for each speaker acquired using the VADER voice activity detector from the CMU Sphinx Project[2]. The database includes speaker pairs who are acquainted with one another and those who are not. Here, the first 20 minutes of speech/non-speech labels six random dialogues were used, since the recordings include an external event at 20+ minutes, which would taint the data. No speaker occurs more than once in this subset of the corpus.

2.2. Process

The speech/non-speech labels (one for each 10 ms of dialogue) for both speakers were used to extract all mutual silences in each dialogue. Each mutual silence was automatically labeled as follows: the *instigator* of a silence is the speaker who last spoke before the silence occurred (or who last spoke alone, in cases of a simultaneous end of speech); the *owner* of the silence is the speaker who breaks the silence (or the instigator, in cases of simultaneous start of speech); a *gap* [12] is a silence with a different instigator and owner (aka *inter-speaker silence*); and a pause [12] is a silence with the same instigator and owner (aka *intra-speaker silence*).

All pause and gap durations were then transformed into the log domain. This was done to address the fact that such distributions are typically positively skewed [7, 13], which makes arithmetic means overestimaters of central tendency. Mean durations in the log domain (or geometric means) may be better suited to describe gap and overlap distributions.

Next, the durations were filtered as follows: the pauses and gaps owned by each speaker were treated separately and filtered using a moving window – a 20 point rectangular mean filter – in order to create a smoother sequence. We note that since pauses and gaps occur at irregular intervals, the length of the filter in the temporal domain varies as a function of current gap/pause frequency, although it always contains 20 data points. For completeness, Figure 2 shows histograms (30 second bin size) over the filter lengths for gaps and pauses for all speakers. The most commonly occurring window lengths *in time* are 3-3.5 minutes for gaps and 1.5-2 minutes for pauses. The mean filter takes 20 points to fill up, and we label each of the 20 first data points of each gap and pause sequence for each speaker *low confidence*, as they are less robust.

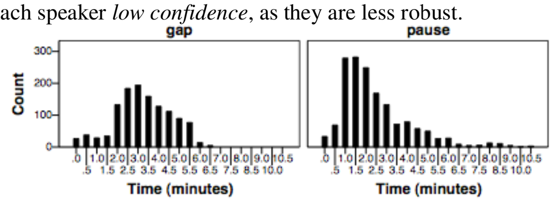

Figure 2: *Histograms over filter lengths of gaps (left pane) and pauses (right pane) in the time domain. The bin size is 30 seconds.*

Finally, we use linear interpolation between the points in each speaker's gap and pause durations, respectively, to make it possible to find a corresponding (derived) value for current mean pause and gap length for each speaker at any given point in time. From these value pairs we calculate the difference between the mean duration of the speakers. To avoid duplicating data, we calculate the differences from the actual instances of one speaker to the corresponding derived values of the other speaker in one direction only (chosen at random for each speaker pair), as illustrated in Figure 3. For each of the calculations involving these differences in the remainder of the paper, we examined reversing the direction and found no noteworthy differences.

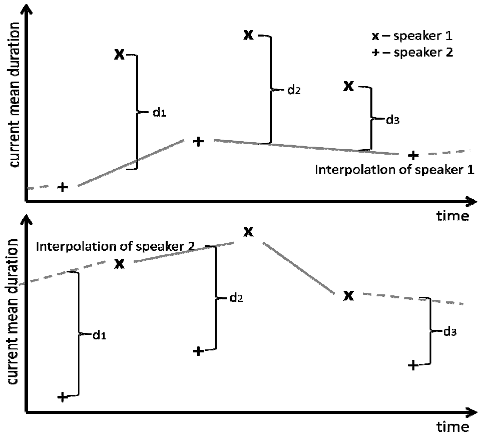

Figure 3: *Schematic illustration of difference calculation. The upper part shows calculation from actual pause/gap instances from speaker 1 to derived values in speaker 2, and the lower pane shows the opposite, given the same data.*

3. Results and discussion

Figure 4 shows plots of the average durations of gaps and pauses in the 6 dialogues calculated using the method of [9] (effects of different conversational partner.

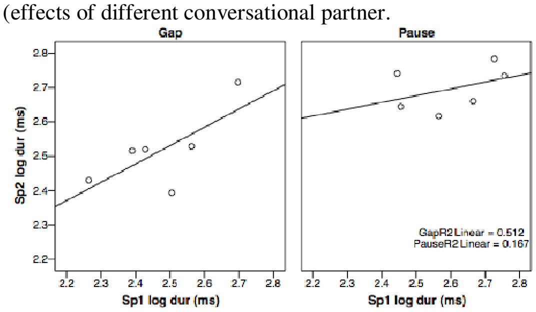

Figure 4: *Scatter plots of mean log duration of gaps (left pane) and pauses (right pane) of speaker 1 on the x-axes, speaker 2 on the y-axes with linear regression lines.*

We note that, for gaps, a linear model explains 51% of the variance (R^2=.51), suggesting that, when temporal information is removed, speakers are more similar to each other than chance would have it. For pauses, however, a linear model explains much less of the variance (R^2=.17) in our data. Using the second method of analysis (effects over time), Figure 5 presents the mean difference between speakers for gaps (left pane) and pauses (right pane) comparing the first ten minutes of dialogue to the following ten minutes in the style of [7, 10].

[1] Swedish Research Council (VR) project 2006-7482

[2] http://cmusphinx.sourceforge.net/

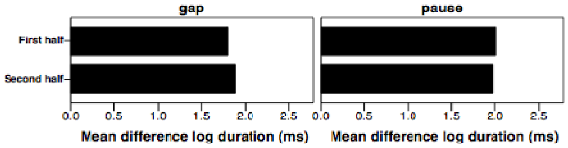

Figure 5: *Mean difference in log duration of gaps (left pane) and pauses (right pane) for all speakers.*

When we compare pauses and gaps in the first and second halves of the dialogues, we observe that pauses appear to converge while gaps diverge. The differences are miniscule, however: on the order of milliseconds. Although Figure 4 suggests that interlocutors converge with regards to gap and pause length, the results in Figure 5 show that this is not a process that can be captured by sampling the distance between speakers at two different stages in the dialogue. It is instead likely that we need dynamic models of convergence to capture the process.

Figure 6 presents plots of each speaker's mean gap and pause lengths over our moving 20-point window over time. In the figure, the x-axes show minutes from beginning of dialogue, while the y-axes show mean duration in milliseconds. The light lines represent the first speaker of each dialogue, and the dark lines the second. The dotted part of each line to the left represents the low confidence values. We note that, in some of the gap panes, particularly in the third and the fourth row, the mean gap length appear to be highly synchronous, whereas in others, particularly in the second (pause) pane, the lines are clearly diverging. In general, the gaps appear to provide evidence of more synchrony than the pauses.

To gauge the strength of the tendencies shown in the figure, we used Pearson correlations to capture 1) convergence (or divergence) by correlating the differences between filtered values from Sp1 and the corresponding interpolated values from Sp2 with the time of their occurrence; and 2) synchrony by correlating filtered values from Sp1 with the corresponding interpolated values from Sp2. Analyses were run separately for gaps and pauses, and split over dialogues (see Table 1) as well as over the pooled dialogues. Low confidence values were excluded.

Regarding convergence, there were no significant tendencies towards convergence when differences were pooled across dialogues. As seen in Table 1, only one dialogue showed significant convergence with respect to gaps and three with respect to pauses. Two dialogues significantly diverged for gaps and one for pauses. Again, this suggests that convergence is not a global phenomenon, and the results in Figure 4 remain unaccounted for. It seems likely, however, that convergence may already have taken place at the beginning of our data, as the interlocutors did have a chance to speak together for several minutes in Spontal recordings.

Table 1: *Pearson correlations of log durations (ms) for speaker 1 vs. interpolated log durations (ms) for speaker 2 for gaps and pauses in the different dialogues. Low confidence values are excluded.*

	Convergence/divergence		Synchrony	
Dialogue	Gap	Pause	Gap	Pause
D1	.568**	.816**	.620**	.305**
D2	-.068	.091	-.262**	-.368**
D3	.086	-.088	.725**	-.008
D16	-.466**	-.751**	.753**	.158
D17	-.170	-.385*	-.301*	.435*
D19	.499**	-.466**	.640**	.697**

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

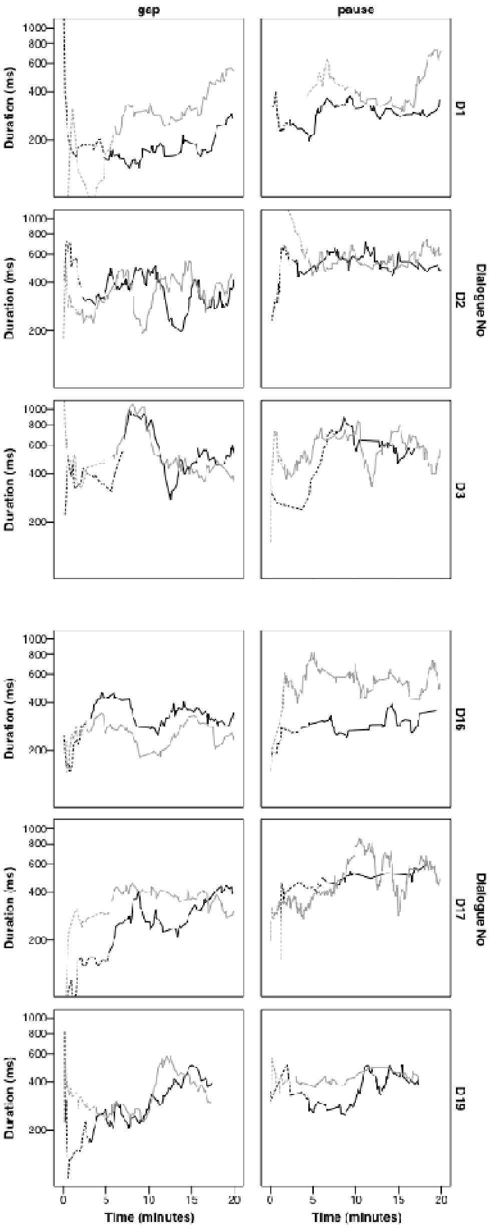

Figure 6: *Mean log duration of gaps (left column of panes) and pauses (right column of panes) over the moving window over the entire dialogue.*

In terms of synchrony, we see in Table 1 that two of the dialogues were negatively correlated with regard to gaps and one with regard to pauses, and that two dialogues showed no significant correlation for pauses. The negative correlations are weaker, however, and when the data is pooled across dialogues, a significant positive correlation appears both for gaps and for pauses (p<.01) – an encouraging sign that dynamic modeling of synchrony may indeed be possible. We suggest then that this approach to measuring the similarity of speaker behavior in dialogue provides a useful objective measure for intuitive observation.

4. Summary and future work

Observations of similarity between interlocutors are plentiful, and numerous studies have investigated whether interlocutors are more similar to each other during a dialogue. There is an underlying assumption that interlocutors in fact *become* more similar over the course of a conversation – an assumption that has been verified in several studies by comparing behavior at the beginning of a conversation to behavior at the end. Although it is clear that temporal aspects and dynamics are at the heart of the phenomena being examined, these are rarely captured in current approaches.

Given the considerable variability of speech in conversation and the large number of factors that influence variation, it is unsurprising that studies of interlocutor similarity have tended to employ gross measures to demonstrate similarity – at the expense of more detailed analysis of how this similarity manifests itself over time. We propose that the time has come to look in more detail at these phenomena and have presented an approach which, we believe, makes such analysis possible. Out of the twelve plots in Figure 6, ten show a significant correlation over time. Given that the plots describe silence durations – a parameter which is inherently discontinuous and which must be transformed in order to make it continuously available in the temporal domain – this gives us hope that similarity between speakers is a phenomenon that can indeed be modeled dynamically.

The present study has proposed a new approach to measuring interlocutor similarity. Our next steps include repeating our analyses on more of the Spontal corpus and on data from other corpora and other languages, including the English Switchboard Corpus [14] and the Columbia Games Corpus [15]. We also plan to investigate different window shapes and lengths. In particular, a decaying window ought to be an improvement over a rectangular, and one involving lighter processing as well as smaller latency so that the measure can be applied in on-line analyses for spoken dialogue systems. We will also test our approach on other parameters that have been shown to become more similar between speakers, including pitch features, energy features, and speaking rate.

Looking further ahead, we are interested in the more general question of measuring the latency of the processes of convergence and synchrony between interlocutors. Do these latencies differ for different parameters, contexts, and speakers or do we find similarities? Do similarity processes differ depending upon which interlocutor precedes the other – that is, is one speaker the leader and the other the follower? We also are interested in explaining why convergence or synchrony over the whole dialogue may be interrupted at certain points, only to return again. For example, the two areas where the curves diverge strongly in the left pane in the second row of Figure 6 or the marked and synchronous rise in the left pane of the third row suggest that similarity in gap behavior has been interrupted for a period of time. We hypothesize that other factors which influence the production of gaps in dialogue may in such cases override the general synchrony of the exchange.

We are ultimately interested in implementing models of similarity in an experimental spoken dialogue system, in order to measure whether such a system is perceived as a better conversational partner, and whether a system producing convergence and synchrony elicits more convergence and synchrony in users than one that does not.

5. Acknowledgements

This research is carried out at KTH Speech Music & Hearing. Funding was provided by the Swedish Research Council (VR) projects 2006-2172 *What makes speech special?* and 2006-7482 Spontal, and by NSF IIS-0307905.

6. References

[1] S. Brennan, "Lexical entrainment in spontaneous dialog," *Proceedings of ISSD*, pp. 41-44, 1996.

[2] M. J. Pickering, and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences,* vol. 27, pp. 169-226, 2004.

[3] K. G. Niederhoffer, and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology,* vol. 21, no. 4, pp. 337-360 2002.

[4] D. Reitter, F. Keller, and J. D. Moore, "Computational modelling of structural priming in dialogue," *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 121–124, New York: Association for Computational Linguistics, 2006.

[5] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context and consequence," *Contexts of Accommodation*, H. Giles, N. Coupland and J. Coupland, eds., pp. 1-68: Cambridge University Press, 1992.

[6] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of the Acoustical Society of America,* vol. 119, no. 4, pp. 2382-2393, 2006.

[7] J. Jaffe, and S. Feldstein, *Rhythms of dialogue*, New York, NY, USA: Academic Press, 1970.

[8] W. S. Condon, and W. D. Ogston, "A segmentation of behavior," *Journal of Psychiatric Research,* vol. 5, no. 3, pp. 221-235, 1967.

[9] L. ten Bosch, N. Oostdijk, and L. Boves, "On temporal aspects of turn taking in conversational dialogues," *Speech Communication,* vol. 47, pp. 80-86, 2005.

[10] N. Suzuki, and Y. Katagiri, "Prosodic alignment in human-computer interaction," *Toward Social Mechanisms of Android Science: A COGSCI 2005 Workshop*, pp. 38-44, Stresa, Italy, 2005.

[11] J. Jaffe, B. Beebe, S. Feldstein *et al.*, *Rhythms of Dialogue in Infancy: Coordinated Timing in Development*: Blackwell Publishing, 2001.

[12] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language,* vol. 50, no. 4, pp. 696-735, 1974.

[13] E. Campione, and J. Véronis, "A large-scale multilingual study of silent pause duration," *Speech Prosody 2002*, pp. 199-202, Aix-en-Provence, France, 2002.

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

[15] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)* pp. 169–172, Columbus, Ohio, USA: Association for Computational Linguistics, 2008.