Interspeech 2024
1-5 September 2024, Kos, Greece

NumberLie: a game-based experiment to understand the acoustics of deception and truthfulness

Alessandro De Luca[1,2], *Andrew Clark*[1], *Volker Dellwo*[1,2]

[1]Linguistic Research Infrastructure (LiRI), University of Zürich, Switzerland
[2]Department of Computational Linguistics, University of Zürich, Switzerland

alessandro.deluca@uzh.ch

Abstract

To record clearly defined natural deceptive speech with precise knowledge of the ground truth and immediate consequences for the lying subject we present here the NumberLie game. The NumberLie design enables simultaneous and isolated audio recording of five players in our state-of-the-art laboratory, or adapted to any number of players in an online setting, playing against each other in a number-based game revolving around deception and trustworthiness. We describe the technical solutions employed to guarantee precise labelling of statements as truths or lies and immediate consequences to each interaction, backed by a performance-based financial reward to motivate participants. The design is easily manipulated to tailor to specific research questions, maintaining constant or eliminating completely additional sources of variability.

Index Terms: experimental design, deceptive speech, computational paralinguistics

1. Introduction

Designing deceptive speech experiments brings many challenges to researchers, most notably being able to know the ground truth of any statement and providing direct consequences for lying. Many experiments have tried to tackle such challenges with different designs and variable success. Here we propose a design clearly addressing these aspects and providing a versatile platform that can be tailored to various research questions.

Deception, defined as "a deliberate attempt to mislead others" [1], is a common phenomenon of human (and animal) communication. DePaulo's definition of deception can be adapted to various experimental designs, often resulting in broad categorization of deceptive statements. Eliciting deceptive speech naturally in a laboratory setting poses significant challenges. Enos [2] gives a list of criteria to be met to design an *ideal* deceptive speech elicitation experiment as the motivation behind the collection of the CSC deceptive speech corpus [3]. These criteria include the ability to verify the validity of any statement, the recording quality, and the experimental design's resemblance to a real-world scenario, particularly simulating a high-stakes situation with a substantial potential gain for successful deception, balanced by the risk of punishment. We propose adding consideration of the cognitive load hypothesis of deception [4, 5, 6], which suggests that deceiving demands greater cognitive processing than truth-telling. Thus, controlling the cognitive demands of the experiment specific to the deceiving subject is essential. Considering all the technical requirements of a controlled experiment eliciting natural deceptive speech often leads to compromises on some aspects. Table 1 presents recent data collections related to deception, particularly in speech, which have attempted to meet these criteria with varying success in our opinion, often due to design or technical limitations.

One of the first deceptive speech experiments that has attempted to fulfil most criteria previously mentioned is the CSC corpus [2, 3]. This design follows an interview scheme, where participants are asked to lie in certain parts of an interview, with a possible financial reward for successful deception as motivation. One facet we argue is problematic in terms of deception here is how lies and truths are labelled. This is done with a pedal that the participant has to press during their statement, which could have impacts on the production under the perspective of the cognitive load account of deception [6]. A more important aspect that is dubious is the fact that although one may lie, how they do so is impossible to elucidate. For example, lies could be a simple exaggeration of a true fact and not a completely false statement. This is common for many other designs, including more modern ones, such as CSX [7], DSN [8], DDD [9], and those based on interactive games like "Secret Hitler" [10] and C2W2D [11]. A final component that is missing from all designs listed in Table 1, is a direct and immediate consequence of deception.

Table 1: *Summary of recent deception experiments. (Type=data medium, Style=type of design; 2P=2 players).*

Corpus	Type	Style
CSC [3]	Audio	Interview
DyViS [12, 13]	Audio	Mock police interview
Meyer [5]	fMRI	Game (2P)
CSX [7]	Audio	Interview ("fake resume")
DSN [8]	Audio	Interview
DDD [9]	Audio	Game (2P - questionnaire)
"Secret Hitler" [10]	Audio	Social Game
C2W2D [11]	Video	Social Game

2. The NumberLie game

The "NumberLie" game (henceforth *NL* game) can be played between three to any number of players (up to five in our physical facility). With this design our goal is to meet the prescribed criteria of an ideal deceptive speech experiment, focusing on capturing clearly defined lies in a minimally demanding setting for participants, with immediate consequences for lying and comparable speech segments between participants. We utilize verifiable statements represented by numbers to enable accu-

10.21437/Interspeech.2024-2082

rate labelling of truths and lies. Additionally, we employ carrier sentences including monophthong long-vowel fantasy names to ensure acoustically comparable speech segments between subjects.

2.1. Gameplay and rules

The gameplay is very simple:

1. At the beginning of the game (or of a round) a player is randomly selected to start. That player receives a random number and then passes *a number* to the next player using a carrier sentence (e.g. "My name is Lee, I have number 10, and I pass it on to Sam").
2. The player that receives the number (e.g. Sam) can either trust the previous player (e.g. Lee) by continuing to play or doubt (challenge) the previous player by explicitly declaring it (e.g. "I doubt!"):
   a. If Sam trusts Lee: Sam receives a random number and passes *a number* on to the next player.
   b. If Sam doubts Lee: Lee's statement (passed number) is checked against the number generated for that turn (i.e. the number drawn by Lee). The game then resumes with a new round (point 1).

We intentionally use "*a number*" and not "the number" to highlight that players are allowed to choose any number to pass (in accordance with the game rules). The drawn number can be lower than the one received by the player, in which case, if they trust the previous player, they must lie. A round ends if a player challenges or receives a penalty. There are only two rules to follow:

1. The number passed must be higher than the one received (resets when a round ends).
2. The players must pass the number using the carrier sentence: "My name is ..., I have number ..., and I pass it on to ...".

Failure to follow these rules results in a penalty: a points deduction equal to the highest possible in the scoring system (in the current design: -2).

2.2. Scoring system and financial incentive

To incentivise deception in a high-stakes situation we decided to create a scoring system inspired by popular examples from game-theory. The scoring system is outlined in Table 2. The primary goal is to create a scenario where deceiving carries significant stakes, offering a large reward for successful deception with an equivalent risk. This principle is mirrored on the receiving player's side, with the same reward for deception detection and a corresponding loss for being deceived. Additionally, we wanted to give a smaller benefit to two players involved in a cooperative interaction (i.e. truth from the presenter and trust from the receiver), with a marginal cost for both players in case of unsuccessful cooperation (receiver challenges truth).
Before the game, players are made aware that their final compensation will depend on their performance in the game. Each player starts with 10 points and gains or loses points based on their interactions with others, as outlined in Table 2. At the game's conclusion, each player receives an additional payment, in addition to the fixed compensation for participation in the experiment, equivalent to their share (calculated in CHF 5 increments) of a prize pool (CHF 50). This share is determined by the proportion of their score to the sum of all strictly positive scores of all players.
The scoring system in combination with the respective financial reward provides an easy to manipulate platform to investigate the impact of perceived incentive on the players' strategies.

2.3. Ethics

This study and experimental design have been approved by the PhF Ethikkommission of the University of Zürich (# 23.08.06). Participants are briefed on the game rules, mechanics, and rewards prior to playing. Participants are also informed that the experimenter does not know the validity of statements during the game and that no judgment on their character will be made.

Table 2: *NL game scoring system. Each cell refers to one interaction.*

Presenter \ Receiver	Trust	Challenge
Truth	$+1$ / $+1$	-1 / -1
Lie	-2 / $+2$	$+2$ / -2

3. Methods

The *NL* design is adaptable for online and laboratory settings. The current implementation lasts ≈ 1 h and operates in the state-of-the-art LiRI laboratory to ensure control over playback quality and achieve the highest recording quality for comprehensive acoustic analysis.

3.1. The LiRI laboratory

The laboratory houses five double-walled recording booths interconnected with both audio and video feeds allowing participants to hear and see each other and the audio feeds to be recorded (during the *NL* game, video feeds between the booths are turned off). Each booth also includes a monitor for presenting stimuli. All booths are fitted with a cardioid condenser microphone and studio-quality headphones. Recordings from every booth are mono 24-bit 48 kHz PCM WAV. Each booth can be simultaneously recorded on a separate channel since the laboratory was specifically designed with this type of study in mind. Therefore, we can record conversations among multiple speakers individually, eliminating overlapping speech. Each booth is fitted with a Raspberry Pi 4[1] equipped with 8 GB of RAM connected to the stimuli monitor. The Raspberry Pi are connected via a LAN to a memcached [14] server and are running a simple script to retrieve player messages from the server and display them on the monitor. A control computer runs the game code and is connected to the memcached server, setting the server messages for each client (booth). The client messages always include the player's name (fantasy name) and score. Whenever it is a player's turn their message will also contain the number they have "drawn", the received number, and the carrier sentence completed with all information apart from the number (i.e. "My name is Lee, I have number ..., and I pass it on to Sam."). A visualisation of the player's view can be found in Figure 1.

[1] Raspberry Pi 4: `https://www.raspberrypi.com/products/raspberry-pi-4-model-b/`

3.2. Game implementation

The game[2] runs on PsychoPy v2023.2.3 [15] and Python v3.9.18 [16]. The components of the game are a "Player" class and the game routines. The "Player" class stores information about each player during the game (name, score, round points, message). The main game routine updates player messages and variables when needed (before and after each interaction with another player). The main game routine draws one (two if it is the first turn in a round) player(s) that has not played yet to be the receiving player or the first player in a round (when everyone has played the pool of possible players is reset). The current player is then assigned a number, which will be delivered to their screen using the aforementioned memcached scheme. The experimenter only has to listen to the number spoken by the player and input it into a text-box from the control computer. This is not only saved in an experiment data table (alongside other turn information such as the current player, drawn number, turn time-stamps, etc.), but also passed as additional information to the next player. In case of a challenge or a penalty, the experimenter can input one of two keys, `c` or `p`, to indicate the relative situation. Penalties are automatically applied when the passed number is lower than the one received, simplifying the process for the researcher. Explicit indication of a penalty is required only when the produced utterance differs significantly from the carrier sentence. Time-stamps for each turn are automatically recorded from the beginning of the turn (coincides with the end of the previous turn) to the end of the turn, which is marked when the experimenter confirms the input by pressing the `return` key once the sentence has been uttered. These time-stamps are approximate but can be used to precisely align the boundaries of spoken sentences in the post-processing phase. The time-stamps are relative to a $200\,\mathrm{ms}$ $2\,\mathrm{kHz}$ alignment tone played on loudspeakers in each booth at the beginning of the game, to synchronise the audio recordings to the game time. An additional task left to the experimenter is to click on a visible button on the screen to initiate the "last turn" routine, in which the last player will receive a message giving only the option to trust or challenge the previous player. The decision of when to initiate the last turn routine is solely based on the experiment's elapsed time since the researcher lacks knowledge of the players' scores and the game's progression.

3.3. Post-processing

Apart from the first step, the post-processing processes to go from the raw ($\approx 1\,\mathrm{h}$) recordings to aligned single sentence-like units (SUs) (utterances for each player's turn) are automated thanks to our experiment design.
The phases for post-processing are as follows:

1. Cutting of all session recordings from the alignment tone to the experiment's end (recordings from the same game session should have the same duration). This is done manually for the first recording of a session, at which point the selection start and end times can be noted and used to cut the remaining recordings from the same session. We do this using Praat v6.3.16 [17], but any other audio editing tool could be used.
2. Rough alignment of SUs using the experiment data (turn time-stamps). Accomplished with a Python script using the experiment data from each session and the individual audio recordings from each player to create Praat TextGrid files with one interval tier. The boundaries are equivalent to the player's turn start and end and the labels are the carrier sentence with the spoken number for that turn (e.g. "My name is Lee, I have number 16, and I pass it on to Sam"). The approximate alignment and orthographic transcription are made only for turns that did not result in a penalty and were not a challenge statement.
3. Precise alignment using the Montreal Forced Aligner (MFA) v2.2.17 [18]. The MFA aligns each utterance in the audio recordings at the word and phone level, using the approximate alignments made in the previous step. The result is a new Praat TextGrid file, for each audio file, containing two to three tiers depending on if using the MFA option to keep the original text or not. The TextGrid tiers include a word level alignment, a phone level alignment, and optionally the original approximate sentence alignment from step 2.
4. Alignment of experiment data to forced-aligned utterances. Automated with a Python script similar to that of step 2. It realigns the turn sentences to the forced-alignments from step 3. Additionally, it also aligns for each turn two binary variables: "truth", describing the truthfulness of the utterance, and "trusted", describing the response from the succeeding player. Finally, the score at the end of the turn is also added to the TextGrid alignments.
5. Extraction of single utterances from each audio recording using the boundaries and meta-information contained in the relative TextGrid files, allowing for different filtering options.

The final result after the post-processing steps (up to step 3) is a collection of audio recordings for each player in each session (cut from the end of the pure tone to the end of the experiment) with their respective Praat TextGrid files containing 6 to 7 tiers (words, phones, [utterances], sentence (re-aligned), truth, trusted, score).

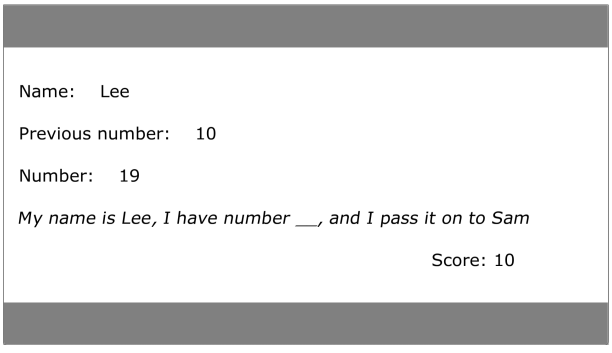

Figure 1: *Current player's view (screenshot of the monitor).*

4. Results

4.1. Audio recording quality

Figure 2 shows an example of an utterance from one of the pilot recording sessions after post-processing. The high signal-to-noise ratio of the recordings is evident given that most formants are clearly visible, thus allowing for accurate classical acoustic analyses and analyses employing MFCCs. The automatic alignments achieved using the MFA are precise both at the word and phone level, although the phonetic transcription is debatable. The phonetic transcription of the MFA is highly dependent on the phonetic dictionary used for the language. As a proof of concept, we used the `english_mfa` acoustic model

[2]Code publicly available at `https://gitlab.uzh.ch/indexical-dynamics/numberliegame`

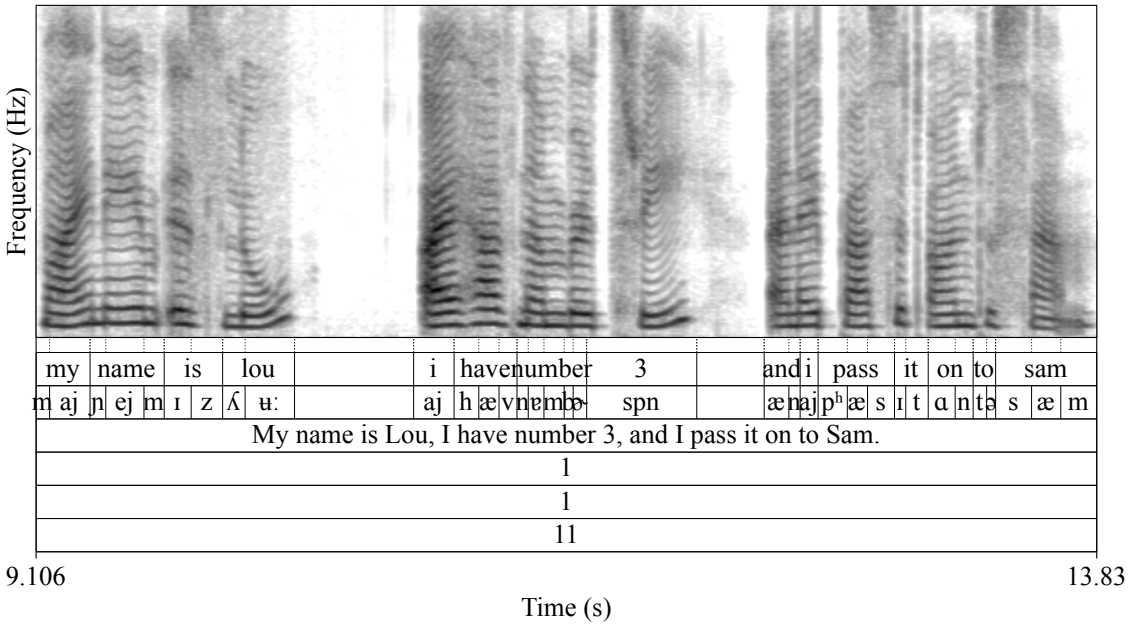

Figure 2: *Example utterance after post-processing steps. Tiers from top to bottom: words, phones, sentence, truth, trusted, score. Spectrogram settings: view range* = 0 Hz-6000 Hz*, window length* = 0.05 s*, dynamic range* = 90.0 dB*.*

in conjunction with the `english_us_mfa` dictionary. The participants in the pilot were all non-native speakers of English but with good proficiency[3].

4.2. Perceived difficulty

The game's level of difficulty from the participants' viewpoint was assessed through a brief survey administered to each pilot study subject. As indicated in Table 3, the game was not perceived as particularly challenging. Of the 11 respondents, 54.5% had experience playing a similar game to the *NL* game (deception games) before the experiment and only three found being deceptive ethically challenging. Based on conversations held with the subjects at the end of the experiment, the game duration (≈1 h) was not considered strenuous. Some subjects even expressed their wish to play longer, possibly implying forgetting the experimental setting and rather enjoying the game itself.

5. Limitations

The design of the *NL* game ensures that players have the freedom to lie in most cases. However, there are instances where lies are forces, such as when the drawn number is lower than the one received while simultaneously trusting the previous player. Although there may be distinctions between these types of lies, we argue that compared to similar experimental designs based on social games, our experiment uniquely allows for accurate labelling of both types of lies, enabling characterization of their differences. Another aspect of spontaneity that is challenging to evaluate is the requirement for players to use the carrier sentence to pass the number to the next player (see section 2.1). Precise production of the carrier sentence is crucial because we have observed that imprecise production can lead to inaccuracies in the MFA alignments. For instance, filled pauses noticeably impact the word alignment accuracy of the MFA in our testing. We have yet to determine whether this issue can be addressed with other techniques, but given the extensive research on the salience of pauses in deceptive speech related to the cognitive load account of deception [1, 6, 10], we believe it is important to explore different (ideally automated) solutions to capture such instances during the *NL* game.

Table 3: *Summary of survey on game difficulty. For each question, the respondent could select a value on a discrete 6-point Likert scale going from "very easy" (1) to "very difficult" (6). N=11. (Q25=*1^{st} *quartile, Q75=*3^{rd} *quartile).*

Difficulty on	Q25	median	Q75
Gameplay	2.0	2.0	3.5
Lying	2.5	3.0	4.5
Lie detection	3.0	4.0	5.0

6. Conclusion

With the *NL* game design we attempt to provide recordings of speech that is precisely labelled as either a lie or a truth in a setting with immediate and high-stakes consequences for deceivers and with implications in the perception of deceptive speech. We believe that due to its versatility and ease of manipulation, this design can be employed to research different facets of deception in many fields, including phonetics, psychology, economics, and social studies. Finally, we plan to release a public database of deceptive speech from the *NL* game in the coming year.

[3]The planned data collection to create a deceptive speech corpus that will be published will include only native speakers.

7. Acknowledgements

This project was funded by the Swiss National Science Foundation under the "The dynamics of indexical information in speech and its role in speech communication and speaker recognition" project #185399. We would like to thank the 11 participants of the pilot study who made this manuscript and the evaluation of our experiment's design possible.

8. References

[1] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychological Bulletin*, vol. 129, pp. 74–118, 2003.

[2] F. Enos, "Detecting deception in speech," Ph.D. dissertation, Columbia University, 2009.

[3] J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, A. Stolcke, and E. Shriberg, "Distinguishing deceptive from non-deceptive speech," 2005. [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/D8697C06https://doi.org/10.7916/D8697C06

[4] M. Zuckerman, B. M. Depaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," *Advances in Experimental Social Psychology*, vol. 14, pp. 1–59, 1 1981.

[5] K. E. Sip, M. Lynge, M. Wallentin, W. B. McGregor, C. D. Frith, and A. Roepstorff, "The production and detection of deception in an interactive game," *Neuropsychologia*, vol. 48, pp. 3619–3626, 10 2010.

[6] M. McDonald, E. Mormer, and M. Kaushanskaya, "Speech cues to deception in bilinguals," *Applied Psycholinguistics*, vol. 41, pp. 993–1015, 9 2020. [Online]. Available: https://www.cambridge.org/core/journals/applied-psycholinguistics/article/speech-cues-to-deception-in-bilinguals/17067316D8C786D9427DECF1248D3F19

[7] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," *WMDD 2015 - Proceedings of the ACM Workshop on Multimodal Deception Detection, co-located with ICMI 2015*, pp. 1–8, 11 2015.

[8] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity native language," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, pp. 2001–2005, 2016. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-129

[9] C. H. Huang, H. C. Chou, Y. T. Wu, C. C. Lee, and Y. W. Liu, "Acoustic indicators of deception in mandarin daily conversations recorded from an interactive game," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, pp. 1731–1735, 2019.

[10] Z. Zhang, C. McGettigan, and M. Belyk, "Speech timing cues reveal deceptive speech in social deduction board games," *PLOS ONE*, vol. 17, p. e0263852, 2 2022. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0263852

[11] C. Yang, X. You, X. Xie, Y. Duan, B. Wang, Y. Zhou, H. Feng, W. Wang, L. Fan, G. Huang, and X. Shen, "Development of a chinese werewolf deception database," *Frontiers in Psychology*, vol. 13, p. 1047427, 1 2023.

[12] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, "The dyvis database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research," *International Journal of Speech, Language and the Law*, vol. 16, pp. 31–57, 9 2009. [Online]. Available: https://journal.equinoxpub.com/IJSLL/article/view/10005

[13] S. C. Jat, K. McDougall, and A. Paver, "Pausing and the 'othello error': Patterns of pausing in truthful and deceptive speech in the dyvis database," *International Journal of Speech, Language and the Law*, vol. 30, pp. 87–118, 8 2023. [Online]. Available: https://journal.equinoxpub.com/IJSLL/article/view/24331

[14] B. Fitzpatrick, "Distributed caching with memcached," *Linux Journal*, vol. 2004, no. 124, p. 5, aug 2004.

[15] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, pp. 195–203, 2 2019. [Online]. Available: https://link.springer.com/article/10.3758/s13428-018-01193-y

[16] G. V. Rossum and F. L. Drake, *Python 3 Reference Manual*. CreateSpace, 2009.

[17] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, 01 2001.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.