Disfluencies and Teaching Strategies in Social Interactions Between a Pedagogical Agent and a Student: Background and Challenges

Tanvi Dinkar[a], Ioana Vasilescu[b], Catherine Pelachaud[c], Chloé Clavel[a]

[a]Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France
[b]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[c]CNRS-ISIR, UPMC, Paris, France
{tanvi.dinkar,chloe.clavel}@telecom-paristech.fr,
Ioana.Vasilescu@limsi.fr, catherine.pelachaud@upmc.fr

Abstract

This paper i) Presents the related work and the challenges regarding the integration of disfluencies in human-agent interactions and, ii) Positions the context and motivations behind our project.

1. Introduction

Disfluencies are breaks, irregularities or non-lexical vocables that occur within the flow of otherwise fluent speech. There are different types of disfluencies, such as word or sound repetitions, fillers/filled pauses (e.g. 'er','um' or 'uh' in English), repairs and so on. They are frequent in spoken language, as spoken language is rarely fluent. An example of their significance in speech can be observed with systems such as Google Duplex: an AI system for accomplishing real world tasks over the phone. A key component to the naturalness of the system was in the incorporation of disfluencies (such as fillers and auto-corrections) in the TTS responses during human-agent interaction (Leviathan et al. 2018). Disfluency has been well studied in cross-linguistic fields and psychology, with a consensus that it is an important tool of speech. They inform us about the linguistic structure of the utterance: such as in the (difficulties of) selection of appropriate vocabulary while circumventing interruption, lexical planning, to build syntactically valid sentences, and to maintain the speaker turn in dialogue. They are linked to deeper meanings of a speaker's emotions, such as fillers and repetitions as an indicator of uncertainty or hesitation (Mifflin, 2000), and to the speaker's Feeling of Knowledge (FOK): i.e the speaker's perception of how knowledgeable they are about a particular topic (Smith and Clark, 1993). Disfluency is also studied as an important tool of communication (Mills, 2014). In speech and language processing, automatic disfluency detection in ASR is typically done with the intent of removing disfluencies from the transcribed text, as subsequent NLP models achieve highest accuracy on syntactically correct utterances. Cleaning speech of disfluency removes the naturalness of speech as well as important information on the cognitive and emotional state of the speaker.

The aim of this project is to study disfluencies in a pedagogical environment in the context of interactions between humans and agents (virtual characters or robots). This project is part of ANIMATAS (Advancing intuitive human-machine interaction with human-like social capabilities for education in schools), an H2020 Marie Sklodowska Curie European Training Network[1]. In this project, we investigate the role of disfluencies in such a context and we will focus on the triangular interaction between the student, teacher and agent, where the agent will learn from both the student and the teacher. An agent could detect and analyse the student's disfluencies, and respond appropriately with (dis)fluent utterances. With the agent's analysis of disfluencies and active use of disfluencies in the student-agent-teacher context, we aim to develop a computational model that will formalise teaching strategies and social interaction based on disfluency, and when to trigger these strategies to help a student in his/her learning phase. Outside of the pedagogical environment, we believe that our work will contribute to

1. http://www.animatas.eu/

dialogue analysis, such as in the agent detecting verbal conflict and measuring the quality of dialogue among interlocutors, as well as in empathetic listening by the agent.

We thus address two research questions in this paper. The first research question is 'What can the agent learn from the user's disfluencies in a learning task?'. For example, disfluency can be an indicator of: i) Uncertainty and feelings of frustration exhibited by the student towards a subject and; ii) The quality of dialogue between the student and teacher and how coordination among them develops. The second research question is 'What are the advantages of the agent's use of disfluencies in speech, where the student is the listener?'. For example, if the agent exhibits uncertainty about a topic through the use of disfluencies, this could help the student to develop important verbal skills by encouraging him/her to respond with better clarity of thought, and participate in topics in which they are not confident. The related work and the challenges pertaining to these two research questions are presented in the two following sections.

2. User's disfluencies in Human-Agent Interactions

In this section, we look at relevant work in cross-linguistics on the functions and factors of disfluency from the user's perspective, and computational studies on the use of disfluencies in speech processing. The research question is the following: How can the agent utilise the user's disfluencies?

There are two main theoretical positions behind the production of disfluencies. One is that disfluencies are accidentally caused in speech due to cognitive burden of the speaker (Bard et al. 2001). Other works study disfluencies as an important communicative function used in dialogue, where convergence on a task is achieved faster due to disfluencies. This is because disfluencies such as clarification requests highlight possible miscommunication that interlocutors may have been unaware of otherwise (Mills, 2014). Often studies will look at both of these positions, by analysing the individual disfluencies of a speaker as well as the collective disfluencies produced by interlocutors. These studies are typically conducted in the context of a task-oriented dialogue between two participants. An unrestrained conversational style dialogue is not usual for this type of study, due to the manual annotation required of the speaker's transcripts. Also, frequency of repairs in dialogue are almost double in task oriented dialogues than in ordinary conversations (Colman and Healey, 2011). Monologues are used to study disfluencies in speakers, but less commonly, because studies have found that speakers are more disfluent in dialogues (Oviatt, 1995). Oviatt (1995) also found that speakers are more disfluent in human-human conversations than human-machine conversations. However, dialogue between human and agent was less sophisticated at the time that the work was published.

Some studies measure disfluency by the frequency of their distribution in dialogue in a particular context. For example, Colman and Healey (2011) show that disfluencies are affected by dialogue role and domain, but not by familiarity or modality (face-to-face versus no eye contact). Measuring speaker intent based on disfluencies is also done by the type of disfluency that occurs in the dialogue. For example, Yoshida and Lickley (2010) studied the effects that disfluencies have on turn taking in establishing common referring expressions between interlocutors, by using a modified HCRC Map task (Brown et al. 1984, Anderson et al. 1991). This task was unlabelled (i.e. landmarks were pictorially represented) to encourage interlocutors to form their own identifying expressions for images, and in doing so produce more disfluencies. They found that fillers frequently occur at the start of discourse, signalling that the subsequent utterance could contain new or unfamiliar information, indicating production difficulties. They also found that self-repairs and speaker modifications tend to occur at the middle of the utterance, indicating a desire for better achievement of the task, showing their communicative function. This shows that the occurrence of different types of disfluencies indicates different speaker intents. Studies also look at the correlation between different factors affecting disfluencies. For example, Branigan et al. (1999) study the non-linguistic factors that affect the rate of disfluency, considering gender, conversational role, ability to see the addressee and practice at the task. Results show that these non-linguistic factors do not steadily affect disfluencies, however they do observe that studying these factors in isolation is an oversimplification: for example repetitions were found to be higher in speakers that cannot see their addressee, though this did not affect the overall disfluency rate.

In emotion detection, Moore et al. (2014) found that disfluency features achieve higher accuracy for emotion detection than lexical or acoustic features. Tian et al. (2015) investigate the usefulness of disfluencies and non-verbal behaviour (DIS-NV) in emotion detection. One finding was that using disfluency features is dependent on the corpus, as the corpus they used contained a mixture of scripted and unscripted data (IEMOCAP database (Busso et al. (2008)); which has fewer examples of disfluencies than the corpus (AVEC2012 database (Schuller et al. 2012)) of spontaneous speech used in Moore et

al. (2014). They conclude that disfluencies could possibly capture high level features in emotion detection that lexical/ acoustic features might omit.

We anticipate challenges in using the above referenced work as a basis to study disfluencies from the user's (student, teacher, or both) perspective in the context of human-agent interaction. We see that different types of disfluencies indicate different cognitive processes of the speaker. However, the rate of different disfluencies is not equal, and hence some types of disfluencies are sparse in data (Moore et al. 2014). Cross-linguistic studies are also conducted on smaller datasets, due to the manual annotation and curation that is required. Apart from insufficient data, there is a question of whether the results of these studies will scale well.

3. Perception and Generation of the Agent's Disfluencies

Many studies focus on the comprehension of disfluent speech, i.e. taking into account the listener's understanding of disfluent speech uttered by the speaker (Corley and Stewart, 2008). This section looks at disfluencies from a listener's perspective. The research question is the following: What are the advantages in the agent's use of disfluencies in speech, where the student is the listener?

Corley et al. (2007) studied the effect of hesitation ('um') on the listener's comprehension using the N400 function of an Event-related potential (ERP), which they establish in predictable versus unpredictable words. The N400 effect can be observed during language comprehension, typically occurring 400 ms after the word onset; and exhibits a negative charge recorded at the scalp consequent to hearing an unpredictable word. In using hesitations preceding the unpredictable word, the N400 effect in listeners was visibly reduced. In a subsequent memory test on the listener, words preceded by hesitation were more likely to be remembered. One drawback however is the processing time hypothesis, i.e. do listeners remember disfluent speech better simply because disfluencies add time to the speech?

Fraundorf and Watson (2011) examined this in a study on how fillers affect the memory of the listeners; by comparing fillers versus coughs of equal duration spliced into fluent speech. Fillers facilitated recall, and coughs negatively hampered recall accuracy. Disfluent speech is hence more likely to be remembered by the listener, and this is not solely based on the additional time of the utterance. They also study comprehension by manipulating the location of the fillers in speech. Fillers typically occur at discourse boundaries, to signal new or upcoming information (Swerts, 1998). However, the authors found that fillers benefit listener's recall accuracy regardless of it's typical or atypical location.

Wollermann et al. (2013) explore the listener's perception of disfluencies using TTS. This is based on the listener's evaluation of how uncertain they think the speaker is regarding a topic, or Feeling of Another's Knowing (FOAK) (Brennan and Williams, 1995). They had the system exhibit 'uncertain' behaviour through disfluent TTS responses in a question-answering context. They found that disfluencies in combination (eg. delays + fillers) increased a listener's perception of uncertainty towards the system's answers. Pfeifer and Bickmore (2009), evaluate an agent that uses fillers 'uh' and 'um' in speech. The motivation behind this was to improve the naturalness of speech in an ECA, as ECAs often try to emulate humans in gestures and facial expressions, yet speak in fluent sentences. Results are mixed, with some participants saying that fillers enhanced the naturalness of the conversation, while others expected that an agent should speak fluently, and fillers were deemed inappropriate. However, further investigation is required, particularly concentrating on the social factors of participants. For example a participants' level of exposure to interacting with an agent could make a difference in their attitude towards the social presence and naturalness of an agent (Goble and Edwards, 2018).

Our goal is for the agent to utilise disfluencies for learning tasks, but also as a response mechanism in human-agent dialogue. For example, when the agent detects a student's possible frustration with a task, responding with similar uncertainty using disfluencies, hence displaying empathy. Although Fraundorf and Watson (2011) extend disfluency studies to a discourse level, these works are not conducted in an active dialogue. The benefits of the agent utilising disfluencies for learning tasks could be dependent on following this format, constraining the student-agent interaction.

4. Conclusion

This paper i) Presented the related work and the challenges regarding the integration of disfluencies in human-agent interactions and, ii) Positioned the context (that is to study disfluencies in a pedagogical environment in the interactions between humans and agents) and motivations behind our project.

Reference

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... & Sotillo, C. (1991). The HCRC map task corpus. *Language and speech, 34*(4), 351-366.

Brown, G., Anderson, A., Shillcock, R., & Yule, G. (1985). *Teaching talk: Strategies for production and assessment.* Cambridge University Press.

Bard, E. G., Lickley, R. J., & Aylett, M. P. (2001). Is disfluency just difficulty?. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech.*

Branigan, H., Lickley, R., & McKelvie, D. (1999, August). Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of the 14th International Conference of Phonetic Sciences* (pp. 387-389).

Brennan, S. E., & Williams, M. (1995). The feeling of another′ s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language, 34*(3), 383-398.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation, 42*(4), 335.

Colman, M., & Healey, P. (2011, January). The distribution of repair in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*(3), 658-668.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass, 2*(4), 589-602.

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.

Goble, H., & Edwards, C. (2018). A Robot That Communicates With Vocal Fillers Has… Uhhh… Greater Social Presence. *Communication Research Reports*, 1-5.

Leviathan, Y., & Matias, Y. (2018). Google Duplex: An AI System for Accomplishing Real World Tasks Over the Phone. *Google AI Blog*.

Mifflin, H. (2000). The American heritage dictionary of the English language. *New York*.

Mills, G. J. (2014). Establishing a communication system: Miscommunication drives abstraction. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp. 193-194).

Moore, J. D., Tian, L., & Lai, C. (2014, April). Word-level emotion recognition using high-level features. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 17-31). Springer, Berlin, Heidelberg.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language, 9*(1), 19-36.

Pfeifer, L. M., & Bickmore, T. (2009, September). Should agents speak like, um, humans? The use of conversational fillers by virtual agents. In *International Workshop on Intelligent Virtual Agents* (pp. 460-466). Springer, Berlin, Heidelberg.

Schuller, B., Valster, M., Eyben, F., Cowie, R., & Pantic, M. (2012, October). AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 449-456). ACM.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of memory and language, 32*(1), 25-38.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of pragmatics, 30*(4), 485-496.

Tian, L., Moore, J. D., & Lai, C. (2015, September). Emotion recognition in spontaneous and acted dialogues. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 698-704). IEEE.

Wollermann, C., Lasarcyk, E., Schade, U., & Schröder, B. (2013). Disfluencies and uncertainty perception–evidence from a human–machine scenario. In *Sixth Workshop on Disfluency in Spontaneous Speech.*

Yoshida, E., & Lickley, R. J. (2010). Disfluency patterns in dialogue processing. In *DiSS-LPSS Joint Workshop 2010.*