Deceptive deception: disfluencies are incorrectly interpreted as cues to deceptive speech

Aurélie Pistono (aurelie.pistono@ugent.be)
Department of Experimental Psychology, Ghent University, Belgium
2 Henri Dunantlaan, Ghent, Belgium

Bram De keersmaecker (bram.dekeersmaecker@ugent.be)
Department of Experimental Psychology, Ghent University, Belgium
2 Henri Dunantlaan, Ghent, Belgium

Robert J. Hartsuiker (robert.hartsuiker@ ugent.be)
Department of Experimental Psychology, Ghent University, Belgium
2 Henri Dunantlaan, Ghent, Belgium

Abstract

There is no consensus in the literature about the role of disfluencies as cues to deception. The current study used an interactive picture-description game to collect speech data of speakers and veracity assessments of listeners engaged in a socially meaningful interaction. The paradigm was implemented so that not only statement veracity (i.e., true or false) could be analysed, but also speaker intention (i.e., wanting or not wanting to be believed) and listener decision (i.e., believing or not believing the speaker). The goal was to test whether veracity, intention, and decision could be predicted based on disfluency patterns, using Multivariate Pattern Analysis. We observed that veracity and intention could not be predicted above chance on the basis of disfluency features, while listeners based their decision on these patterns. These results suggest that listeners wrongly interpret disfluencies as cues to deception.

Keywords: language production; disfluency; deception; MVPA

Introduction

The question of whether disfluencies, such as pauses, repetitions, or repairs, can reveal deception is still under debate. Two types of theories have been proposed in deception research: cue theories and non-cue theories (Levine & McCornack, 2014; Levine, 2018). Non-cue theories posit that the psychological processes underlying deceptive and truthful communication are indistinguishable. Therefore, cues cannot be relied upon to detect deception. On the contrary, cue theories assume that lying and truth-telling are psychologically different processes, resulting in different observable behaviours (Levine, 2018). Specifically, a liar experiences cognitive loads such as maintaining coherence, chronology, and logic of a story that did not occur, as well as managing a sincere appearance towards an audience (Levine, 2018).

Within cue theories, the Cognitive Demand Hypothesis (Loy et al., 2018; Vrij, 2000) states that lying requires more cognitive resources than truth-telling. As a result, fewer resources are available to maintain the fluency of the stream of speech, causing liars to be more disfluent. Several studies have supported this hypothesis. For instance, DePaulo et al. (1982) instructed participants to either tell the truth or lie about their feelings towards a person from their personal lives. The participants produced more disfluencies when pretending to dislike someone compared to when they were truthful. Mann et al. (2002) analysed the speech of suspects during police interviews and reported that suspects used longer pauses when lying compared to when they were telling the truth.

In contrast, the Attempted Control Hypothesis (Loy et al., 2018; Vrij, 1995) suggests that liars may be more fluent. Liars often strive to create an honest impression on their audience, appearing as truthful as possible. Furthermore, they may strategically plan their speech and monitor their impression on the listener. As a result, liars may appear more fluent. Indeed, several studies showed that lies elicit *fewer* disfluencies. For instance, Davis et al. (2005) collected videotaped criminal confessions and found that um's and uh's occurred primarily with true statements. Villar and Castillo (2017) analysed the speech produced by personalities on a TV show during a game that required participants to deceive their opponents. They found that true statements contained three times as many um's compared to false statements. Loy et al. (2018) used the treasure hunting paradigm, which we will also employ. During this interactive game, where the goal is to deceive one's opponent, both filled and silent pauses were more frequent during honest speech. In sum, several studies support the use of actual cues to detect deception but the direction of this effect is still debated and it is therefore unclear whether these cues reflect honest speech or deception.

On the contrary, there is a consensus in the current literature regarding perceived cues to deception. Indeed, authors tend to agree that listeners identify perceived cues as indicators of deception (DePaulo et al., 1982; Levine, 2018;

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Loy et al., 2018). More specifically, listeners perceive liars as hesitating more, completing their sentences less, using more redundant repetitions, and uttering more disfluencies such as 'um', 'uh', 'ah', and 'er' (DePaulo et al., 1982; Zuckerman et al., 1981). They also infer that liars talk slower and with a higher pitch (Zuckerman et al., 1981). These results suggest that listeners tend to interpret disfluency as a sign of deception, although it is still uncertain whether this is truly the case.

The differing results regarding actual cues to deception may be attributed to three factors. First, it is possible that individuals exhibit unique behaviours when lying and vary in the degree to which this behaviour is affected by Attempted Control and/or Cognitive Demand. Second, previous studies analysed all disfluencies produced during deceptive speech, but it is highly possible that some of these disfluencies were related to other mechanisms, such as speech encoding issues. Third, some studies did not control for the intention of the speaker (Figure 1). As a result, the truth condition was actually deceptive, with speakers bluffing instead of telling the truth. In other words, participants were producing an utterance that was technically true, but with the intention of misleading the listener. This type of speech is an alternative form of deception because the speaker's intention is to deceive, despite the fact that they are speaking the truth.

The aim of this study is to expand research on the role of disfluencies as cues to deception, both actual and perceived cues, while accounting for the aforementioned limitations. This will allow current theories of verbal cues to deception, namely the Attempted Control and Cognitive Demand hypotheses, to be disentangled. The treasure hunting paradigm (Loy et al., 2018; Vandenhouwe, 2021) will be used, in which two participants play an interactive game against each other. One player takes the role of the speaker, and the other takes the role of the listener. Each round, they are both shown two similar images side by side but only the speaker knows the location of the treasure (Figure 2). The speaker can make a true or false statement about the location, and the listener must then guess which image hides the treasure. If the listener guesses correctly, she will score a point; otherwise, the speaker will score a point.

To dissociate statement veracity and intention, we included an additional condition where the speaker was instructed to tell the truth and was only awarded points if the listener believed them. By comparing this 'honest truth' vs. lie conditions, we can evaluate the effect of veracity while controlling for intention (Figure 1). By comparing this 'honest truth' vs. bluff conditions, we can evaluate the effect of intention while controlling for veracity (Figure 1). To determine the basis on which listeners make their truth assessments, we compared the trials where listeners believed the speaker with those where they did not (Figure 1).

To measure inter-individual differences in disfluency patterns related to veracity or intention, we used Multivariate Pattern Analysis (MVPA; Haynes & Rees, 2006). These analyses not only determine whether two conditions can be distinguished from each other, but also identify whether the pattern of variables is relevant across participants to distinguish (i.e., classify) these conditions. Indeed, speech disfluency measures are typically treated as dependent variables and tested one by one to determine whether they vary between experimental conditions when generalized across participants. In contrast, MVPA tests whether two experimental conditions can be distinguished based on all the information available to the classifier, such as different disfluency measures (Pistono & Hartsuiker, 2021, 2023). Therefore, we applied MVPA to determine if a classifier can predict whether a speaker is lying, bluffing, or telling the truth, and if a listener determined whether a lie or the truth was told, based on disfluency measures. To determine whether veracity and intention can be better classified by certain disfluencies rather than disfluencies produced overall, we conducted two levels of classification: whole utterance versus on the informative (sub)utterance (i.e., only when the speaker was potentially being deceitful in the utterance). This resulted in a total of six classifications.

Informative (sub)utterance analyses were defined as follow (in bold): "*Stereotyped structure favoured by the participant / **part of the utterance that changed every trial***"

- Example 1: *the treasure is behind / **the bird that whistles***
- Example 2: *the money is with / **the broken chain***

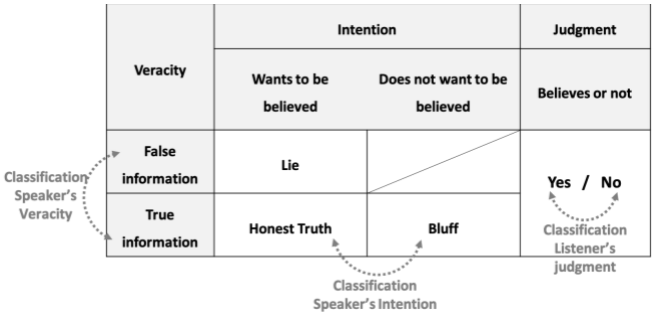

Figure 1: Schematic overview of differences and similarities between lying, bluffing and judgment and corresponding classifications. Each classification will be performed at two levels: whole utterance level vs. Informative (sub)utterance level.

Methods

Participants

Forty-eight native speakers of Dutch (24 duos) were recruited via the Ghent University Research Participation System (14 men, 33 women, 1 other, age=22.5±3.67). None of the duos knew each other prior to the experiment. All participants gave their informed consent.

Material

The material was identical to that of Loy et al. (2018) and Vandenhouwe (2021). The study included 51 pairs of black and white images, each consisting of two visually related images (see Figure 2). This pairing method was chosen to encourage speakers to produce as many, as long, and as complex utterances as possible. Following the design of Vandenhouwe's (2021) study, the trials involved presenting two images of the same pair on separate screens to both the listener and the speaker. One of the images was superimposed on a pile of treasure, while the other was superimposed on a pile of dirt on the speaker's screen. The experiment began with three practice trials to allow participants to become familiar with the game, followed by three blocks of 48 trials each. To establish the honest truth condition, 16 out of the 48 trials in each block were randomly designated as honest truth trials.

Procedure

At the beginning of the session, the speaker and listener were determined by a coin toss. Both participants were informed that they would be playing a game against each other with the goal of earning as many points as possible. A reward of 10 euros would be given to the highest scoring participant at the end of the experimental sessions.

During each trial, the listeners were instructed to identify the image behind which the treasure was hidden. Correct guesses would earn them points. The speakers were given different instructions. They were informed that, unlike the listeners, they could see behind which image the treasure was hidden. In 'normal' trials, the fixation cross appearing before a trial was grey, indicating that they would only receive points if they successfully misled the listener into selecting the incorrect image (the one with the pile of dirt behind it). Participants were given the option to deceive or mislead about the location of the treasure. In 'honest truth' trials, the fixation cross was red, indicating that they were required to truthfully disclose the location of the treasure. In this case, speakers were informed that they would only receive points if they successfully convinced the listener, and both the speaker and the listener would receive points if the listener correctly identified the location of the treasure. The listener saw a grey fixation cross for all trials.

Both participants were fully informed of their opponent's instructions and the reward system. The speakers were instructed to use complete sentences, adequately explain the image content, and remain mindful of their truthfulness throughout the experiment. The listeners were advised to continue making decisions based on the explanations provided by the speakers.

The speaker and the listener sat opposite each other with a screen displaying the experiment, so that they were able to see each other's face but not each other's stimuli. The speech and screens of the speakers were recorded using Open Broadcaster Software (Bailey, 2017), and the decisions, reaction times, and accuracy of the listeners were recorded using Psychopy (Peirce et al., 2019).

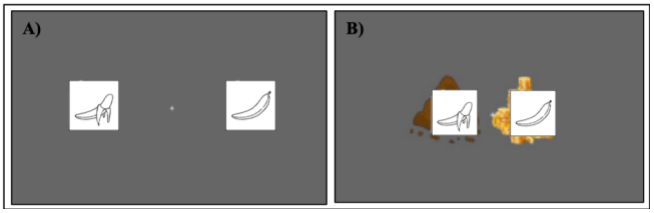

Figure 2: Example of a trial. A) Screen of the listener (left). B) Screen of the speaker (right).

Before each trial, an audio recording announced which trial number was about to start. Then, a fixation cross was displayed on both the speaker's and listener's screens. After one second, an image pair appeared on both computer screens. Upon the appearance of the images on the screen, the speaker could make their statement and then press the 'X' key on the keyboard in front of them. The listener was then able to make a decision between the left or right image on the screen by pressing either the F or J key respectively. Once the listener had decided, the images on both screens disappeared and an inter-trial screen was displayed. The listener could advance to the next trial by pressing the space bar. Every six trials, a feedback message was displayed, informing the participants about the interim score. This feedback message was intended to ensure that participants remained highly motivated during the game.

Variables

A native Dutch speaker (BD) transcribed and categorized all utterances. Veracity was coded as honest truth, deceptive truth (bluff), or deception (lie). 'Honest truth' trials were automatically categorized as honest truth condition, except for the trials where the speaker failed to follow the instructions to tell the truth, in which case the trial was excluded from the dataset. The decision was coded as either 'believe' or 'not believe', indicating the listener's belief in the speaker. The speaker's speech disfluencies were coded and grouped into five categories: filled pauses, silent pauses, repetitions, repairs, and prolongations. Table 1 provides definitions and examples of each disfluency.

Table 1: Definitions and examples of speech disfluencies in each category. The disfluencies themselves are underlined in each example. The English translation is in italic. Filled pauses were broken up further into uh's and um's.

Disfluency category	**Definition**	**Example**
Filled pauses	When the speaker delays the speech stream by inserting a filler (uh, um).	uh de schat zit achter *uh the treasure is behind*

Silent pauses	When the speaker delays the stream of speech by being silent.	de fles met … een *the bottle with … a*
Repetitions	Repetitions of entire words, or part of a word.	dat is zo het het kleinste *that is the the smallest*
Repairs	When the speaker stops a sentence and resumes with a substitution for a word or with the addition of new material	de kast is – de lade van één kast is open *the closet is – the drawer of one closet is open*
Prolongations	When the speaker delays the stream of speech by prolonging a speech sound.	de vis met dee kleine staart *the fish with thee little tail*

After the initial coder rated all utterances for speech disfluency variables, four additional coders, who were also native Dutch speakers, independently transcribed and rated three randomly selected blocks each. This was done to assess intercoder reliability for coding disfluencies. Throughout the transcription and labelling process, all coders were unaware of the veracity of the speakers' utterances, the listeners' decisions, and the nature of the trials. After transcribing all trials and labelling disfluencies, intercoder reliability was calculated between the first coder and the other coders. On average, the other coders agreed with the first coder on 88.69% (SE = 0.012) of all instances across all categories of disfluency (see Table 2).

Table 2: Table of average percentage agree with the first coder (coder1) across disfluencies.

Coder	Coder2	Coder3	Coder4	Coder5
Mean %	89.70	89.47	86.34	89.24

Data analysis

On the speakers' side, MVPA was used to investigate whether the information contained in the pattern of disfluency could be used to accurately classify statement veracity and speaker intention. Likewise, on the listener's side, MVPA was used to investigate whether the same information could be used to accurately classify the listener's decision (i.e., believed or not believed). Classifiers were trained for each participant individually. The training was done using a linear discriminant analysis classifier on all disfluencies. The classifications were performed using a leave-one-out cross-validation approach to ensure unbiased evaluation of classification performance. In each cross-validation fold, the data was split into different folds. The classifier was then trained on data from all but one fold and used on the left-out fold to predict its class membership (respectively, honest truth vs. lie; honest truth vs. bluff; believed vs. not believed). This procedure was repeated until each fold has been the left-out fold once. The accuracy measure used was the proportion of correctly classified trials. Classification accuracies for each analysis were compared to the chance level, which is 50% for a two-class problem, using a one-tailed t-test. Additionally, we identified the disfluencies that consistently contributed to the classification at the group level, by testing whether their mean weight was significantly different from zero, as in Pistono & Hartsuiker (2021, 2023). Two classifications were performed for each analysis: one at the utterance level and one at the informative (sub)utterance level, leading to 6 classifications in total.

Results

Descriptive

Due to a malfunction in the recording hardware during one of the experiment sessions, 60 trials (1.74% of all trials) were excluded from the dataset as they had no viable video recording; 16 trials (0.46%) were excluded due to ambiguity in the described image; 44 trials (1.27%) were excluded as the speaker provided false information during an honest truth trial; and 111 trials (3.21%) were excluded due to unusually extreme reaction times (51 trials) or statement lengths (60 trials) exceeding 3 standard deviations from the mean. A total of 231 trials, which represents 6.68% of all trials, were excluded from the dataset.

The final dataset consisted of 3225 viable utterances (i.e. trials). Of these 3225 utterances, the proportions of honest truths, bluffs and lies were comparable: 1049 utterances (32.53%) were honest truths, 1044 utterances (32.37%) were bluffs, and 1132 utterances (35.1%) were lies. On the listeners' side, listeners decided not to believe 1259 utterances (39.04%; ranging from 7.2% to 62.7% across the group).

Classifications whole utterances

Veracity The six speech variables used to predict the veracity of the utterances were: um's, uh's, silent pauses, repetitions, repairs, prolongations. The mean classification accuracy was below chance level (49.5% ($t(23) = -0.32$, $p = 0.37$, Figure 3A). In other words, it was not possible to discern lying from truth telling based on the patterns in these disfluencies.

Intention When classifying honest trust vs. bluff, the mean accuracy was at chance level (49.8% ($t(23) = -0.19$, $p = 0.43$, Figure 3B). In other words, it was not possible to discern bluffing from truth telling based on the patterns in these disfluencies.

Judgment When classifying listener's judgment based on disfluency pattern, the mean accuracy was above chance level

(61.7% (t(23) = -5.68, p < 0.001). In other words, listeners relied on disfluency patterns to base their decisions (i.e., whether the other participant was lying or not). When looking at the contribution of each variable independently, only repetitions contributed to this accuracy consistently across participants (t(23) = 2.62, p = 0.02, Figure 3C).

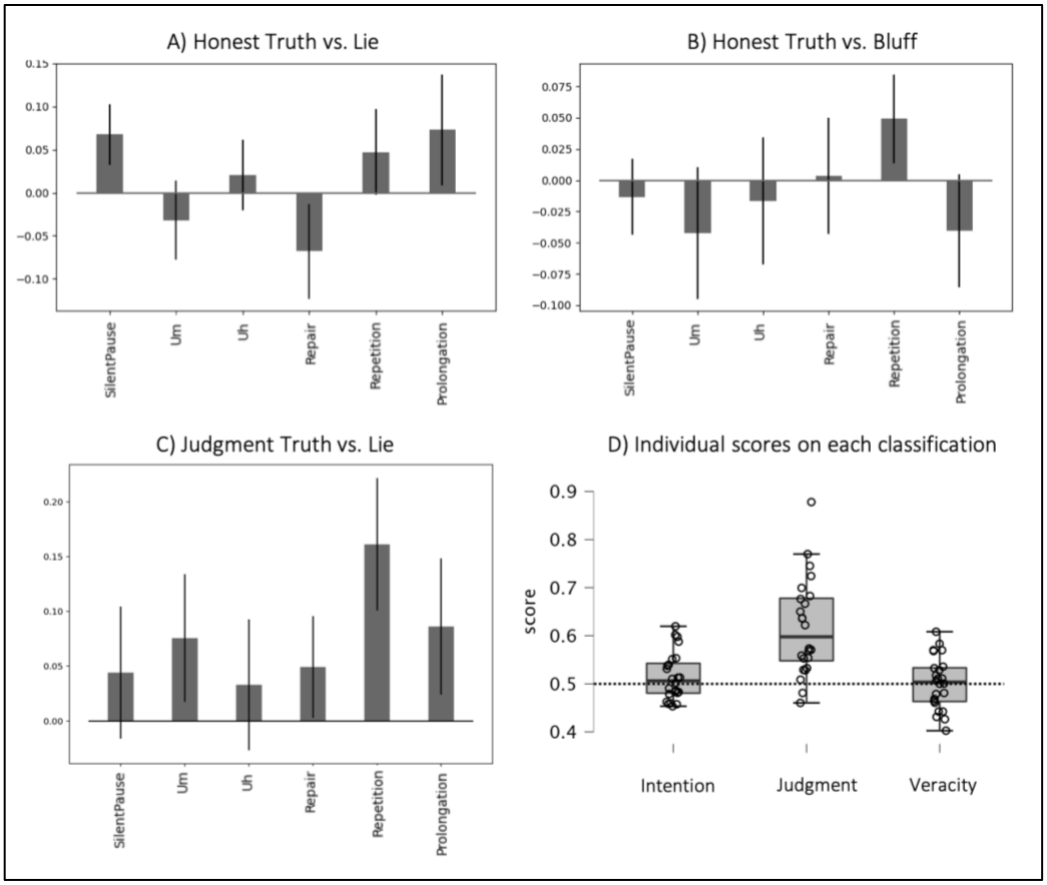

Figure 3: A) Contribution (weight) at the group level of each speech related measure respectively in classifying veracity. Positive means more of the measure in truths compared to in lies. B) Contribution (weight) at the group level of each speech related measure respectively in classifying intention. Positive means more of the measure in honest truths compared to bluff. C) Contribution (weight) at the group level of each speech related measure respectively in classifying decision. Positive means more of the measure in not believed compared to believed trials. D) Classification accuracy for each participant individually on each classification (each point represents a participant).

Classifications Informative (sub)utterance

Similar results were obtained when performing classification at the informative (sub)utterance level compared to the utterance level. Disfluency patterns were not able to predict Veracity and Intention above chance (respectively, 51.6%, t(23) = 1.25, p = 0.11; and 51.5%, t(23) = 1.44, p = 0.08), while they predicted listener's judgment (61.5%, t(23) = 5.34, p < 0.001)).

Discussion

The purpose of the current study was to investigate disfluency patterns as cues to deception, both on the speaker's side and the listener's side. There are two main theories in the current literature on disfluencies as cues to deception, namely the Cognitive Demand Hypothesis and the Attempted Control Hypothesis. In order to differentiate between these two theories, we sought to account for potential confounding factors that have been present in previous studies, namely bluff/truth distinction, inter-individual variability and disfluency location. However, by controlling for these three variables in our analyses, we found that disfluency was not a cue to deception.

First, we tested whether disfluency patterns could classify honest truth vs. deceptive speech. In order to investigate this, the existing treasure hunting paradigm introduced by Loy et al. (2018) was used. This paradigm allows researchers to elicit truths and lies from participants who are embedded in a socially meaningful interaction while something is at stake. The addition of an honest truth condition allowed for a proper investigation of the role of veracity while controlling for

intention. Similarly, the role of intention could be investigated while controlling for veracity. By comparing honest truth with deception and honest truth with bluff, we were able to investigate the hypothesized processes of Cognitive Demand and Attempted Control in a unique way.

MVPA was used to test all measures on the level of individual participants, in order to provide a better understanding of inter-individual variability. However, based on the pattern in the speech disfluency measures, it was not possible to distinguish truth telling from lying above chance level, nor truth telling from bluffing. Similar results were obtained when classifying disfluency patterns at the utterance or informative (sub)utterance level. Initially, two levels of classification were performed to better understand the discrepancy of results regarding disfluencies as cues to deception and to adjudicate between the Cognitive Demand Theory and the Attempted Control Hypothesis. However, our results support non-cue theories -at least for the 6 phenomena under study- as disfluencies did not vary with the experimental condition. These results contradict previous literature, possibly due to the multivariate analyses employed. Figure 3D shows that, for some participants, classification accuracy was above chance. In a univariate framework, this could lead to significant differences due to *some* participants. However, the current results do not show a clear and consistent pattern of disfluency across participants who are lying or bluffing. It is important to note that the study did not consider reaction times before speaking, which could have been a crucial factor to consider. While it is possible that disfluency does not vary with veracity or intention, one can predict that speakers will take more time when they want to mislead their opponent. This prediction would be in line with the Attempted Control hypothesis. Vandenhouwe (2021) used MVPA in a similar paradigm and found significant classifications when comparing truth vs. lie. These classifications indicated that lying was associated with increased fluency, which supports the Attempted Control hypothesis. However, this study did not control for speakers' intention. Additionally, classifications were based not only on disfluency phenomena but also on other related variables, such as utterance duration, disfluent duration, and fluent duration. These variables are highly correlated and therefore increase the performance of the classifier. Nonetheless, it is possible that focusing exclusively on disfluencies is not sufficient, and other variables contributing to speech fluency play a significant role (e.g., duration of pauses, as in Mann et al., 2002; vocal pitch, as in Zuckerman et al., 1981; etc.).

Second, we analyzed disfluencies as perceived cues to deception. These results are in contradiction with classifications performed on the speakers' side, since classifications of listeners' decision based on the patterns in speakers' behavioral cues, were above chance level. In other words, it is predictable to a degree of at least 60% on an individual level if a listener will or will not believe a statement given the pattern of disfluencies produced by the speaker. MVPA showed that classifications are quite stable at the group level, since only two participants did not rely on disfluency patterns above chance. In particular, the production of repetitions was reliably interpreted as a deceptive cue on the listeners' side. These findings are consistent with previous research (DePaulo et al., 1982; Loy et al., 2018; Vandenhouwe, 2021; Vrij, 2008; Zuckerman et al., 1981) and reinforce the idea that listeners tend to agree on which behavioral cues are indicative of deception.

In conclusion, listeners may perceive disfluency as a sign of deception and used disfluency patterns to predict the talker's veracity, even though it may not necessarily indicate this. Future research should consider other temporal variables such as speech onset or speech rate, which may provide insight into speakers' veracity or intention. This would not only provide additional support for existing cue or non-cue theories of deception, it would also enable better analysis of whether veracity and intention elicit different monitoring processes and, therefore, different behavioral patterns.

References

Bailey, H. (2017). the OBS Project Contributors. *Open Broadcasting Software*. Retrieved from https://www.obsproject.org/

Davis, M., Markus, K. A., Walters, S. B., Vorus, N., & Connors, B. (2005). Behavioural cues to deception vs. topic incriminating potential in criminal confessions. Law and Human Behaviour, 29(6), 683-704.

DePaulo, B. M., Rosenthal, R., Rosenkrantz, J., & Green, C. R. (1982). Actual and perceived cues to deception: A closer look at speech. Basic and Applied Social Psychology, 3(4), 291-312.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. Nature Reviews Neuroscience, 7(7), 523–534.

Levine, T. R. (2018). Scientific evidence and cue theories in deception research: reconciling findings from meta-analyses and primary experiments. International Journal of Communication, 12, 19.

Levine, T. R., & McCornack, S. A. (2014). Theorizing about deception. Journal of Language and Social Psychology, 33(4), 431-440.

Loy, J. E., Rohde, H., & Corley, M. (2018). Cues to Lying May be Deceptive: Speaker and Listener Behaviour in an Interactive Game of Deception. Journal of Cognition, 1(1), 42.

Mann, S., Vrij, A., & Bull, R. (2002). Suspects, lies, and videotape: An analysis of authentic high-stake liars. Law and Human Behaviour, 26(3), 365–376.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behaviour made easy. Behaviour research methods, 51(1), 195-203.

Pistono, A., & Hartsuiker, R. J. (2023). Can object identification difficulty be predicted based on disfluencies and eye-movements in connected speech?. *Plos one*, *18*(3), e0281589.

Pistono, A., & Hartsuiker, R. J. (2021). Eye-movements can help disentangle mechanisms underlying disfluency. *Language, Cognition and Neuroscience*, *36*(8), 1038-1055.
Vandenhouwe, N & Hartsuiker, R. J. (2021) Speech disfluencies as actual and believed cues to deception: Individuality of liars and the collective of listeners. in The 10th Workshop on Disfluency in Spontaneous Speech (DiSS 2021), St. Denis, France, August 2021, 17-22.
Villar, G., & Castillo, P. (2017). The Presence of 'Um'as a Marker of Truthfulness in the Speech of TV Personalities. Psychiatry, Psychology and Law, 24(4), 549-560.
Vrij, A. (1995). Behavioural correlates of deception in a simulated police interview. The Journal of Psychology, 129(1), 15–28.
Vrij, A. (2000). Detecting lies and deceit: The psychology of lying and implications for professional practice. New York: Wiley.
Vrij, A. (2008). Detecting lies and deceit: Pitfalls and opportunities. John Wiley & Sons.
Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In Advances in experimental social psychology, 14, 1-59. Academic Press.
Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. Journal of Nonverbal Behaviour, 6(2), 105-114.