ARE WE ALL DISFLUENT IN OUR OWN SPECIAL WAY AND SHOULD DIALOGUE SYSTEMS ALSO BE?

Simon Betz, Soledad López Gambino

Bielefeld University, Phonetics and Phonology Workgroup, Dialogue Systems Group
simon.betz@uni-bielefeld.de, m.lopez_gambino@uni-bielefeld.de

Abstract: This study explores inter- and intra-speaker variation in use of time-management strategies. How do speakers differ in their use of pauses, fillers and other resources aimed at managing time while they plan their next contribution? Taking a rather qualitative approach, we describe individual speakers' production, using a small limited-domain corpus of task-oriented German speech. It is assumed that spoken dialogue systems can benefit from mimicking individual speaking styles. This study is a first step in that direction, aiming to describe in detail speaker-specific productions of selected time-buying disfluencies for later use in synthesis.

1 Introduction

Disfluencies, such as hesitations or fillers, are a useful feature of spoken dialogue systems that are capable of real-time interaction with human users [10]. They have the potential to aid incremental systems in buying time and correcting erroneous output [4], they can make synthetic speech sound more natural [1], and can aid reference resolution and operate as indirect clarification requests [5].

For future studies, we plan to address the fundamental question whether characteristics of a particular speaker should be mimicked for disfluency synthesis. [2] have shown that individual synthesized gestures outperform those based on averages of multiple sources. The same might hold true for disfluencies, which like gestures play a key role in conversational speech. This study provides a framework for describing such characteristics as a basis for later modeling.

Previous studies seldom addressed the issue of the "disfluency character" explicitly. [8] notes speaker-specific pause ranges in their study on timing patterns in spontaneous speech, [9] conducts a large-scale analysis of the Switchboard corpus, finding that in terms of repair disfluencies, speakers seem to form two distinct groups. We take the speaker as a starting point and analyze all speech material uttered by each speaker in order to be able to model them individually.

We present analyses of timing and patterns of spontaneous speech elements in a limited-domain corpus [6] as well as a system to describe and compare individual speakers.

2 The TAKE Corpus

2.1 General description

The data consists of approximately 4 hours and 45 minutes of audio and video collected during a Wizard-of-Oz experiment (6 speakers, between 40 and 50 minutes for each[1]), together with

[1]The original corpus collected contains one more speaker, whose data we excluded from the current analysis due to linguistic selection criteria.

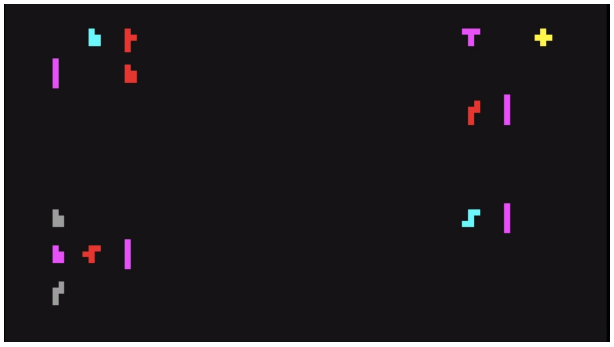

Figure 1 - Screen shown to participants. These are instructed to choose a piece and describe it.

the corresponding transcriptions [6]. Subjects were shown a screen with 15 Pentomino pieces (geometric shapes made up of five squares) of different colors and they were expected to select one of the pieces and describe it (see figure 1). They were encouraged to use pointing gestures alongside speech. Subsequently, the wizard clicked on a piece and, as a consequence, the piece was highlighted on the screen as if it had been selected by the system. The participant then either confirmed the selection verbally or rejected it. In case of rejection, the speaker was expected to reformulate the description until the right tile was chosen. Once a selection had been confirmed, a new screen with a different distribution of pieces was shown and the task was repeated.

2.2 Disfluencies, time management and speech patterns in TAKE

One of the most interesting aspects of this corpus is how easily the connection between time management and disfluencies can be perceived and studied. The simple reference task elicits data which render the analysis of time-management devices more straightforward than in other, more complex speech corpora. In other corpora, fillers and silent pauses are also used for correction purposes whereas, in the TAKE corpus, they are employed mainly to assist time-management.

Moreover, a regular conversational structure for communicating the task emerges in all sessions. For all speakers, there is a strong tendency for utterances to follow the pattern:

1. Episode-initial stretch of silence and optional filler
2. Stretch of low-content words (optional)
3. Stretch of high-content words

1. is the time span from the presentation of a new screen to the first non-filler word a speaker utters. While not being an utterance-initial disfluency, it is still closely related to the topic of analysis, as it reveals how speakers manage to bridge the conversational gap that inevitably occurs while they have to articulate the instructions for the task. It is observable that speakers are struggling with the trade-off between silence being only bearable for a given period of time and content not being ready for articulation. Speakers can solve this conflict in different ways. They can ignore the time pressure and remain silent, they can relax the tension of silence by inserting fillers, they can utter a sequence of low-content words until the actual content is available, or they can find a balance between these strategies. It is assumed that speakers internally strive to achieve said balance but cannot always do so, depending on the complexity of the scene to solve.

2. is an (optional) stretch of words that do not carry much relevant information for the solution of the task. Its size varies depending on how the speaker facilitates the transition from initial

Initial	Low / High Content
Duration of stretch	Number of words
Default: Only silence	Number of Disfluencies
Probability for filler	Type of Disfluencies
Filler duration	Duration of Disfluencies

Table 1 - Information to be extracted per speaker for generating characteristic disfluencies

computing to presenting the content. Speakers with a preference to avoid fillers or silences will produce more low content words. Speakers with a higher tolerance for silences might omit this part entirely. In addition, this part is interesting for speech rate analysis, as lengthening of phones is a common strategy to buy time which manifests preferably on low-content words [8],[3](this volume).

3. is the stretch of words containing the actual piece description. In this task the most important ones are color, shape and location words. The amount of words in this stretch will again vary from speaker to speaker. It is a matter of strategy whether to describe the tile in question as precisely as possible or to try to solve the task with minimal articulatory effort, providing more descriptions only in case of failure.

2.3 Sketching a framework for description

Based on the three stretches identified above, we created a framework for description of generic utterances in this corpus. This framework was used to program a simple generator that uses duration and probability metrics from the analysis to predict number, distribution and duration of disfluencies given a speaker and a number of words per stretch. We plan to expand the generator for inserting disfluencies into incrementally synthesized speech. Table 1 summarizes the information desired for each stretch. Note that low and high content stretches only differ in terms of content but have the same structure. Table 4 summarizes the corresponding values obtained from the corpus. In the next section we describe speaker characteristics that emerged during analysis of the corpus data as well as overlaps and differences between speakers that could hint at candidate strategies to mimic in synthesis.

3 Analysis

3.1 Episode-initial silences

The audio data is segmented into episodes, each one starting when the participant is shown a new screen and ending once the described piece has been identified and the next screen is presented. Most episodes begin with a period of silence while the speaker selects a piece and plans how to describe it. The length of these episode-initial silences shows considerable variability across speakers, with mean values between 1.56 and 4.76 seconds (see table 2). Furthermore, the speech of individual participants also exhibits noticeable variation. As an example, for the participant with a mean duration of 4.76 (speaker 2), the duration of the actual instances go from slightly over one second to almost 12 seconds.

These considerations make it difficult to characterize the time-management style of these speakers with regard to the time they remain silent before they start talking, since no clear tendencies are visible. Further inspections regarding the interplay of initial silence and fillers and of the initial stretch with the subsequent low-content stretch are required.

Speaker	Mean	Std Dev	Min	Max
2	4.762371	1.945617	1.103125	11.847240
3	3.112759	1.420842	1.099911	7.841845
4	2.062363	0.546863	1.134643	4.146375
5	1.557412	0.852793	0.202000	5.234000
6	1.716684	0.650728	0.402625	5.170875
7	4.494371	1.860924	1.829625	10.847511

Table 2 - Duration of episode-initial silences by speaker, in seconds

Speaker	# Initial silences	Initial silences + filler	Initial silences + other
2	66	11 (16.67%)	55 (83.33%)
3	125	2 (1.6%)	123 (98.4%)
4	215	91 (42.33%)	124 (57.67%)
5	131	9 (6.87%)	122 (93.13%)
6	155	10 (6.45%)	145 (93.55%)
7	165	0 (0%)	165 (100%)

Table 3 - Instances of initial silences for each speaker

3.2 Initial silences + fillers

Speakers sometimes add a filler such as ”äh...” or ”ähm...” (”uh” and ”uhm”) after an initial silence, which grants them extra time to complete planning of the upcoming description. It is worth noticing the differences in the degree to which speakers rely on this strategy, since some of them show a stronger preference for fillers in this context whereas others appear to avoid them. This becomes clear by looking at table 3. Speaker 4 has produced 91 instances of fillers after initial silences, whereas none of the other speakers shows more than 11; two speakers (3 and 7) produced only 2 and 0 fillers respectively.

The tendency of some subjects to avoid initial fillers raises an interesting question. Do these speakers exhibit a preference for other strategies instead? The case which most clearly stands out in this respect is speaker 7, who has not produced instances of fillers at the beginning of any episodes. If we look back at the data on initial silence durations, we notice that this speaker has one of the highest mean silence durations (4.49 seconds) and some of the longest maximum values (up to 10.84 seconds). Considering that degree of comfort with silence varies based on individual, cultural and situational factors [7], it could be hypothesized that speaker 7 might be more comfortable with longer silences than most of the other speakers in the corpus, and therefore not experience the need to interrupt initial silences with fillers. Whereas other speakers make use of combinations of silence, fillers and low-content phrases in order to stall for time while planning, this particular speaker shows a tendency to remain silent until he is able to begin describing a piece.

Another interesting observation can be made when observing the structure of the descriptions produced by this speaker. The vast majority of them follow constant, relatively rigid patterns, generally ABOVE/BELOW + LEFT/RIGHT + COLOR + SHAPE, almost without any other words than the ones used to convey these features. Although the data available is not enough to draw conclusions regarding this, it seems clear that having a fixed template of how to present the information and reusing it for virtually every episode has the potential to reduce the cognitive load of the task, thus eliminating the need for fillers and other time-buyers. These observations render the data in this corpus useful for modeling distinctive time-management styles.

Speaker ->	2	3	4	5	6	7
d INI	5.57 / 2.55	3.11 / 1.44	2.63 / 1.16	1.99 / 1.85	1.86 / 0.97	4.45 / 1.91
p Filler in INI	0.17	0.02	0.42	0.07	0.06	0.00
d Filler in INI	0.28 / 0.10	0.60 / 0.21	0.58 / 0.20	1.2 / 0.71	0.43 / 0.06	0.00 / 0.00
p any DF in LC or HC	0.18	0.03	0.13	0.15	0.09	0.04
n words in LC	2.39 / 1.77	1.13 / 0.33	1.45 / 1.51	2.04 / 1.14	4.41 / 3.32	0.06 / 0.24
p FP / SP in LC	0.16 / 0.84	0.50 / 0.50	0.49 / 0.51	0.47 / 0.53	0.04 / 0.96	0.00 / 1.00
d FP in LC	0.42 / 0.09	0.66 / 0.10	0.68 / 0.22	1.09 / 0.51	0.50 / 0.06	0.00 / 0.00
d SP in LC	0.49 / 0.75	1.86 / 2.95	0.31 / 0.32	0.73 / 0.62	0.52 / 0.61	4.86 / 0.00
n words in HC	9.64 / 7.59	3.79 / 2.96	3.45 / 3.33	3.17 / 2.64	3.67 / 3.50	6.28 / 1.66
p FP / SP in HC	0.02 / 0.98	0.00 / 1.00	0.06 / 0.94	0.10 / 0.90	0.04 / 0.96	0.00 / 1.00
d FP in HC	0.40 / 0.10	0.00 / 0.00	0.31 / 0.06	0.69 / 0.47	0.37 / 0.01	0.00 / 0.00
d SP in HC	1.66 / 1.33	0.26 / 0.06	0.18 / 0.13	0.31 / 0.28	0.34 / 0.26	0.71 / 0.62

Table 4 - Overview, all speakers, all stretches. p=probability. d=duration and n=number reported as mean/std. DF = disfluency, INI = Initial Stretch, LC = Low Content Stretch, HC = High Content Stretch.

3.3 Content stretches

3.3.1 General considerations

The information on the initial stretch alone yields no systematic differences between speakers. The interplay between all stretches has to be taken into account for character description. The clear recurring structure of these three subsequent stretches indicates that speakers need a certain amount of time to deliver the relevant content. We assume the computing to be done once the first high-content word is uttered. Our analysis shows that this is a robust indicator: In all high-content stretches of all speakers we find a near or total lack of low-content function words. No speaker had more than four of these words in the high content stretch.

There are, however, different strategies to reach the point where the first high-content word can be uttered. First we check if there is a (negative) correlation between initial stretch duration and low-content stretch size. We suspected that speakers who take more time to formulate in silence would need smaller low-content stretches because the first high-content word is ready earlier. A detailed analysis of this matter is still in progress. First inspections could only reveal slight negative correlations as expected, with the exception of speaker 7, who produced very long initial stretches and avoided the low-content stretch almost entirely.

3.3.2 Disfluency characteristics across speakers

In terms of disfluency usage, speakers seem to form two groups as far as the different stretches are concerned.[2] In the low-content stretches speakers either exhibit a balanced use of silent and filled pauses (Speakers 3,4,5) or a clear preference of silent pauses over filled pauses. In the high content stretch, all speakers, if they produce disfluencies, produce silent pauses only (minimum 94% silent pause rate among all speakers). It might be an explanation for speakers 2 and 7 to avoid fillers because they take most time computing in the internal stretch, but speaker 6 appears to employ a different strategy, having no need for either fillers or long silences in the beginning.

Speaker 3 is another example of avoiding the low-content stretch almost entirely. He also avoids fillers and silences to a great degree, although without a preference for one type or the other. In general he attempts to communicate the goal with minimal verbosity, at the cost of syntax - it is the example farthest away from conversational speech.

Speaker 2 is the last example for a speaker to avoid low-content words. As a striking difference to speakers 3 and 7, this speaker explores the high-content space much more, producing a mean of 9.64 (std: 7.59) words there. This speaker is interesting as he exhibits the greatest total probability of disfluencies (18% of utterances contained one or more) with a preference for silent pauses over filled ones (84% in the lc and 98% in the hc stretch). Since this speaker, despite the long stretches of high-content words, also includes almost no low-content words in between, a more detailed close-up could be fruitful for future analysis.

Speakers 4 and 5 are similar in many ways. The only difference is that speaker 4 uses initial fillers in 42% of the cases compared to 7% of speaker 5, who in turn produces very long initial fillers. In general, these speakers' distribution of information and time-managing low-content words appears balanced over the course of an utterance.

Speaker 6 is markedly different from the others in the sense that he makes much more use of the low-content stretch compared to all others. It is apparent that his very short initial silent periods and his avoidance of fillers is due to the fact that he is using low-content words as his primary time-buying strategy. Among all speakers he is the only one to have the same relatively wide size range for both low and high-content stretch. This is a very interesting subject to base a disfluency character model upon that dynamically produces low-content output until the actual content is available, using lexical items and silences instead of fillers.

3.4 Characters to model

To conclude the descriptive analysis, speakers loosely fall into three groups: One group (speakers 2,3,7) is characterized by a general avoidance of fillers, a high tolerance to remain silent and a clear dispreference of low-content words. The second group (speakers 4,5) allows for a more balanced information structuring and more smoothly bridges conversational gaps. Speaker 6 forms a group of his own, being able to avoid both conversational gaps and fillers by centering on low-content words.

4 Conclusions and future work

We analyzed a corpus of spoken German utterances resulting from a reference resolution task with the aim of exploring differences in time-management strategies and disfluencies to facilitate those. We found that some of the participants exhibit clear tendencies towards either

[2] analyses of the high-content and low-content stretches are based on utterances that contain at least one disfluency. The probability for an utterance to be disfluent is given in the overview table 4.

avoiding or relying on fillers, producing longer or shorter silences while planning their utterances, and preferring either hesitation-like vocalizations ("äh...", "hm...") or sets of low-content words. Speakers can be loosely assigned to three styles of time management. This analysis provides valuable information for modeling speaker characters for conversational speech synthesis, especially the concrete values for duration and distribution obtained here are valuable for future modeling. As a final remark, this study shows that speakers are indeed disfluent "in their own special way", even if, or, especially when, engaged in a simple reference resolution task. With the insights gained here, the question "if dialogue systems should also be" will be adressed in future studies.

5 Acknowledgments

The second author was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

[1] ADELL, J., A. BONAFONTE and D. ESCUDERO-MANCEBO: *On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms*. In *Proceedings of Interspeech*, 2008.

[2] BERGMANN, K., S. KOPP and F. EYSSEL: *Individualized Gesturing Outperforms Average Gesturing – Evaluating Gesture Production in Virtual Humans*. Lecture Notes in Computer Science, pp. 104–117, 2010.

[3] BETZ, S. and P. WAGNER: *Disfluent Lengthening in Spontaneous Speech*. Proceedings of ESSV, 2016.

[4] BETZ, S., P. WAGNER and D. SCHLANGEN: *Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis*. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015, Dresden)*, pp. 2222–2226.

[5] GINZBURG, J., R. FERNÁNDEZ and D. SCHLANGEN: *Dysfluencies as intra-utterance dialogue moves*. Semantics and Pragmatics, 2014.

[6] KOUSIDIS, S., C. KENNINGTON and D. SCHLANGEN: *Investigating speaker gaze and pointing behaviour in human-computer interaction with the 'mint.tools' collection*. Proceedings of Short Papers at SIGdial, 2013.

[7] LUNDHOLM FORS, K.: *Production and Perception of Pauses in Speech*, 2015.

[8] O'SHAUGHNESSY, D.: *Timing patterns in fluent and disfluent spontaneous speech*. In *International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995*, vol. 1, pp. 600–603. IEEE, 1995.

[9] SHRIBERG, E.: *Disfluencies in switchboard*. In *Proceedings of International Conference on Spoken Language Processing*, vol. 96, pp. 11–14, 1996.

[10] SKANTZE, G. and A. HJALMARSSON: *Towards incremental speech generation in conversational systems*. Computer Speech and Language 27, 2013.