The ventriloquist paradigm: Studying speech processing in conversation with experimental control over phonetic input

E. Felker, A. Troncoso-Ruiz, M. Ernestus, and M. Broersma

Citation: The Journal of the Acoustical Society of America **144**, EL304 (2018); doi: 10.1121/1.5063809
View online: https://doi.org/10.1121/1.5063809
View Table of Contents: http://asa.scitation.org/toc/jas/144/4
Published by the Acoustical Society of America

The ventriloquist paradigm: Studying speech processing in conversation with experimental control over phonetic input

E. Felker,[a] **A. Troncoso-Ruiz, M. Ernestus, and M. Broersma**
Centre for Language Studies, Radboud University, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
e.felker@let.ru.nl, m.troncosoruiz@let.ru.nl, m.ernestus@let.ru.nl, m.broersma@let.ru.nl

Abstract: This article presents the ventriloquist paradigm, an innovative method for studying speech processing in dialogue whereby participants interact face-to-face with a confederate who, unbeknownst to them, communicates by playing pre-recorded speech. Results show that the paradigm convinces more participants that the speech is live than a setup without the face-to-face element, and it elicits more interactive conversation than a setup in which participants believe their partner is a computer. By reconciling the ecological validity of a conversational context with full experimental control over phonetic exposure, the paradigm offers a wealth of new possibilities for studying speech processing in interaction.

[DDO]
Date Received: August 9, 2018 **Date Accepted:** September 28, 2018

1. Introduction

This paper presents a novel experimental paradigm that, for the first time, enables the study of speech processing in interaction while maintaining full experimental control over phonetic exposure. Speech perception and production are doubtless shaped by experiences in conversation, as demonstrated by research on perceptual adaptation (e.g., Norris *et al.*, 2003) and phonetic alignment and accommodation (e.g., Pardo, 2006). To reach a fuller understanding of the mechanisms underlying language processing in interactive contexts, researchers have called for studying language perception and production in more contextualized, ecologically valid settings, such as informal face-to-face communication centered on joint tasks (e.g., Tanenhaus and Brown-Schmidt, 2008; Tucker and Ernestus, 2016; Willems, 2017). For experiments investigating the underlying mechanisms of perceptual learning and phonetic alignment, in which the quantity, context, and timing of exposure to critical speech sounds are theorized to play a key role, control of phonetic detail is crucial. However, controlling phonetic input in a natural conversation poses a methodological challenge.

All approaches to studying sound learning and adaptation make trade-offs between ecological validity and experimental control. Traditional phonetics experiments that control the type and presentation of stimuli (e.g., categorization, discrimination, shadowing, lexical decision, and judgment) have led to fundamental insights into how speech processing works in individuals when tested in isolation but do not address naturalistic interaction. Other research methods provide more ecological validity (e.g., Map task, Brown *et al.*, 1983; Diapix task, Van Engen *et al.*, 2010; spontaneous dialogue, Torreira and Ernestus, 2010; Pardo *et al.*, 2012) but do not control the phonetic exposure participants receive.

To study *syntactic* alignment, the "confederate-scripting" paradigm (Branigan *et al.*, 2000) combines natural interaction with experimental control of language input by fully scripting the linguistic input at the syntactic and lexical level. To investigate *sound* learning mechanisms, however, the relevant level to control is phonetics. Whereas phonetic studies often involve artificial accents, manipulated speech sounds, or avoidance of specific sounds, even a phonetically trained confederate cannot perfectly control all the phonetic details of their speech during a live experiment. Furthermore, since subtle phonetic alignment often occurs in dialogue (e.g., Pardo, 2006), the confederate's accent risks converging toward that of participants, such that not all of them receive comparable phonetic input. In fact, variability in the speech input can only be avoided if the speech is pre-recorded.

[a]Author to whom correspondence should be addressed.

We introduce the new ventriloquist paradigm, which solves the problem of variable phonetic input in live speech by employing pre-recorded speech covertly in a real-time conversation. In this paradigm, a participant and confederate work together face-to-face on a cooperative computer-based task. While the participant believes they are having a normal conversation, the confederate does not actually speak but plays pre-recorded utterances to the participant's headphones while briefly hiding her face behind a screen. As in a ventriloquist performance, the true source of the confederate's speech is thus disguised. The pre-recorded speech meets the experiment's phonetic requirements and includes all phrases necessary for the joint task and various other phrases to respond to whatever the participant says.

This paper presents the methodology of the ventriloquist paradigm and the steps required to incorporate pre-recorded speech in an experiment while convincing participants they are having a live conversation. To illustrate how the paradigm can be used to study sound learning in speech perception and production, we describe its implementation in two dialogue elicitation tasks and an auditory lexical decision test. We also evaluate the ventriloquist paradigm's effectiveness and compare it to two control setups that vary in how present or personal the confederate is: In one version, we removed the face-to-face aspect of the interaction by putting the participant and confederate in separate testing booths. In another, we further reduced the "human" nature of the interaction by not only having participants alone in a booth but also telling them they were interacting with a computer, thus implementing a "Wizard of Oz" experiment (Fraser and Gilbert, 1991; Riek, 2012). By analyzing the conversational interaction produced with the ventriloquist paradigm and these control methods, we assess how effective the ventriloquist paradigm is at creating a convincing, interactive dialogue.

2. Ventriloquist paradigm methodology

2.1 General procedure

At the beginning of a session, the participant is told that he will play a cooperative computer game with a partner. The experiment leader explains that both players will speak into microphones and that their speech will be transmitted to each other's noise-canceling headphones, which they must keep on throughout the session. To prevent the participant from engaging with the confederate before she can play her pre-recorded speech, the experiment leader holds the conversational floor so that the players cannot speak to each other until their headphones are on.

During the cooperative game, the participant and confederate sit at a table across from each other, each facing their own computer monitor, but with ample room between the monitors for them to see each other. Every time the confederate needs to speak, she leans toward a dummy microphone next to the table, thereby hiding her entire face behind her monitor, and surreptitiously presses a key on a hidden numeric keypad corresponding to a desired speech function. The computer then plays a pre-recorded utterance, which the participant hears in his headphones.

2.2 Software and speech materials

The experiment software implements a structured, collaborative two-player game that requires the players to communicate orally to share information or give each other instructions. Each key of the numeric keypad is mapped to a different audio category so that when it is pressed, an audio file from the associated speech category is played. A visual reference of the number key-audio category mappings is overlaid on the confederate's screen as a memory aid. The audio files consist of various categories of pre-recorded utterances that are scripted to meet the researcher's desired phonetic constraints. The utterances can be one of two types: trial-linked or flexible.

Trial-linked utterances can only be played on specific trials or time points within the experiment. For instance, a recording of the speaker introducing herself may be linked to the welcome screen and a recording of her saying goodbye to the end screen. Most trial-linked utterances relate to visual stimuli that occur on specific trials, such as descriptions of a displayed picture or instructions for the participant to click on a displayed word. In case participants ask the confederate to repeat herself, trial-linked utterances have follow-up versions that can be played in succession if necessary. For example, if the first utterance for a trial is "Now we want the word *flower*," a follow-up version could be "I said *flower*," and a second follow-up could be "*Flower*" with even more emphasis. The phrases vary in structure and wording to avoid repetitiveness and contain some disfluencies to make them sound more natural, but they are nevertheless kept short to reduce the chance of the participant interrupting them. To

facilitate the confederate's task of playing the audio files, the software links all trial-linked utterances to a single numeric key, and pressing that key will play only the utterances linked to the current trial, in the pre-specified order.

Other pre-recorded utterances are flexible, meaning they are playable throughout the experiment to respond to whatever the participant might ask. Important flexible utterance categories include affirmative responses, negative responses, backchannels such as "mm-hm," variations of "I don't know" (also useful for responding to off-topic remarks or open-ended questions), requests to elaborate, reassuring remarks, thank-you's, utterances of surprisal about the appearance of new trials (if the confederate cuts a trial short to unblock the conversation), and reminders of the task rules. Each category contains numerous recordings that serve the same communicative function, and there are enough utterances to ensure that no audio file is repeated within a session.

2.3 Physical setup and equipment

The ventriloquist paradigm is set up in a large booth or testing room, ideally with a window through which the experiment leader can monitor the activity. A single computer runs the experiment software and displays graphics on two wide monitors situated side by side, facing opposite directions across the table. A numeric keypad with silent keys is just below the table (e.g., resting on a cabinet), hidden from the participant's view. At the center of the table rests an active microphone aimed toward the participant and connected to an audio mixing console. The confederate's dummy microphone stands at the outside edge of the confederate's side of the table.

Audio output from the computer is split into two channels: one to the participant's noise-canceling over-ear headphones, and one to the audio mixing console. The console combines audio input from the computer and participant's active microphone and sends it to the confederate's headphones, an audio recorder, and a pair of headphones outside the testing booth for the experiment leader.

3. Examples of ventriloquist paradigm implementation

To illustrate how the ventriloquist paradigm can be used to answer specific research questions about speech perception or production in interaction, this section presents two dialogue elicitation tasks and an auditory lexical decision task we have implemented with it.

3.1 Dialogue elicitation task: Code Breaker game

The Code Breaker game is designed for research into various types of phonetic learning, such as perceptually adapting to an unfamiliar accent's vowel shift or learning to more clearly produce a difficult non-native sound contrast. While critical speech sounds in the ventriloquist's speech repertoire are controlled to provide the desired type and amount of phonetic input for participants to learn from, various task- and interaction-related variables, such as the presence and type of feedback from the confederate, can also be manipulated to test specific hypotheses about learning mechanisms.

In the Code Breaker game, the participant and confederate work together to solve puzzles and tell each other to click on words belonging to phonological minimal pairs with or without feedback. In each trial [Fig. 1(a)], Player A sees a sequence of colored shapes followed by a question mark, above an array of four words, and he must tell his partner what shape comes next. Player B finds the specified shape on her screen and tells her partner to click on the target word linked to that shape. When the ventriloquist is Player A, trial-linked utterances refer to a puzzle's solution (e.g., "I think we need a black square"); when she is Player B, the trial-linked utterances contain the target words (e.g., "So you should click on *land*"). For the study of speech perception, the participant acts as Player A, as their challenge is to accurately perceive the target words. For production, the participant acts as Player B, as their challenge is to pronounce the target words accurately.

3.2 Dialogue elicitation task: Picture description

Another interactive game involving more elaborate and contextualized speech is the picture description task [Fig. 1(b)], which can be used in combination with Code Breaker trials to give the participant different types of phonetic exposure (e.g., hearing words in various semantic contexts, with or without their phonological neighbors, with or without spelling cues, etc.). In each picture description trial, Player A sees a picture while Player B sees an array of four words consisting of two phonological minimal pairs. Player A describes the picture until Player B is able to select the word matching

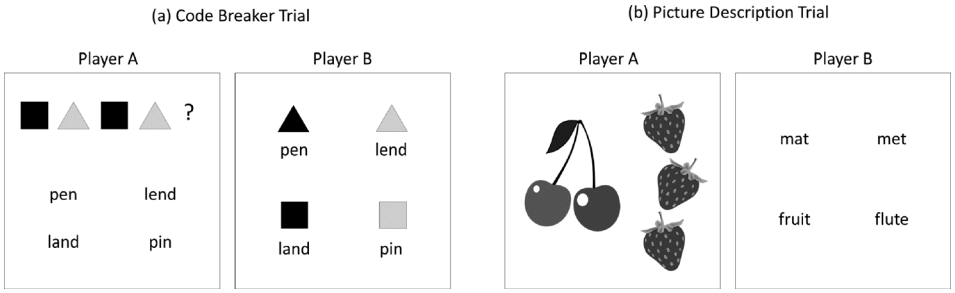

Fig. 1. Sample screens for two players in one trial of the Code Breaker game (Sec. 3.1) and for one trial of the picture description game (Sec. 3.2).

the described picture. Optionally, Player B is also instructed to read aloud their four word options before making a final choice. If the ventriloquist is Player A, the trial-linked utterances are the picture descriptions; if she is Player B, they are the speaker declaring her answer (e.g., "I have *mat*, *met*, *fruit*, and *flute*, so I'm going to choose *fruit*").

3.3 Auditory lexical decision test

An auditory lexical decision task can be employed to measure the participant's perceptual adaptation to the pre-recorded speaker after a dialogue elicitation task. This method is identical to a regular lexical decision test except the participant believes they are responding to words being read aloud in real-time by their conversation partner. The participant is instructed not to request repeats or clarification to ensure that he does not try to interact during this test, and the confederate remains hidden behind her monitor the entire time to avoid visual distraction. The trial-linked audio consists of the auditory lexical decision stimuli. Rather than being triggered by the confederate's button presses, it is played automatically at pre-determined inter-stimulus intervals, randomized within a small range to give the impression that the items are being read in real time.

4. Validity of the ventriloquist paradigm

The validity of the ventriloquist paradigm depends on how reliably it convinces participants they are engaged in a genuine conversation. We analyzed the participant-ventriloquist interaction using data from 101 Dutch participants (aged 18–30 yr) speaking their highly proficient L2 English in sessions of 15 to 30 min in one of two experiments, each with a different confederate and different pre-recorded native English speaker. One experiment (56 participants) used the Code Breaker (production) and picture description task, and the other (45 participants) used the Code Breaker (perception) and lexical decision task.

All participants engaged with the cooperative tasks, and nobody overtly questioned the genuineness of the conversations during the session. Questionnaires administered at the end of each session, without the confederate present, showed that 79.2% of participants reported no suspicion that their partner's speech was pre-recorded. The most common reasons given for suspecting pre-recorded speech were that the timing of the ventriloquist's speech or body movements felt slightly off, phrase structures were repeated, or the speech sounded "too perfect."

Interestingly, we found two differences between those who did and did not report suspicions. For the former group, interactivity (as measured by the total number of ventriloquist utterances played during the entire Code Breaker game) was lower than for the latter [$M = 88.5$ utterances vs $M = 96.6$ utterances, $t(42.075) = 2.20$, $p = 0.03$]. This suggests either that hearing more ventriloquist speech increased believability, or, alternatively, that participants sought less interaction when they suspected their partner's speech was not live. Furthermore, self-reported English proficiency (speaking, listening, reading, and writing) was higher for those who did report suspicions than for those who did not [$t(33.56) = 2.28$, $p = 0.03$], suggesting either that greater task difficulty increased participants' susceptibility to the illusion or that discovering the truth increased self-ratings. Between the two experiments, the proportion of participants who bought into the illusion did not differ; $X^2(1, N = 101) = 0.50$, $p = 0.48$.

5. Evaluating the importance of face-to-face context

To determine whether the face-to-face setting of the ventriloquist paradigm affected the extent to which participants believed in the genuineness of the conversation, we collected data from 22 new participants from the same population using an alternative setup in which the participant and confederate did the same tasks together but in separate testing booths from which they could not see each other.

In these experiments, importantly, the tasks, confederates, audio setup, software, and pre-recorded speech materials were the same as in Sec. 4, except no dummy microphone was needed since the participant never saw the inside of the confederate's booth. For the interactive games, the confederate used her keyboard to play pre-recorded speech exactly as in the ventriloquist paradigm, aiming to be just as interactive as in the ventriloquist setup to enable a fair comparison. While the confederate never entered the participants' testing booth, participants could see her walking by the window of their booth and heard the experiment leader speaking to her as if she was another participant. Furthermore, whenever the experiment leader gave instructions to the participant, she then stopped inside the confederate's booth to create the impression that she instructed her as well.

In the post-experiment questionnaire, only 32% of the participants reported no suspicion that the speech was pre-recorded, a significantly lower proportion than in the ventriloquist paradigm [$X^2(1,\ N=123)=17.375,\ p<0.001$]; moreover, all the participants who noticed the pre-recorded speech also believed their partner was actually a computer or robot. These results suggest that the face-to-face aspect of the ventriloquist paradigm strongly contributed to making the pre-recorded speech sound live. The separate-booth setup, on the other hand, does not seem viable for studying natural conversation, as it convinced few participants that they were having a live conversation or were even talking to another human.

6. Evaluating the importance of beliefs about interlocutor's humanness

To examine whether the ventriloquist paradigm creates more engaging and interactive conversation than when people believe they are talking to a computer, 36 additional participants from the same population were tested in a new setup in which they were told upfront that they were interacting with a computer.

The setup and procedure were as described in Sec. 5, except that the experiment leader told participants that their partner was a smart computer player and no attempt was made to hide the fact that the speech was pre-recorded. Participants completed the tasks from the second experiment (Code Breaker perception and lexical decision task) described in Sec. 4, and the pre-recorded speech was played by the same person as before, who again aimed to be just as interactive as in the ventriloquist setup.

Nobody reported any suspicion that they had been playing with a real person rather than with a smart computer. We compared the interactions during the computer player version of the Code Breaker game to those from the matching ventriloquist paradigm sessions. The total number of pre-recorded utterances played per session was similar in the ventriloquist setup [$M=112.1$, standard deviation (SD) $=15.8$] and the computer-player setup ($M=120.1$, SD $=21.1$), [$t(63.3)=1.904,\ p=0.06$], confirming that the computer player and ventriloquist were played in a comparable way. To assess the interactivity of the conversation, we measured how often and for how long participants spoke, excluding two participants due to recording malfunctions. The *number* of participant utterances (utterances being defined as any stretches of speech bounded by either a pause of at least 0.6 s or an intervening pre-recorded utterance) was higher in the ventriloquist setup ($M=178.8$, SD $=50.8$) than in the computer-player setup ($M=136.6$, SD $=41.6$); $t(76.46)=4.06$, $p<0.001$. For participants' speech duration, we analyzed the ratio of participant-to-confederate speaking time, rather than participant speaking time alone, to control for the influence of any between-session variability in confederate speech duration. This ratio was significantly higher in the ventriloquist setup (2.05:1) than in the computer-player setup (1.76:1); $t(58.94)=2.05$, $p=0.04$. These results demonstrate that the ventriloquist paradigm increases participants' engagement in the conversation, as measured by their speaking behavior, relative to the computer-player control setup.

7. General discussion

This report described the ventriloquist paradigm, a novel experimental method that incorporates pre-recorded speech in real-time, face-to-face conversation. The results showed that the ventriloquist paradigm convinces most participants that they are having a genuine dialogue. The face-to-face aspect of the interaction appears to be instrumental in maintaining the illusion, as participants were much less likely to notice that the speech was pre-recorded in the ventriloquist paradigm than in a control setup

utilizing separate testing booths. Participants may assume, possibly based on prior experience with experiments, that the speech they hear from headphones in a testing booth is pre-recorded unless they have strong evidence to the contrary, such as the confederate's physical co-presence. Furthermore, analyses showed that the ventriloquist paradigm elicited more interactive, engaging conversation than a setup in which participants believed they were interacting with a computer.

Practical challenges associated with the ventriloquist paradigm are that scripting and recording the ventriloquist's utterances is time-consuming, and the paradigm requires a confederate with some degree of acting ability who can think on her feet. Moreover, researchers might have to discard some data from participants who did not buy into the ventriloquist illusion. Furthermore, compared to ordinary conversation, the spontaneity and complexity of interaction with the ventriloquist will always be somewhat limited, given that the pre-recorded speech is only designed to handle conversation around highly-structured, predictable tasks. However, we believe that the paradigm can be adapted to incorporate more complex dialogue tasks than we have used so far, such as the Map Task (Brown *et al.*, 1983), although extensive pilot testing would be needed to determine what trial-linked and flexible utterances would be necessary to make the interaction convincing. Finally, it should be noted that using pre-recorded speech precludes any level of linguistic alignment from the confederate to the participant, and this lack of reciprocal alignment, while enabling full control over the phonetic characteristics of the input, necessarily makes the interaction less natural than if the confederate were speaking spontaneously.

In short, the ventriloquist paradigm can be used to study how people learn from and adapt to each other's speech in everyday communication centered on cooperative tasks, which affords more ecological validity than many traditional experimental paradigms. As the paradigm can be used with a variety of different cooperative tasks, numerous task- and interaction-related variables can be manipulated to study various aspects of speech perception and production. Most importantly, the ventriloquist paradigm allows researchers to fully control the phonetic input participants receive in the conversation, thereby facilitating research into the underlying mechanisms of sound learning. By combining this fine-grained control of the input with a naturalistic dialogue, the ventriloquist paradigm opens up a wealth of new possibilities for studying speech processing in interaction.

Acknowledgments

This research was supported by a Vidi grant from the Netherlands Organisation for Scientific Research (NWO), awarded to M.B. The authors would like to thank Bob Rosbag for his invaluable help with the technical development of the paradigm.

References and links

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (**2000**). "Syntactic co-ordination in dialogue," Cognition **75**, B13–B25.

Brown, G., Anderson, A., Yule, G., and Shillcock, R. (**1983**). *Teaching Talk* (Cambridge University Press, Cambridge, UK).

Fraser, N. M., and Gilbert, G. N. (**1991**). "Simulating speech systems," Comput. Speech Lang. **5**, 81–99.

Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cognitive Psychol. **47**(2), 204–238.

Pardo, J. S. (**2006**). "On phonetic convergence during conversational interaction," J. Acoust. Soc. Am. **119**(4), 2382–2393.

Pardo, J. S., Gibbons, R., Suppes, A., and Krauss, R. M. (**2012**). "Phonetic convergence in college roommates," J. Phonetics **40**(1), 190–197.

Riek, L. D. (**2012**). "Wizard of Oz studies in HRI: A systematic review and new reporting guidelines," J. Human-Robot Interaction **1**(1), 119–136.

Tanenhaus, M. K., and Brown-Schmidt, S. (**2008**). "Language processing in the natural world," Philos. Trans. R. Soc. B **363**, 1105–1122.

Torreira, F., and Ernestus, M. (**2010**). "The Nijmegen corpus of casual Spanish," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, edited by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, European Language Resources Association (ELRA), Paris, pp. 2981–2985.

Tucker, B. V., and Ernestus, M. (**2016**). "Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon," Mental Lexicon **11**(3), 375–400.

Willems, R. (editor) (**2017**). *Cognitive Neuroscience of Natural Language Use* (Cambridge University Press, Cambridge, UK).

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (**2010**). "The Wildcat Corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles," Lang. Speech **53**(4), 510–540.