Synchronization of (Dis)fluent Speech and Gesture: A Multimodal Approach to (Dis)fluency

Loulou Kosmala, Maria Candea, Aliyah Morgenstern

Synchronization of (Dis)fluent Speech and Gesture: A Multimodal Approach to (Dis)fluency

Loulou Kosmala, Maria Candea, Aliyah Morgenstern

Sorbonne Nouvelle University
13, rue Santeuil. 75005 Paris. France

{Loulou.kosmala, maria.candea, aliyah.morgenstern}@sorbonne-nouvelle.fr

Abstract

Disfluency is verbally expressed by several markers (filled, unfilled pauses, repetitions, self-repairs, etc). This study is grounded in the functionally ambivalent view of *(Dis)fluency* following Crible, (2017) and Götz (2013), but with a multimodal and interactional approach. Previous research has shown a coordination between speech and gesture suspension (Gullberg, 2013, 2018; Seyfedinnipur 2006). The aim of our paper is thus to examine how (dis)fluent speech and gestures can be synchronized, and how visual-gestural features can provide a finer understanding of (dis)fluency. Our analyses are conducted on 3 pairs of French and American speakers interacting both in their L1 and their L2. (Dis)fluency markers were annotated according to their multimodal features. Qualitative analyses revealed how the notions of time suspension and planning associated with (dis)fluency were also found in gesture. This strongly supports the idea that (dis)fluency is to be considered a multimodal phenomenon, and its visual cues are essential for a closer examination of its pragmatic functions.

1 Introduction

In spontaneous typical speech, the course of human language can never be largely continuous, as speakers do not know in advance the specific content they are going to deliver, and how they are going to formulate it. They end up producing a number of "disfluent" utterances in the midst of their discourse. Verbal disfluency is usually defined as a temporary suspension of the speech flow (Ferreira & Bailey, 2004) through filled pauses, silence, repetitions, or whole new utterances. Disfluency is not only a vocal phenomenon and it can be signalled through other modalities: (1) facial expressions, (2) head movements, (3) shoulder movements (Jokinen & Alwood, 2010 p.57). Disfluency markers can also be considered as devices used by speakers to achieve fluency. This paper thus focuses on (dis)fluency as an ambivalent process and on its multimodal features.

A number of studies have been conducted on the relation between speech and gesture following Kendon (2004) or McNeill (1985), but less is known about the relationship between gesture and (dis)fluency specifically. Gullberg (2006) points out two opposite functions served by gestures: (1) an interactional function – gestures that can be useful for turn taking regulation, agreement marking, and attention directing; (2) a self-directed function—gestures addressed towards oneself, dealing with the organization of thought. Similarly, a certain duality can be found in disfluency. Two main views emerge from the literature: (1) Disfluency is the result of speech production "problems" linked to a cognitive load, which disrupt the fluidity of utterances (e.g. Bortfeld et al. 2001; Finlayson & Corley, 2012; Schachter, Christenfeld, & Bilous, 1991); (2) (Dis)fluency markers are communicative strategic devices and time-buying tools which serve discourse planning and structuring functions, and therefore restore continuity in speech. (Allwood, Nivre, & Ahlsén, 1990; Crible, Degand, &

Gilquin, 2017; Götz, 2013; Kjellmer, 2003; Kosmala & Morgenstern, 2019; Swerts, 1998; Tottie, 2014).

In the first view, disfluencies are seen as mostly self-directed, as speakers are trying to deal with production problems, while in the second one, they are mostly interactional as disfluencies can also positively contribute to the interaction. Therefore, recent approaches to *(dis)fluency* highlight their functional ambivalence (Crible, Dumont, Grosman, & Notarrigo, 2019; Götz, 2013): (dis)fluencies can both show signs of *fluency* (more other-directed, contributing to the interaction) and *disfluency* (self-directed, disrupting speech). In line with this approach, this paper is grounded in a multimodal, interactive, functional approach to language captured in situated discourse, and aims to explore the ambivalence of (dis)fluency markers conveyed in various modalities. We examine how discourse suspension and planning associated with (dis)fluency markers can also be conveyed in the visual-gestural channel.

Previous studies have demonstrated the importance of a multimodal approach. Seyfeddinipur (2006) investigated the coordination of speech disfluencies and gestures. Her analysis of speech interruptions and gesture phrases in a corpus study indicated that out of 432 speech suspensions, 306 were accompanied by gestures. This suggested that speech disfluency could affect gesture execution as gestures were suspended at the same time as speech. In an earlier study conducted by Seyfeddinipur & Kita (2001), similar results were found, as they concluded that gestures tended to be suspended prior to the production of speech disfluencies. Gaze could also be seen as an indicator of (dis)fluency. Goodwin & Goodwin (1996) found that speakers frequently gazed away from their interlocutor during word search. They explained that gaze withdrawals usually occurred near "perturbations in the talk displaying initiation of a word search" (p.57). Gestures can also be used to compensate for linguistic problems. In L2 acquisition for example, Gullberg (2006: 108) argues that L2 learners do not only need to acquire grammar and vocabulary, but also "appropriate language use in a broader sense in order to be communicatively competent in a new language". She further suggests that gesture production reflects the planning load. In this perspective, the use of gesture may help L2 learners to keep talking.

Our analysis is conducted on French and American speakers in L1 and L2 productions. Our hypothesis is that the notion of time suspension, which is inherent to speech (dis)fluencies, is also reflected in other modalities; and that the combination of vocal and gestural features can show overt traces of speech processing. The multimodal features of (dis)fluency can thus provide a better understanding of these ambivalent processes.

2 Data, methods, and results

2.1 Data

The data used for our analysis is drawn from the SITAF Corpus (Horgues & Scheuer, 2015) which comprises tandem interactions between French and English native speakers (undergraduate students, aged 18-21) engaged in production tasks in L1-L1 or L1-L2. Our analysis was conducted on 10 video recordings comprising 6 L1-L2 pairings and 4 L1-L1 pairings, and involved 6 subjects: A03, A07, and A13 (American speakers), and F03, F07, and F13 (French speakers). The pairings included either one native speaker and one non-native speaker (tandem condition) or two native speakers (control condition).The participants performed two tasks which involved telling a story and inserting three lies that the partner had to identify (task 1) and discussing a controversial topic and deciding on their degree of agreement (task 2). The tasks were done respectively in their L1 and in their L2. The duration of our selected corpus is approximately 40 minutes. It should be noted that the purpose of this study was not to compare Task 1 and Task 2 nor the tandem and control conditions specifically but rather to focus on the relationship between (dis)fluency and gesture, so the two tasks were grouped together in the results.

2.2 Methods and annotation

The methodology used for our analysis is derived from a previous pilot study (Kosmala & Morgenstern, 2017). In line with Crible 2017 and Crible, Dumont, Grosman, & Notarrigo (2019), the term "sequence" was adopted to refer to the cluster of immediately adjacent (dis)fluency markers

which include: (1) filled pause (uh/um), (2) unfilled or silent pause , (3) syllable prolongations, (4) non-lexical repetitions, (5) self-repairs, (6) self-interruptions, and (7) non-lexical sounds, such as tongue clicks, creaky voice. They were coded according to their position in the utterance, their duration (in ms), and their level of complexity (whether they appear isolated or combined, e.g. *filled pause + unfilled pause*). Their accompanying (total overlap) gestural features were also analysed based on the "gestural phrases" taken from Kendon (2004) and Seyfeddinipur (2006). Gestures were also classified into three functional types, adapted from Kendon (2004) and Gullberg (2011): (1) referential gestures (2) deictic gestures, and (3) pragmatic gestures – gestures not related to the content of discourse but on its structure or "breakdown". 48% of the (dis)fluencies (230 observations) were annotated by a second coder, and received Cohen's Kappa measure of 0.84 for the gesture phrase, and 0.78 for the gesture type. The video recordings were transcribed and coded using ELAN (Sloetjes & Wittenburg, 2008).

2.3 Quantitative results

A total of 475 (dis)fluent sequences were found (279 in L1, and 196 in L2), along with 164 accompanying gestures (80 in L1, 94 in L2). Results show that all speakers mostly kept their hands in rest position (64% of the time, ($p < 0.05$). This is consistent with the view that speakers tend not to gesture when they produce (dis)fluencies (Christenfeld, Schachter, & Bilous, 1991) and that gestures occur much more frequently during fluent speech (Graziano & Gullberg, 2013). However, in cases when speakers did produce gestures, they tended to be suspended or interrupted at the same time as speech, (48% of the time overall ($p < 0.05$). Additionally, there were more gestures co-occurring with disfluent speech in L2 (47%) than in L1 (25%) ($p < 0.05$). This is consistent with the idea that L2 learners produce more gestures in their L2 than in their L1 (Graziano & Gullberg, 2013, 2018). As shown in Table 2, all speakers mostly produced pragmatic gestures during (dis)fluent speech (approx. 70% both in L1 and L2), which stresses the fact that (dis)fluencies do not necessarily relate to the content of the interaction but rather to its structure, or its "breakdown", in line with Gullberg (2011).

Table 2. Number of completed gestures during (dis)fluent sequences

	L1	**L2**
Gest. Type	**45**	**50**
referential	7	8
deictic	4	5
pragmatic	34	26

Cases of interrupted or held gestures indicate a synchronization between speech suspension and gesture suspension, while cases of completed gestures show a relation between gesture activity and planning activity. This will be analysed in detail in the next section.

Speakers averted their gaze 81% of the time when producing (dis)fluent sequences, with no significant differences between L1 and L2 (85% and 82%), which is consistent with Goodwin & Goodwin (1996). Speakers did not often display salient facial expressions during disfluent speech (14%) but they were more prominent in L2 speech (21%) than in L1 speech (9%). This may confirm that L2 is more cognitively demanding than L1, and that speakers are more likely to produce "thinking" gestures in L2 in order to seek help from their interlocutors (Gullberg, 2011). Due to the limited size of the data, and the fact that gestures rarely accompany (dis)fluencies, this paper will now focus on a few qualitative examples from the corpus, drawing more specifically on the relationship between (dis)fluency and multimodality. The notions of time suspension and planning will be explored in further detail.

3 On the relation between (dis)fluency and multimodality: qualitative examples

3.1 Gesture and Speech Suspension

One characteristic of speech (dis)fluencies is that they embody a delay in speech (Schegloff, 2010, Clark & Fox Tree, 2002), as their very presence causes a suspension in speech. The following examples will show how this same suspension is also conveyed in gesture with *holds*. Let us consider

Example (A), which is an utterance taken from the American speaker A13 speaking in his L2 (French)

(A) Je suppose que c'est important de:e (1650) [//] d'être là pour ton ami.
I suppose it's important to:o (1650) [//] to be there for your friend.

The underlined part shows the complex (dis)fluent sequence, which contains a prolongation, (*de:e*) an unfilled pause (*1650 ms*) and a self-repair (*de:e [//] d'être*); the total duration of the sequence is 2.164 ms, which represents a fairly long time of suspension. When looking at its gestural manifestation (Fig.1, first picture) we can see that the speaker is holding his left hand in the same position, and then slowly moves up his right hand until they are both aligned. It is only then that the speaker returns to fluent speech.

Figure 1. Gestural expression in the (dis)fluent sequence

There seems to be a synchrony between the suspension of speech indicated by the unfilled pause and the suspension of the hand gesture epitomized in the hold gesture; but more interestingly, there seems to be a relation between the complexity of the (dis)fluent sequence, which is composed of several different (dis)fluency markers (an unfilled pause, a prolongation, and a self-repair) and the gestural activity which is a combination of a hold (left hand) and a (right) hand movement. Both the production of the (dis)fluent sequence and the manual gesture are then followed by fluent speech.
In example (B), taken from another American speaker (A03) in her L1 (English), the same notion of suspension is found, but this time with a simple (dis)fluent sequence (i.e. no combination).

(B) we:e went afterwards [/] we:e went to his aunt's house/ which is closer to my house/**um** his house is further away.

In this case, the (dis)fluent sequence (the filled pause *um*) lasts 465 ms, so is not as vocally perceptible as in (A), but it is clearly visible in her hand gestures (Fig. 1, second picture). Here the speaker uses two deictic gestures, one directed towards her chest, which points to her house, and the other one directed towards her right, which refers to her boyfriend's house. Between the two descriptions, her hands momentarily return to the same rest position during the short length of the production of the filled pause *um*.
These examples have shown how the retraction, or suspension of a gesture can be synchronized with the production of the (dis)fluency, which corroborates Graziano & Gullberg (2013, 2018)'s findings. This suggests that (dis)fluency is a multimodal phenomenon, as time suspension is conveyed in the two modalities.

3.2 Planning activity

The moment of suspension signaled by (dis)fluencies can also be used for planning purposes; (dis)fluencies can thus be seen as time-buying tools for planning (see Nicholson, 2007; Tottie, 2014). While (dis)fluencies carry no semantic weight, the accompanying gestures can provide visual cues and help understand the pragmatic functions served by the verbal markers. We will be looking at two cases of pragmatic cyclic gestures accompanying the (dis)fluent sequences. Cyclic gestures can be used "in the transition from non-fluent to fluent speech when finding the word/concept" (Ladewig 2011, p.8). The following examples illustrate this point.

(C) Um he was staying at **ou:ur** like dormroom you know there's like 6 beds in there.

This utterance is taken from Participant A07 when performing the production task in her L1. She is talking about her stay in South Africa in a youth hostel and the people she met there. In this example, she produces a simple (dis)fluent sequence (marked by a prolongation of 448 ms) before retrieving

the noun “dormroom”. Figure 2 (first picture) shows that she is producing a cyclic gesture at the same time as the production of the (dis)fluency, and prior to the lexical item to be retrieved. As soon as the target word is found, she gazes back at her interlocutor, and completes her gesture.
The circular movement of the gesture may be an indication that the prolongation serves a word finding function. Therefore, it could be interpreted as a way to facilitate lexical retrieval, as the cyclic movement could refer to the lexical item to be retrieved. Producing the movement may thus help retrieve the word more quickly, following Krauss (1998) and is in synchrony with the prosodic expression (the 448 ms phonemic prolongation).
A similar example is found in F03’s multimodal utterance, also performing the production task in her L1 (French) with another French speaker (F07):

(A) F03 : Alors personnellement pendant les dernières vacances donc pas celles de Noël mais celles **(0.662)** après le petit trou qu’on a eu.
F07: ouais.
*So personally during the last vacation so not Christmas vacation but **(0.662)** after the short gap we had.*

Here, the speaker produces an unfilled pause before planning a rather long prepositional phrase (“après le petit trou qu’on a eu”/after the short gap we had) which probably refers to reading week

Figure 2. Cyclic gesture during word search

at university. However, she does not seem to know (or has perhaps forgotten) how that short break is called, and therefore describes it by using her own words “le petit trou” (the short gap). While she is trying to retrieve the words, she also produces a cyclic gesture for the duration of her pause (Fig. 2, second picture). But as opposed to (C), her gaze is fixed on her interlocutor while she is producing the cyclic gesture. This could indicate that she is seeking help from her interlocutor, but it may also show that the two speakers share common ground; that is, both are students from the same university, so both are aware of what “the short gap we had” refers to. The fact that she is gazing at her interlocutor thus serves an additional interactive function. As a result, her interlocutor answers with the use of verbal backchanneling (“ouais”/yeah), and nods in agreement.
These examples have shown how cyclic gestures, used in similar word-finding contexts, co-occurring with a (dis)fluency marker may indicate that the speaker is currently planning parts of the utterance, but can also determine whether the planning process was more self-oriented, therefore more DISfluent (in C) or other-oriented, more communicative, contributing to the fluency of the interaction (in D).The multimodal features of (dis)fluencies thus allow for a finer understanding of these ambivalent processes.

4 Conclusion

This study of L1 and L2 speakers of French and English has shown that (dis)fluent speech and gestures can be synchronized, as speech and gesture production were sometimes suspended at the same time. Moreover, the gestural features have proven to be useful indicators of the pragmatic planning functions associated with (dis)fluencies. However, gestures were not frequent with disfluent speech. A comparative analysis of fluent speech will thus be explored in a larger dataset for future studies. The quantitative findings suggested a higher gestural activity in L2 than in L1 during disfluent speech, and a higher number of pragmatic gestures during (dis)fluencies, which supports previous findings (e.g. Graziano & Gullberg, 2013, 2018), but more quantitative and qualitative work

needs to be done on those differences. Overall, the findings provide strong support for the idea that (dis)fluency should not only be viewed as a purely verbal and vocal process, but as a multimodal one as well. While vocal (dis)fluency markers are typically non-lexical as they lack propositional content, their co-occurring visual-gestural features can add visual content and richer meanings, thus providing a finer understanding of these ambivalent processes, typical of spontaneous interactions.

References

Allwood, J., Nivre, J., & Ahlsén, E. (1990). Speech Management—on the Non-written Life of Speech. *Nordic Journal of Linguistics*, *13*(1), 3–48.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147.

Christenfeld, N., Schachter, S., & Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, *20*(1), 1–10.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses. *Languages in Contrast*, *17*(1), 69–95.

Crible, L., Dumont, A., Grosman, I., & Notarrigo, I. (2019). (Dis)fluency across spoken and signed languages: Application of an interoperable annotation scheme. In L. Degand, G. Gilquin, & A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (Corpora and Language in Use-Proceedings 4). Louvain-la-Neuve: Presses universitaires de Louvain.

Ferreira, F., & Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, *8*(5), 231–237.

Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: an intentional signal from the speaker? | SpringerLink. *Psychonomic Bulletin & Review*, *19*(5), 921–928.

Goodwin, C., & Goodwin, M. H. (1996). Seeing as a situated activity: Formulating planes. In D. Middleton & Y. Engestrom (Eds.), *Cognition and Communication at Work*. Cambridge: Cambridge University Press.

Götz, S. (2013). *Fluency in native and nonnative English speech* (Vol. 53). John Benjamins Publishing.

Graziano, M., & Gullberg, M. (2013). Gesture production and speech fluency in competent speakers and language learners. *Presentado En TIGER, Tilburg University, Holanda*.

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology*, 9, 879.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL-International Review of Applied Linguistics in Language Teaching*, *44*(2), 103–124.

Gullberg, M. (2011). Multilingual multimodality: Communicative difficulties and their solutions in second-language use. *Embodied Interaction: Language and Body in the Material World*, 137–151.

Horgues, C., & Scheuer, S. (2015). Why some things are better done in tandem. In *Investigating English Pronunciation* (pp. 47–82). Springer.

Jokinen, K., & Allwood, J. (2010). Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging. In *Culture and computing* (pp. 55–70). Springer.

Kjellmer, G. (2003). Hesitation. In defence of er and erm. *English Studies*, *84*(2), 170–198.

Kosmala, L., & Morgenstern, A. (2019). Should "uh" and "um" be categorized as markers of disfluency? The use of fillers in a challenging conversational context. In L. Degand, G. Gilquin, & A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (Corpora and Language in Use-Proceedings 4). Louvain-la-Neuve: Presses universitaires de Louvain.

Kosmala, L., & Morgenstern, A. (2017). A preliminary study of hesitation phenomena in L1 and L2 productions: a multimodal approach. *TMH-QPSR*, 37.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, *7*(2), 54–54.

Ladewig, S. H. (2011). Putting the cyclic gesture on a cognitive basis. *CogniTextes. Revue de l'Association Française de Linguistique Cognitive*, (Volume 6).

Nicholson, H. B. M. (2007). *Disfluency in dialogue: attention, structure and function*. University of Edinburgh.

Schachter, S., Christenfeld, N., & Bilous, F. (1991). Speech Disfluency and the Structure of Knowledge. *Journal of Personality and Social Psychology*, *60*(3), 362–367.

Schegloff, E. A. (2010). Some other "uh (m)" s. *Discourse Processes*, *47*(2), 130–174.

Seyfeddinipur, M. (Ed.). (2006). *Disfluency: Interrupting speech and gesture*. MPI-Series in Psycholinguistics.

Seyfeddinipur, M., & Kita, S. (2001). Gesture as an indicator of early error detection in self-monitoring of speech. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

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

Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, *21*(1), 6–29.