Piloting A Novel Framework for Robust Multimodal Disfluency Detection

Elif Ecem Özkan
Queen Mary, University of London
Mile End Rd, London E1 4NS, UK
e.ozkan@qmul.ac.uk

Julian Hough
Swansea University
Bay Campus, Swansea SA1 8EN, UK
julian.hough@swansea.ac.uk

Patrick G. T. Healey
Queen Mary, University of London
Mile End Rd, London E1 4NS, UK
p.healey@qmul.ac.uk

Lorenzo Jamone
Queen Mary, University of London
Mile End Rd, London E1 4NS, UK
l.jamone@qmul.ac.uk

CHI'23,, April 23–28, 2023, Hamburg, Germany
ACM 978-1-4503-XXXX-X/23/04.
https://doi.org/10.1145/XXXXXXX.XXXXXXX

Abstract

Any agent that interacts with humans on a regular basis should be capable of natural and efficient communication. This is useful for any situation but a key challenge is understanding non-fluent speech. Human dialogue is naturally disfluent, and in some circumstances and for some people it can become pervasive. Importantly, disfluency can be also be a useful resource for building shared understanding. A strategy that is based on the real-time adaptability of human-human interaction could pave the way for a better interaction when the speech recognition fails. We stress the importance of people's ability to make *running repairs* in conversation and discuss how disfluency detection could potentially enhance the performance of voice user interfaces (VUIs) in sustaining mutual understanding. Based on the findings of a human-human conversation study, we outline the elements of a novel multimodal disfluency detection framework.

Author Keywords

human-human interaction; conversational agents; voice user interfaces; human-agent interaction; multimodal communication

CCS Concepts

•Human-centered computing → Natural language interfaces;

Introduction/Background

Conversational Agents (CAs) or Voice User Interfaces (VUIs) face particular design challenges and expectations in the context of natural human dialogue. Improving such interfaces is not only important for establishing more dynamic and satisfactory interactions with people, but also to enhance their ability to adapt to natural human diversity. Currently, the common usage of these systems is mainly transactional, within very simple and well-defined tasks. The limited capabilities of the VUIs is prone to failure even in the slight shift from the expected user input. As as result, they are being used only by people who manage fluent speech, have clear pronunciation and who are mostly native speakers of 'standard' English. Even then, it is still for very limited and clearly defined tasks. The current technical capabilities of speech recognition and language processing -which were previously blamed for all of these faults- have improved drastically, but VUIs have not been able to show the level of flexibility that is expected in natural human dialogue. The main issue, we argue, is that VUIs are not designed according to the realities of natural human conversation.

Recent reviews have criticised how the current state of conversational agents is still far from achieving natural conversation [1], let alone efficient, dynamic and inclusive human-like interaction. Where do we start enhancing the efficiency and the human-likeness of conversation with agents? We suggest to start from looking at *repair* phenomena, more specifically self-repairs (i.e. disfluencies). Repair term comes from the field of Conversation Analysis (CA) which focuses on natural human dialogue. It is the mechanism that people very regularly and creatively use to deal with any "troubles of speaking, hearing and understanding" [25]. Repairs are very frequent, contribute to a dynamic establishment of mutual understanding (common ground) and pervasive in conversation [4, 7] across languages [8] including sign language [20]. Hence, it the *running-repairs hypothesis* proposes that "the coordination of language use depends primarily on processes used to deal with misunderstanding on-the-fly and only secondarily on those associated with signaling understanding" [14]. We focus on the most common type, disfluencies, which are when speakers modify their utterances within their turn by restarting, repeating, or changing their words. They have received significant attention in NLP and there are now systems that can recognise and parse them [18], potentially even in live conditions [22].

Disfluencies can be accompanied by non-verbal signals, such as gestures [17, 13]. In particular cases, bodily signals, specifically manual gestures could constitute 50%-70% of the meaning [10, 15, 16]. The human non-verbal behaviour accompanying repairs are very likely to have additional meaning, or can be used as an additional signal for the detection of disfluencies. We know that listeners actively use disfluencies as a signal for repair and react to it accordingly [4]. Incorporating a multi-modal disfluency detection in VUIs, not only would enable novel interaction strategies for more human-like agent responses [28], but could also significantly benefit the elderly, neurodiverse, non-native speakers or people with hearing impairments. These groups of people may have, in conventional terms, less fluent language.

More specifically, research has shown that older adults are more prone to word finding difficulties [5] and may also produce more disfluencies under stressful conditions [6]. This is not only in the case of human-human interaction. There are specific differences between younger and older adults when speaking to a task-oriented conversational agent [9]. Even though the disfluencies were not annotated or

***Repair* Facts**

Frequent: An estimate for the rate of disfluencies in speech is 6 words per 100 [26]. This number could be higher in other individuals (bilinguals or children have different rates).

Universal: "Huh?" is a universal word, and it is a word [8]. It sounds similar and serves the same purpose for 31 languages, other initiated repair, to signal problems of hearing and understanding.

An example: - Can you get the yell-eh-orange box?

quantitatively analysed in this study, authors point at the importance of disfluencies in these interactions and how it is more likely to be a distinguishing factor for older users [9]. In addition to increasing the quality of interaction for older users, the disfluency detection features in VUIs could even contribute to useful applications within interfaces, for example Alzheimer's Dementia recognition [24].

The Assessment of Challenges and Requirements for Multi-modal Disfluency Detection

Findings from analyses involving dyads in natural dialogue have shown specific motion patterns during disfluencies [27, 28, 29] which could be used in designing novel multi-modal disfluency detection frameworks. The data that provide evidence for this come from a study in which 15-minute face-to-face conversations about designing an apartment happen between 13 dyads. In separate sessions, pairs engage in natural dialogue to design an apartment for them to share within budget (for task specifics, please see [19]). An example of a participant engaging in this task can be seen in Figure 1. During this task, they wear head-trackers on hats and hold controllers which was used to measure the 3D positions of their head and hands. They are also recorded by cameras and microphones. The data was labelled using automatic disfluency detection [18], and time windows containing disfluencies were compared with time windows where no speech disfluency happens (for methodology details, please see [27]). The motion data revealed that the head and hand heights and velocities are significantly higher during windows containing disfluency instances [29].

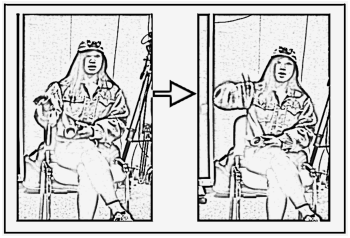

Figure 1: This is a screenshot of consecutive moments from a video on the dataset. The participant uses their hands to explain a furniture that she cannot remember the name of.

A participant is shown in two consecutive frames of a video while she is talking and gesturing.

The mentioned findings have been argued to present evidence for the importance of repair and the way it manifests through non-verbal behaviour in speakers [12]. Arguably, the motion data could be used as additional signals to enhance human-agent interaction. But, how can VUIs, such as Alexa, Apple Siri or Google Home really detect disfluencies easily in real time and work with them, rather than discarding them out as errors?

For successful multi-modal detection in real time and in real life, the VUIs should have certain capabilities. We discuss this briefly in the following sub-sections.

Human Motion Detection

Even though, the aforementioned data have shown statistically significant differences in the motion data, it is unfavourable to measure gestures with big hand-held sensors that may obstruct the natural gesture use. Therefore, the motion data should be collected through non-intrusive methods. Some possible scenarios would be through cameras embedded in the VUI and computer vision algorithms or light devices such as smart watches that could detect the hand positions when worn by the user.

Computer vision based systems and/or algorithms can be susceptible many factors such as camera angles, lighting or objects obstructing the view. It might also be difficult to capture head or hand detection continuously. Is the interface is designed with a camera and expected to capture gestural information this way, it should be communicated to the user to place the device in certain angles and not to obstruct the view. Head, pose and hand detection or activity recognition algorithms should have interpolation strategies for the missed movement information.

Automatic Speech Recognition in Real-time

For automatic speech recognition (ASR) results used for disfluency detection, word-level timing is vital, not only for identifying the exact point at which the repair or hesitation disfluency occurs for more output accuracy, but for our purposes here for using word timing in conjunction with the

concurrent non-verbal behaviour for the input data. Various modern ASRs make this available directly, along with some output for hesitation utterances like ‘uh’ and ‘um’ in English models (see [2]). Word-timing information has shown to improve disfluency in uni-modal disfluency detection [18, 21].

Multimodal Fusion
Needless to say, multimodal fusion of data and decisions are key challenges at both the data integration stage, and the actual disfluency detection stage. The detection system should integrate multiple streams of data with both regular (video and audio) and irregular (words and motion detection) time-stamps. Various types of multimodal fusion technique could be used depending on the need for live results: *early/feature-based fusion* could be used at the acoustic frame level continuously as data arrives into the system, or if live processing is not needed, different forms of *late fusion* can be used whereby results from different detection algorithms working on different modalities separately detect disfluencies, then the results are collated for an ensemble or voting-based detection system for the final decisions. Different combinations of these can be found in *model-based fusion* of different kinds, including *word-based fusion* [23] whereby acoustic and vision features are collated at the end of every recognized word from the ASR for use in a sequence classifier.

Personalisation for Inclusivity
People design the way they communicate according to their recipients and they adapt to them [3]. It is referred to as *recipient design* or *audience design*, and even 5 year-olds do it [3]. They change the way they talk when they are addressing a younger child. Most of us speak louder or gesture more when interacting with someone who has hearing impairment. Even though it is not always clear to humans, people with schizophrenia, on the spectrum or at an early stage of Alzheimer’s differ in their fluency, or use of disfluent speech.

Rule-based Responses from the Agent
It has been innumerably mentioned in human-computer/agent interaction literature how annoying it is to repeatedly hear "Sorry, I don’t understand what you mean by ..." when the speech recognition fails. People who stutter or generally disfluent find it to be very discouraging and they usually end up not using the device. We even hear it makes people "sad" when Alexa does not understand their accent [11].
A main improvement to this feature would be re-modelling the rule-based error response in a way that allows VUIs to work with people dynamically and interactively to achieve mutual understanding. As human beings, we miscommunicate all the time. In fact, human dialogue is rarely fluent without errors [3]. But, we find ways to understand each other through repairing these misunderstandings in creative ways. Moreover, there is information available in disfluencies which actually help listeners compensate for disruptions, delays or reformulations [4].

Conclusion
A closer look at natural human-human interaction can reveal some real-time strategies people use to scaffold shared understanding in communication. Drawing on the findings of a dyadic natural conversation study, we explain our outlook on an important feature for a universal and inclusive VUI design: multimodal disfluency detection. Fundamentally, disfluencies in verbal and non-verbal performances should not be discarded during recognition or data processing. Recognising this is an important step for a collaborative approach to human-agent interaction which is closer to natural human-human interaction.

Acknowledgements

Work supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology at Queen Mary University of London (EP/L01632X/1).

REFERENCES

[1] Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling Feedback in Interaction With Conversational Agents—A Review. *Frontiers in Computer Science* 4 (2022). DOI: http://dx.doi.org/10.3389/fcomp.2022.744574

[2] Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. *Dialogues with Social Robots: Enablements, Analyses, and Evaluation* (2017), 421–432.

[3] Susan Brennan. 2004. Conversation with and through computers. *User Modeling and User-Adapted Interaction* 1 (2004), 67–86.

[4] Susan E. Brennan and Michael F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language* 44, 2 (2001), 274–296. DOI: http://dx.doi.org/10.1006/jmla.2000.2753

[5] Deborah M Burke and Meredith A Shafto. 2004. Aging and language production. *Current directions in psychological science* 13, 1 (2004), 21–24.

[6] Anthony J Caruso, M Troy McClowry, and Ludo Max. 1997. Age-related effects on speech fluency. In *Seminars in speech and language*, Vol. 18. © 1997 by Thieme Medical Publishers, Inc., 171–180.

[7] Marcus Colman and Patrick Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

[8] Mark Dingemanse, Francisco Torreira, and N. Enfield. 2013. Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PloS one* 8 (11 2013), e78273. DOI: http://dx.doi.org/10.1371/journal.pone.0078273

[9] Kallirroi Georgila, Maria Wolters, Johanna D. Moore, and Robert H. Logie. 2010. The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation* 44, 3 (2010), 221–261. http://www.jstor.org/stable/40666360

[10] Jennifer Gerwing and Meredith Allison. 2009. The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture* 9, 3 (2009), 312–336. DOI: http://dx.doi.org/https://doi.org/10.1075/gest.9.3.03ger

[11] Drew Harwell. 2018. The accent gap. *The Washington Post* (2018). https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/

[12] Patrick GT Healey, Marcus Colman, and Mike Thirlwell. 2005. Analysing Multimodal Communication: Repair-Based Measures of Human Communicative Coordination. *Advances in natural multimodal dialogue systems* (2005), 113–129.

[13] Patrick George Healey, Nicola Jane Plant, Christine Howes, and Mary Lavelle. 2015. When Words Fail: Collaborative Gestures During Clarification Dialogues.. In *AAAI Spring Symposia*.

[14] Patrick G. T. Healey, Gregory Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in cognitive science* 10 2 (2018), 367–388.

[15] Judith Holler and Geoffrey Beattie. 2003. How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica* 146 (2003), 81–116. DOI: http://dx.doi.org/10.1515/semi.2003.083

[16] Judith Holler and Katie Wilkin. 2009. Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes* 24, 2 (2009), 267–289. DOI: http://dx.doi.org/10.1080/01690960802095545

[17] Judith Holler and Katie Wilkin. 2011. An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics* 43, 14 (2011), 3522–3536.

[18] Julian Hough and David Schlangen. 2017. Joint, Incremental Disfluency Detection and Utterance Segmentation from Speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 326–336. https://aclanthology.org/E17-1031

[19] Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 1784–1788. https://aclanthology.org/L16-1281

[20] Elizabeth Manrique and N. Enfield. 2015. Suspending the next turn as a form of repair initiation: evidence from Argentine Sign Language. *Frontiers in Psychology* 6 (2015), 1326. DOI: http://dx.doi.org/10.3389/fpsyg.2015.01326

[21] Morteza Rohanian and Julian Hough. 2020. Re-framing Incremental Deep Language Models for Dialogue Processing with Multi-task Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 497–507. https://www.aclweb.org/anthology/2020.coling-main.43

[22] Morteza Rohanian and Julian Hough. 2021. Best of Both Worlds: Making High Accuracy Non-incremental Transformer-based Disfluency Detection Incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 3693–3703. DOI: http://dx.doi.org/10.18653/v1/2021.acl-long.286

[23] Morteza Rohanian, Julian Hough, and Matthew Purver. 2019. Detecting Depression with Word-Level Multimodal Fusion. *Proc. Interspeech 2019* (2019), 1443–1447.

[24] Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs. (2021). DOI: http://dx.doi.org/10.48550/ARXIV.2106.15684

[25] Emanuel Schegloff. 1987. *Recycled turn beginnings; A precise repair mechanism in conversation's turn-taking organization*.

[26] Jean E.Fox Tree. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34, 6 (1995), 709–738. DOI: http://dx.doi.org/https://doi.org/10.1006/jmla.1995.1032

[27] Elif Ecem Özkan, Tom Gurion, Julian Hough, Patrick G.T. Healey, and Lorenzo Jamone. 2021. Specific hand motion patterns correlate to miscommunications during dyadic conversations. In *2021 IEEE International Conference on Development and Learning (ICDL)*. 1–6. DOI:http://dx.doi.org/10.1109/ICDL49984.2021.9515613

[28] Elif Ecem Özkan, Tom Gurion, Julian Hough, Patrick G.T. Healey, and Lorenzo Jamone. 2022. Speaker Motion Patterns during Self-Repairs in Natural Dialogue. In *Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 24–29. DOI:http://dx.doi.org/10.1145/3536220.3563684

[29] Elif Ecem Özkan, Patrick G.T. Healey, Tom Gurion, Julian Hough, and Lorenzo Jamone. 2023. Speakers Raise their Hands and Head during Self-Repairs in Dyadic Conversations. *IEEE Transactions on Cognitive and Developmental Systems* In press (2023).