Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-022-02218-6

BRIEF REPORT

Automatic imitation of human and computer-generated vocal stimuli

Hannah Wilt[1] · **Yuchunzi Wu**[2,3] · **Antony Trotter**[4] · **Patti Adank**[1]

Accepted: 10 November 2022

Abstract
Observing someone perform an action automatically activates neural substrates associated with executing that action. This covert response, or *automatic imitation*, is measured behaviourally using the stimulus–response compatibility (SRC) task. In an SRC task, participants are presented with compatible and incompatible response–distractor pairings (e.g., an instruction to say "ba" paired with an audio recording of "da" as an example of an incompatible trial). Automatic imitation is measured as the difference in response times (RT) or accuracy between incompatible and compatible trials. Larger automatic imitation effects have been interpreted as a larger covert imitation response. Past results suggest that an action's biological status affects automatic imitation: Human-produced manual actions show enhanced automatic imitation effects compared with computer-generated actions. Per the integrated theory for language comprehension and production, action observation triggers a simulation process to recognize and interpret observed speech actions involving covert imitation. Human-generated actions are predicted to result in increased automatic imitation because the simulation process is predicted to engage more for actions produced by a speaker who is more similar to the listener. We conducted an online SRC task that presented participants with human and computer-generated speech stimuli to test this prediction. Participants responded faster to compatible than incompatible trials, showing an overall automatic imitation effect. Yet the human-generated and computer-generated vocal stimuli evoked similar automatic imitation effects. These results suggest that computer-generated speech stimuli evoke the same covert imitative response as human stimuli, thus rejecting predictions from the integrated theory of language comprehension and production.

Keywords Imitation · Speech perception · Speech production · Vocal

Introduction

Observing others' manual or vocal actions activates neural mechanisms required to perform that action (Buccino et al., 2004; Fadiga et al., 1998). For vocal actions, this type of covert—or automatic—imitation occurs whenever we hear and/or see someone speak and involves activation of speech production mechanisms and associated neural substrates (Fadiga et al., 2002; Nuttall et al., 2016; Watkins et al., 2003). Covert imitation is generally referred to as *automatic imitation* and is measured behaviourally using the stimulus–response compatibility (SRC) paradigm for manual and for vocal actions (Brass et al., 2000; Cracco et al., 2018; Heyes, 2011; Kerzel & Bekkering, 2000, Jarick & Jones, 2009). In a manual SRC task (Brass et al., 2000), participants are instructed to perform a manual action following a prompt (e.g., lift index finger for "1" and middle finger for "2"). The prompt is presented superimposed on a distractor image or video of a hand lifting the index or middle finger. When the prompt is presented in the presence of a compatible distractor ("1" with a video of a lifting index finger), participants are faster to perform the correct response than when the prompt is presented together with an incompatible distractor ("1" with a lifting middle finger video). For compatible distractors, action observation is thought to facilitate motor or production mechanisms required for performing the prompted action, thus reducing response times (RTs) and

✉ Patti Adank
p.adank@ucl.ac.uk

1 Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1N 1PF, UK

2 Department of Neural and Cognitive Sciences, New York University Shanghai, Shanghai, China

3 NYU-ECNU Institute of Brain and Cognitive Sciences at New York University Shanghai, Shanghai, China

4 Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

Published online: 28 November 2022

errors. In contrast, incompatible distractors result in competition between the facilitated motor mechanisms and those required to produce the prompted response, thus increasing RTs and errors. A larger automatic imitation effect (i.e., a larger RT difference between incompatible and compatible trials) indicates that production mechanisms were more engaged for the distractor, thus showing a measure of covert imitation (Heyes, 2011). In vocal SRC paradigms, participants produce a speech response following a prompt (e.g., say "ba" when seeing "£") while ignoring a distractor (e.g., an audio recording and/or a video of someone saying "ba"). As for manual SRC tasks, RTs are slower for incompatible ("da") than compatible ("ba") distractors (Galantucci et al., 2009; Kerzel & Bekkering, 2000; Roon & Gafos, 2015).

The integrated theory of language production and comprehension proposes that covert imitation serves a specific purpose in action processing (Pickering & Garrod, 2013). When planning to produce a speech action (e.g., an instruction to say "ba" or "da"), a set of motor (articulatory) commands is formulated plus two control signals. A perceptual control signal processes the proprioceptive (and/or visual/auditory) experience in the speech production system of the action being executed. This perceptual signal is used as sensory, or reafferent, action feedback. The second control signal is an efference copy of the action ultimately resulting in a predicted percept of producing the planned utterance. The perceptual feedback signal and the predicted percept are compared in real time during action execution. Whenever a discrepancy is detected, an error signal is sent to the action planning mechanism to update motor plans.

When perceiving speech, a similar mechanism is presumed to operate, except, here, forward models generate predictions regarding upcoming speech utterances. Pickering and Garrod (2013) propose that this predictive process operates via the simulation or the association route. The simulation route is based on listeners' experiences of producing speech utterances. Here, listeners use a process of covert imitation to generate a forward perceptual model. This covert imitation crucially involves the engagement of forward models predicting incoming speech and the engagement of production mechanisms. The association route is based on listeners' experience of perceiving others' utterances. This route also involves engagement of prediction via forward models, except the links to sensory processing systems are activated rather than production systems. The simulation route is preferred when the speaker is similar/familiar (e.g., culturally, speech production style, anatomy) to the listener. When the speaker is less similar (e.g., when they are a non-native speaker), the association route is used.

The integrated theory predicts that observing speech utterances produced by a similar or familiar a speaker involves covert imitation, as it will preferentially engage the simulation route (Adank et al., 2010). Listening to a speaker who is dissimilar or unfamiliar will engage less covert imitation, as it will preferentially engage the association route. The integrated theory predicts that speech utterances for dissimilar speakers will vary in evoked covert imitation. Listening to similar speakers will evoke larger automatic imitation effects than listening to less similar speakers. Thus far, no speech SRC studies have been conducted to test this prediction, but various SRC studies using manual actions tested a related question—namely, whether observation of computer-generated manual actions results in less automatic imitation compared with human-produced actions (Cracco et al., 2018; Gowen & Poliakoff, 2012; Longo et al., 2008; Press et al., 2005; Press et al., 2006; Stürmer et al., 2000).

Press et al. (2005) presented participants with an opening or closing hand produced by a human or a robot. They found more automatic imitation for human stimuli (33-ms automatic imitation effect) versus robotic stimuli (6 ms). When testing was repeated on the second day, automatic imitation decreased for human stimuli (24 ms) and increased for robotic stimuli (11 ms). Press et al. concluded that decreased automatic imitation for robotic stimuli shows that human movement stimuli are more effective visuomotor primes than robotic stimuli. Press et al. (2006) manipulated stimuli depicting a human hand by adding a metal wire wrist and informed participants that these stimuli were produced by a robot. Participants in Experiment 1 produced a prespecified response (hand opening/closing) for compatible and incompatible distractors while RTs were measured as in Press et al. (2005). They report no differences between SRC effects for human (16 ms) and robotic (26 ms) stimuli. Experiment 2 aimed to disentangle beliefs about the stimuli from stimulus animacy. Participants in a between-group design saw a genuine human or robotic hand (blue animated silhouettes) producing the actions. Participants presented with the genuinely human stimulus were told that the hand was either human or robotic in the two sessions of testing, and participants who were presented the genuinely robotic stimulus were told that the movement was generated by human or robotic movement. The participants who saw genuine human stimuli displayed similar automatic imitation for stimuli they were told were human (15 ms) and for stimuli they were told were robotic (14 ms). Participants who saw genuine robotic stimuli showed similar automatic imitation for stimuli they thought were human (5 ms) and for stimuli they believed to be robotic (5 ms). Experiment 2 demonstrated that stimulus properties modulate automatic imitation, with less automatic imitation for genuinely robotic stimuli.

Longo et al. (2008) presented participants with two sets of computer-generated, human-simulating manual stimuli: biomechanically possible and impossible hand actions. The possible actions consisted of a hand lowering either the index or middle finger. The impossible actions consisted of a hand with the index of middle finger bending in an

impossible angle. In Experiment 1, participants completed an SRC task that included both types of stimuli, without any instructions regarding the difference between possible and impossible actions. They found similar automatic imitation for possible (7 ms) and impossible (8 ms) actions. In Experiment 2, participants were informed that some finger movements were possible, and some were impossible. They found that automatic imitation for impossible stimuli nearly disappeared (1 ms), while automatic imitation persisted for possible movements (10 ms). Longo et al. show an effect of beliefs on automatic imitation of impossible actions not reported by Press et al. (2006).

Nevertheless, while it has been established that automatic imitation is generally decreased for nonhuman- than for human-generated manual actions, it is not evident that vocal actions show the same pattern in automatic imitation of actions with a different biological origin as manual actions. The current study aimed to address this gap and test if computer-generated vocal stimuli evoke decreased automatic imitation compared with human-produced vocal stimuli. Computer-generated, or synthetic, utterances are ideal for testing the prediction of the integrated theory regarding the use of covert imitation during perception. Computer-generated speech, especially when generated by a less sophisticated speech synthesizer, represents a speaker who will be dissimilar to the listener. Synthetic speech is generated by a speech synthesis programme running on a computer, which is physically and conceptually dissimilar from a human speaker. Consequently, the integrated theory predicts more covert imitation when listening to a human speaker for computer-generated speech. We conducted an online SRC experiment in which participants were presented with syllables ("ba" and "da") produced by a male human speaker or generated by a computer programme. Based on the integrated theory and results from Press et al. (2005; Press et al., 2006), we predicted that automatic imitation would vary depending on the speaker's biological status and be decreased for computer-generated speech.

Method

Participants

The study was conducted online, participants were recruited on Prolific (prolific.co) and the experiment hosted on Gorilla (gorilla.sc). We used the Bayes stopping rule to determine our sample size. We set our minimum sample size to 32 participants; the number required to fully counterbalance the design. After this minimum sample was collected, we calculated the BF10 based on model fit (Jarosz & Wiley, 2014) for a model including the two-way interaction between compatibility and speech type versus one including only the main effects. BF10 > 3 were considered evidence in favour of the alternative hypothesis, and we considered BF10 < 0.2 as evidence in favour of the null hypothesis (Raftery, 1995). Participants had completed a minimum of five studies on Prolific with an approval rate of ≥95% and were required to run the study on a computer, using Chrome, with wired headphones. A total of 191 participants were recruited for the eligibility screening, to obtain the final sample of 32 participants (16 female, average age = 23.7 years, $SD = 3.6$ years, range: 18–31 years). Exclusions are listed in detail in the Supplementary Materials. Participants received £0.75 upon completion of the eligibility study, and those who passed the eligibility study received a further £4 upon completion of the main task, a rate commensurate to £7.50 per hour. The experiment was approved by the Research Ethics Committee of University College London (UCL) and conducted under Project ID #15365.001.

Materials

The human SRC stimuli were audio recordings of a 40-year-old male British-English speaker from the North-West of England saying /ba/ and /da/ in a neutral tone of voice. The recording were made using a RØDE NO1-A Condenser Microphone and a Focusrite Scarlett 2i4 USB Computer Audio Interface preamplifier plugged into the sound card input of a Dell PC in a sound-proofed room at 44.1 kHz. Audio recordings were amplitude-normalized, down-sampled to 22.050 kHz, saved as mono, and scaled to 70 dB SPL (sound pressure level) using Praat (Boersma & Weenink, 2018).

The computer-generated vocal stimuli were created using a Klatt synthesizer (Klatt, 1980) using in-house software. We decided against the use of a modern, sophisticated speech synthesizer, which can produce speech near-indistinguishable from human-produced speech utterances (Wagner et al., 2019). In contrast, the Klatt synthesizer and other synthesizers developed in the 1980s (e.g., MITtalk, DECTalk; successors of the Klatt synthesizer) produce speech that is clearly distinguishable from human-produced speech (Pisoni et al., 1985; Ralston et al., 1991).

Using the Klatt synthesizer, we constructed a continuum between /ba/ and /da/ in 10 steps. All computer-generated syllables were converted to .wav files and matched to the human speech sounds as much as possible (Table 1) in terms of their sampling frequency, intensity, average fundamental frequency (f_o), and lowest three formant frequencies. We matched the acoustics characteristics of the human and computer-generated stimuli closely to avoid possible confounds introduced by a difference in intelligibility between the two stimulus types, or by paralinguistic issues such as perceived accent. We ran a pilot study ($N = 10$, not reported) to establish which of the computer-generated syllables were

Table 1 Stimulus characteristics of human and computer-generated stimuli

	Human /ba/	Human /da/	Computer-generated /ba/	Computer-generated /da/
Duration	473 ms	473 ms	470 ms	470 ms
Intensity	75 dB SPL	75 dB SPL	75 dB SPL	75 dB SPL
Average f_o	121.9 Hz	123.2 Hz	131.8 Hz	131.6 Hz
Average F_1	661.8 Hz	656.6 Hz	761.6 Hz	768.5 Hz
Average F_2	1040.3 Hz	1042.3 Hz	1150.9 Hz	1178.1 Hz
Average F_3	2553.6 Hz	2595.5 Hz	2458.2 Hz	2559.1 Hz
Consonant duration	37 ms	42 ms	34 ms	39 ms
Vowel duration	435 ms	420 ms	436 ms	431 ms

F = formant frequency; f_o = fundamental frequency; dB = decibel; SPL = sound pressure level

most frequently classified as "ba" and "da." The two selected stimuli were classified as /ba/ and /da/ in >95% of instances.

Materials consisted of audio clips of speech sounds /ba/ and /da/, and response prompts "&" and "£" (Arial, font size 48). Stimulus videos were generated in Microsoft PowerPoint and included both the prompt and the distractor audio signal to align audio stimuli and visual prompts precisely, to overcome stimuli onset lags observed in online experiments (Bridges et al., 2020). We presented the prompt at two stimulus-onset asynchronies (SOAs) relative to auditory distractor onsets. The SRC videos started with a blank screen, followed by the presentation of the symbol prompt (& or £) in the centre of the screen at 550 ms or 600 ms for 200 ms. The audio syllable started playing 750 ms from the start of the video, for 473 ms for the human stimuli and for 470 ms for the computer-generated stimuli. The video stopped playing after 3,000 ms had elapsed. We chose to use negative SOAs of −200 (SOA1) and −150 ms (SOA2) relative to the presentation of the audio distractor (Ghaffarvand Mokari et al., 2020, 2021; Roon & Gafos, 2015). This procedure ensured that the distractor stimuli would be presented at a time point at which participants would most likely be preparing their response. Thus, SOA1 started at 550 ms relative to video onset, and 200 ms before the audio start of the audio distracter, and SOA2 was presented 600 ms relative to video onset and 150 ms before the audio distracter. Note that we use '*SOA*' to indicate the onset of the prompt relative to distractor onset, following Adank et al. (2018), Kerzel and Bekkering (2000) and Wu et al. (2019). Other SRC studies have reversely defined SOA as the onset of the distractor relative to response prompt onset (Galantucci et al., 2009; Ghaffarvand Mokari et al., 2020, 2021; Roon & Gafos, 2015). Negative SOAs in the current experiment hence indicated that response prompts appeared before the auditory distractor. On Gorilla, the video was preceded by a 500-ms fixation cross and followed by a 100-ms interstimulus interval (ISI). Recordings were set to start with distractor video onset and last 3,000 ms. The trial design for the SRC task is detailed in Fig. 1.

To ensure that participants were paying attention and did not mute their sound, we included a "catch" trial in each block of the SRC task. In each catch trial, participants heard tone stimuli consisting of one, two, or three tones and were asked to report how many tones they had heard via button press. Using Praat (Boersma & Weenink, 2018), we created a 350-Hz tone with a length of 200 ms, with a sampling frequency of 44.1k Hz, saved as a mono sound. We created three .wav files for the catch trials: The first consisted of a single tone, the second of two tones spaced 32-ms apart, while the third file contained three tones, all spaced 32-ms apart.

Procedure—Eligibility screening

Participants first completed an eligibility screening study to assess the adequacy of their hardware in providing quality audio recordings. In this eligibility study, participants read the information sheet and provided consent before completing a headphone check (Woods et al., 2017). In each of six trials, participants judged which of three pure tones was the quietest. Performance below 83% correct (chance level: 33.33%) led to the immediate rejection of participants from the experiment. Participants having passed the headphone check then completed an audio recording task. In all eight trials, participants viewed a fixation cross for 500 ms followed by an SRC video without the symbol prompt. 1,500 ms from trial onset, participants produced either /ba/ and /da/ as indicated by written instructions on the screen, as quickly as possible. For 16 further trials, participants were instructed to remove their headphones and placed them close to their microphone and turning their system volume to maximum. Trials began with a 500-ms fixation cross, followed by an SRC video (3,000 ms) and a 100-ms interstimulus interval (ISI). Recordings were set to start at video onset and last 3,000 ms. This task was included to collect recordings of the participants' voice to check sound quality. Participants were considered eligible if they followed task instructions and provided good-quality audio recordings, if audio stimuli onsets recorded from their headphones in the final 16 trials fell within an 80-ms range, and if recordings were consistently clear of static or excessive background noise.

Procedure—Main task

Participants who passed the eligibility test were invited to take part in the main experiment. After reading the information sheet and providing consent, participants

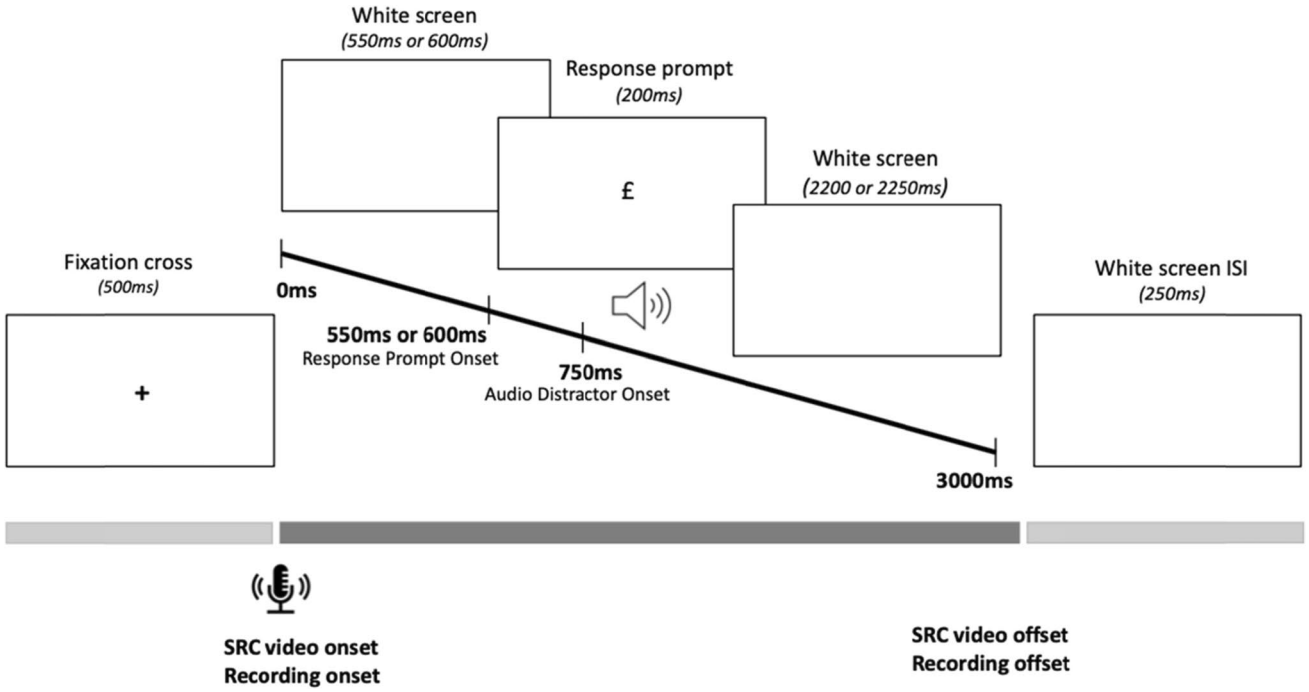

Fig. 1 Trial design for the SRC task. On Gorilla, a 500-ms fixation cross was created to precede a 3,000-ms SRC video, followed by a 250-ms interstimulus interval (ISI). Audio recordings were set to co-occur with SRC video presentation. Timings in parentheses indicate durations of specific events. The response prompt appeared at either 550 ms (SOA1) or 600 ms (SOA2) from SRC video onset

completed the headphone check as per the eligibility study. Participants then completed a two-alternative forced-choice (2AFC) phonetic identification task where they classified all four auditory stimuli (computer-generated /ba/ and /da/ and human /ba/ and /da/) as /ba/ or /da/. This task was included to establish the relative intelligibility of both types of stimuli. Each trial in the 2AFC task started with a 250-ms fixation cross, followed by a 100-ms ISI after which an audio file played. Participants indicated via key press (left or right arrow key) which sound they heard. Human and computer-generated stimuli were presented in separate 2AFC blocks. Each block consisted of two practice trials and 20 main trials (2 syllables × 10 repetitions). After the 2AFC task, participants moved on to the SRC task. They first read detailed instructions on the task, including prompt–response pairings and instructions to produce the prompted response as quickly and accurately as possible. Participants were informed that they would hear distractor sounds that were produced by a male human speaker or computer generated and were provided with examples of both types of stimuli. Participants were also informed of the catch trials. They then completed an SRC block with human stimuli and an SRC block with computer-generated stimuli, in randomized order. Before either task, participants were reminded of the prompt-response pairings and completed eight practice trials (2 prompts × 2 distractors × 2 SOAs). The main SRC task consisted of five blocks of 24 trials. The mapping of the prompt (& or £) onto the response (ba or da) was counterbalanced across participants, who were reminded of the prompt–response mappings between blocks.

A catch trial was included within each block. Catch trials started with a 500-ms fixation cross, after which one of the catch trial auditory stimuli was played and participants indicated via button press whether they heard one, two, or three tones. A minimum of 75% correct catch trials was required to be included in the final data analysis.

After the SRC task, participants completed a video onset detection test aimed at estimating latencies between recording onsets and SRC video distractor onsets. Participants received the same instructions as in the second part of the eligibility test (i.e., to place their headphones close to their microphone and turn their system volume to maximum so the sound of the stimuli could be recorded from their headphones). As in the SRC tasks, audio recordings were set to start with SRC video onset and last 3,000 ms. Here, each SRC trial type (2 conditions × 2 prompts × 2 syllables × 2 SOAs) was presented four times and recorded, for a total of 64 trials. The average delay between the expected audio onset in the SRC video stimuli (750 ms) and the recorded audio onset obtained in the recordings of the video onset detection task was used to adjust reaction time (RT) estimations for the SRC tasks.

Data processing and analysis

Responses were recorded via the audio recording zone on Gorilla.sc. Recordings were initially encoded as stereo .weba files and recoded offline as mono .wav files. Responses were coded manually and RTs annotated manually on Praat (Boersma & Weenink, 2018). The syllable detection script from the Prosogram plugin (Mertens, 2004) was first used to delimitate participants' productions for each recording (i.e., for each trial). The acquired boundaries were checked and adjusted manually to account for noise and detection errors. While audio recordings were set to coincide with the presentation of SRC video stimuli, piloting revealed delays between recording and video onsets. This asynchrony was problematic, as latencies in video stimuli onsets—and hence response prompt onsets—inflated RT estimations. To estimate these video onset latencies and correct RT computations, audio onsets were measured for each trial in the video onset detection test and compared with the audio onset in the video stimuli (750 ms). Mean video onset latencies (*measured audio onset − expected audio onset [750 ms]*) were computed for each participant and combination of syllable (ba or da), biological status, and SOA. Video onset latencies averaged 118 ms across participants (SD = 49 ms, range: 43–295 ms). We also computed standard deviations of the video onset latencies for each combination of participant, syllable, biological status, and SOA as an indicator of the variability in video onset latencies per condition. These standard deviations averaged 12 ms. To obtain RTs from prompt onsets in the SRC task, RTs measured manually from recording onsets were corrected for SOA (550 ms or 600 ms from video onset) as well as for mean video onset latencies (*RT from prompt onset = measured RT from recording onset – SOA – mean video onset latency*).

Data were analyzed with generalized linear mixed-effects models (GLMMs) using the *lme4* package in R (Bates et al., 2014). GLMMs enable the modelling of independent variables of interest (*fixed* effects) whilst considering unexplained variability within and across participants (*random* effects). A further appeal of GLMMs is that they allow for the analysis of non-normally distributed data through specifications of a distribution and link function, which is considered preferable to transformation for RT data (Balota et al., 2013; Lo & Andrews, 2015; Schramm & Rouder, 2019). In addition, a further advantage was that this analysis allowed us to avoid potential issues reported with log-transforming and subsequently back transforming RT data (Feng et al., 2013; Lo & Andrews, 2015; Manandhar & Nandram, 2021; Molina & Martín, 2018). Instead, we chose to use a gamma-distribution and identity link function to account for positive skew in RT data without the need to transform the data, following Lo and Andrews (2015).

In the current experiment, fixed effects were compatibility (compatible vs. incompatible), biological status of the distractor stimulus (human vs. computer generated), SOA (SOA1 vs. SOA2), and their interactions. The random effect structure consisted of by-participant intercepts. Errors were excluded from the RT analyses. Errors included productions of wrong or multiple responses, missing answers, anticipatory responses (RT <200 ms) and late responses (RT >1,000 ms). For each participant, observations with RTs outside of three median absolute deviations (MADs) from their median RT in each condition (combination of biological status, compatibility, and SOA) were removed from the analyses.

The error analyses used a binomial distribution and logit link function. We used a forward model building strategy to determine the best fitting model. Starting with a model with random effects only, we performed chi-squared tests to assess improvement of model fit after the inclusion of each fixed factor, from lower order to higher order effects. A factor was only maintained in the model if it significantly improved model fit. To supplement our analyses, Bayes factors (BF_{10}) were computed at each step following Jarosz and Wiley (2014) to evaluate evidence for each factor. BF_{10} quantifies the likelihood of the alternative hypothesis (H_1) over the null hypothesis (H_0). Further, Cohen's d (Cohen, 2013) was computed to estimate the effect size of the significant effects in the final model using the following formulas (Equations 1 and 2).

$$Cohen's\ d = \frac{\left(M_2 - M_1\right)}{Pooled\ SD}. \tag{1}$$

$$Pooled\ SD = \frac{\sqrt{{SD_1}^2 + {SD_2}^2}}{2}. \tag{2}$$

Results

The full data set for 32 participants consisted of 7,667 observations. For the reaction time (RT) analyses, erroneous trials were removed (687 trials, 8.96%). These included 252 erroneous responses, five missing answers, six anticipatory responses, and 424 late responses. We removed 567 observations with RTs outside of three median absolute deviations (MADs) from each participant's median at each experimental level (i.e., each combination of biological status, compatibility, and SOA). The remaining 6,413 trials were included in the RT analyses. The final model included the main effects compatibility, biological status, and SOA, as well as the interaction between compatibility and SOA (Table 2; see the Supplementary Materials for the model selection process).

Table 2 Mean raw reaction times (RTs) and standard deviations (*SD*s) in milliseconds (ms) per condition

Biological Status	SOA	Compatibility	*M* (ms)	*SD* (ms)
Human	SOA1	Compatible	686	143
		Incompatible	699	141
	SOA2	Compatible	672	128
		Incompatible	693	131
Computer-generated	SOA1	Compatible	684	139
		Incompatible	686	138
	SOA2	Compatible	662	131
		Incompatible	679	121

SOA stimulus-onset asynchrony

There was a significant main effect of compatibility with slower RTs for incompatible trials ($M = 689$ ms, S$D = 132$ ms) than for compatible trials ($M = 676$ ms, $SD = 134$ ms), $BF_{10} = 3269017$, $d = 0.10$. The results showed an overall automatic imitation effect of 13 ms (Fig. 2). The main effect of biological status was also significant, with slower RTs in response to human ($M = 688$ms, $SD = 135$ms) than to computer-generated stimuli ($M = 678$ ms, $SD = 131$ ms), $BF_{10} = 665.14$, $d = 0.08$. The main effect of SOA was significant, with slower RTs at SOA1 ($M = 684$ ms, $SD = 154$ ms) than at SOA2 ($M = 674$, $SD = 144$ ms), $BF_{10} = 54.60$, $d = 0.07$. The effect of compatibility was modulated by SOA, with larger automatic imitation effects at SOA2 ($M = 19$ ms, $SD = 27$ ms) than at SOA1 ($M = 6$ ms, $SD = 27$ ms), $BF_{10} = 20.09$, $d = 0.48$. Crucially, there was no evidence for an interaction between compatibility and biological status with similar automatic imitation effects (incompatible minus compatible RTs) in the human ($M = 16$ ms, $SD = 28$ ms) and in the computer-generated ($M = 10$ ms, $SD = 27$ ms) condition, $BF_{10} = 0.02$. The low BF_{10} provides strong evidence for the null hypothesis. Results for the errors and the 2AFC task can be found in the supplementary materials, and while there was a significant difference in identification of human

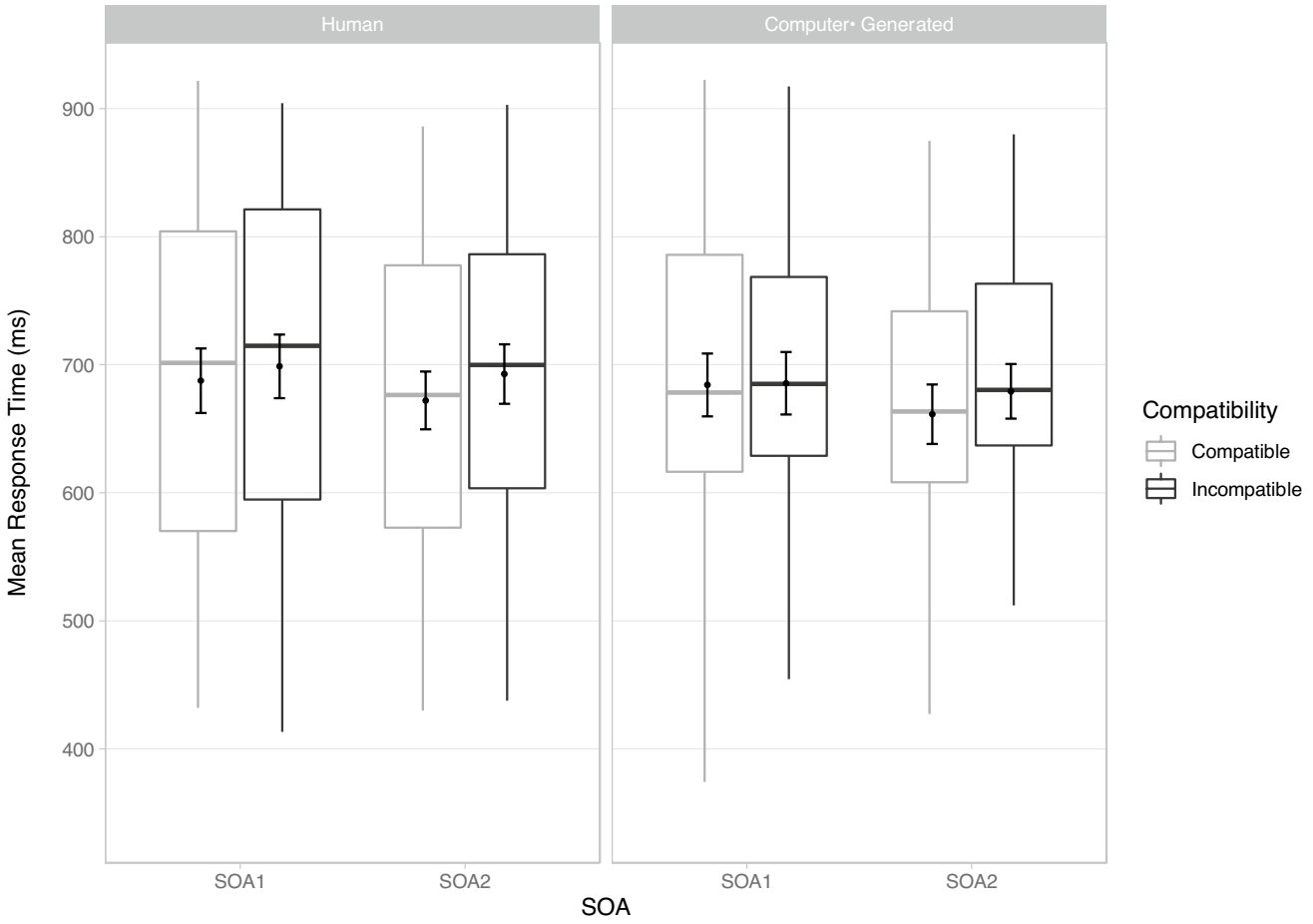

Fig. 2 Mean response times (RTs) in milliseconds (ms) for correct trials in the stimulus–response compatibility (SRC) tasks for each experimental condition. Points in the background show the raw mean RTs for each participant (points are offset on the *x*-axis for clarity). The boxplots indicate the first, second (median), and third quartiles and whiskers indicate 1.5 times the interquartile range of the distribution. Black points in the foreground show the mean and error bars indicate standard errors

and computer-generated speech, both were extremely high and well above chance (>90%).

Discussion

The present study aimed to establish whether computer-generated vocal actions evoke an automatic imitation response. Second, it aimed to determine whether this evoked response was smaller or similar to responses measured for human vocal actions. Participants responded slower and with more errors to incompatible trails than to compatible trials, displaying a clear automatic imitation effect of 13 ms. Participants also responded slower (and with lower accuracy) to the human-produced vocal actions than the computer-generated actions. There were no further significant main effects or interactions, showing that human vocal stimuli are not treated differently in terms of automatic imitation of vocal actions by the speech perception system as proposed by the integrated theory. There was a trend towards a smaller automatic imitation effect for the computer-generated vocal stimuli, but adding the interaction between compatibility and biological status did not improve model fit.

In contrast to Longo et al. (2008), we do not find any effects of top-down manipulation. Press et al. (2005, 2006) report no effect of informing participants about the human or robotic status of their stimulus materials and neither did we find a difference in automatic imitation effects between the two stimulus types, despite informing participants that the stimuli were produced by a human or computer generated. Based on Press et al. (2005, 2006) and Longo et al. and predictions from the integrated theory, we expected smaller automatic imitation effects for the computer-generated vocal stimuli, but our results did not confirm this prediction.

Integrated theory of language production and comprehension

Results did not confirm the prediction from the integrated theory that covert imitation is reduced for speakers less similar to the listener. One possibility is that both types of stimuli were perceived as equally similar to the listener and therefore both processed via the simulation pathway, as the result of a learning effect. Participants may have become overly familiarized during the 2AFC identification task (including twenty repetitions of the computer-generated stimuli) before the main SRC task. We therefore conducted an additional analysis (Supplementary Materials) to evaluate whether participants showed changes in the size of the automatic imitation effect over the course of the SRC task, but we found this effect remained stable. Therefore, it seems unlikely that any learning effects—participants getting more accustomed to the computer-generated stimuli—affected results. In addition, it seems likely the computer-generated stimuli were perceived as dissimilar to the human actions, as they were classified correctly less often than the human actions, as also reported in (Pisoni et al., 1985) for speech stimuli generated using comparable speech synthesizers.

An alternative explanation is that the theory's predictions are incorrect (or underspecified with respect to similar, dissimilar, familiar, or less familiar speech). For instance, it could be the case that all vocal stimuli are processed using the simulation route only. Using a single pathway for all incoming speech would be more parsimonious compared with two pathways, as no mechanism is needed to decide whether an incoming vocal stimulus is similar enough to be processed through the simulation route or whether needs to be processed using the association route. Processing of dissimilar speakers might thus rely as much on engagement of production mechanisms as familiar speakers, in contrast to predictions of the integrated theory.

Automatic imitation of vocal versus manual actions

Our results did not replicate Press et al. (2005, 2006), who report smaller automatic imitation effects for computer-generated manual stimuli, whereas we did not. This discrepancy could be due to differences in how computer-generated manual and vocal stimuli are perceived. Manual computerized actions (e.g., the pincher used in Press et al., 2005, 2006) may be processed as inanimate objects such as tools, while vocal computer-generated actions may instigate us to assign an identity to its bearer, as for faces (Lavan et al., 2019). While it is feasible to manipulate manual stimuli to evoke specific stereotypes (e.g., by using racial cues; cf. Correll et al., 2015), we did not explicitly implement such a manipulation and neither did we ask participants whether they assigned an identify to either type of stimulus. It is also unclear whether hands evoke stereotypical connotations as is the case for voices. Upon hearing an unfamiliar voice, listeners create a mental image of the speaker's physical appearance (Krauss et al., 2002), and imagine the physical appearance of computer-generated voices (McGinn & Torre, 2019). Moreover, voice identity perception studies demonstrated that listeners attribute traits (trustworthiness or aggression) to unfamiliar voices after brief exposures (100–400 ms; Mileva & Lavan, 2022). It is unclear whether listeners also attribute voice identity traits to computerized voices. Therefore, it seems plausible that the difference between our results and those reported in Press et al. (2005, 2006) is due to differences in how manual and vocal stimuli evoke stereotypical notions related to their origin or bearer.

Limitations and future directions

The Klatt synthesizer used was unsophisticated compared with current speech synthesizers (Wagner et al., 2019). We intended to establish a baseline of the capacity of synthetic stimuli to evoke an automatic imitation effect, so we selected a synthesizer that created synthetic-sounding stimuli, so stimuli were not confused with human-produced stimuli and avoid a situation where participants believed that both stimuli were human produced. We intended to keep stimulus intelligibility stable for human and computer-generated conditions. The computer-generated stimuli were less intelligible than human stimuli (96.48% computer-generated stimuli vs. 99.53% human stimuli; cf. Supplementary Materials). Future experiments could first disentangle biological status and intelligibility, as these factors were confounded in our study. For instance, human and computer-generated stimuli could be degraded parametrically using noise. Using computer-generated and human stimuli equalized for intelligibility, it would be feasible to also introduce a top-down manipulation (e.g., by informing one group of participants that all stimuli are human vs. telling a different group that all stimuli are computer-generated). This way, separate effects of top-down (stimulus beliefs) and bottom-up (intelligibility) on automatic imitation of vocal actions could be established. Second, future experiments could clarify the relationship between familiarity/similarity and biological status. For instance, per the integrated theory, processing of computer-generated speech matching the listener's regional or foreign accent could rely more on the simulation than on the association route and vice versa. An experiment with a two-factor factorial design with familiarity and biological status as factors could establish whether this prediction can be confirmed.

Participants responded 10ms faster to computer-generated trials. This effect counters findings for similar condition-related effects reported in Press et al. (2006) and Longo et al. (2008), who found slightly faster RTs (Press et al., 2006) for human compared with computer-generated stimuli or no RT difference between impossible and possible actions (Longo et al., 2008). We suspect that this difference between our study and the studies might modality specific. Perhaps observers respond differently to manual and speech stimuli differing in biological status, because our speech stimuli can be categorially perceived (Liberman et al., 1967) and the manual stimuli used in Press et al. (2006) and Longo et al. cannot. This possibility could be explored in future experiments—for example, using a factorial design with modality (manual or speech responses) and biological status (human and computer-generated) fully crossed.

Conclusions

Our results demonstrated that vocal stimuli from human and non-human origins evoke similar covert imitation. The speech perception system is therefore not tailored towards vocal stimuli produced by fellow humans and may treat any speech-like signal the same. Our results further demonstrate that computer-generated vocal stimuli evoked a covert imitative response. Finally, our results have implications for the integrated theory of language comprehension, but also for research in human-computer interactions and voice identity research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-022-02218-6.

Acknowledgements We would like to thank Mark Huckvale for generating the computer-generated stimuli and Nadine Lavan for constructive discussion. This work was supported by a Project Grant to Patti Adank by the Leverhulme Trust under project number RPG-2018-043, Mechanisms Governing Imitation of Speech.

References

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science, 21*(12), 1903–1909.

Adank, P., Nuttall, H. E., Bekkering, H., & Maegherman, G. (2018). Effects of stimulus response compatibility on covert imitation of vowels. *Attention, Perception, & Psychophysics, 80*(5), 1290–1299.

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1563–1571.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv:1406.5823 [Stat]*. http://arxiv.org/abs/1406.5823

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer (Version 6.0. 37) [Computer program]. http://www.praat.org/

Brass, M., Wohlsläger, A., Bekkering, H., & Prinz, W. (2000). Compatibility between observed and executed finger movements: Comparing symbolic, spatial and imitative cues. *Brain and Cognition, 44*, 124–143.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ, 8*, Article e9414.

Buccino, G., Binkofski, F., & Riggio, L. (2004). The mirror neuron system and action recognition. *Brain and Language, 89*, 370–376.
Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology, 108*(2), 219–233.
Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., De Coster, L., Radkova, I., Deschrijver, E., & Brass, M. (2018). Automatic imitation: A meta-analysis. *Psychological Bulletin, 144*(5), 453–500.
Fadiga, L., Buccino, G., Craighero, L., Fogassi, L., Gallese, V., & Pavesi, G. (1998). Corticospinal excitability is specifically modulated by motor imagery: A magnetic stimulation study. *Neuropsychologia, 37*(2), 147–158.
Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*(2), 399–402.
Feng, C., Wang, H., Lu, N., & Tu, X. M. (2013). Log transformation: Application and interpretation in biomedical research. *Statistics in Medicine, 32*(2), 230–239.
Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics, 71*(5), 1138–1149.
Ghaffarvand Mokari, P., Gafos, A., & Williams, D. (2020). Perceptuomotor compatibility effects in vowels: Beyond phonemic identity. *Attention, Perception, & Psychophysics, 82*, 2751–2764.
Ghaffarvand Mokari, P., Gafos, A., & Williams, D. (2021). Perceptuomotor compatibility effects in vowels: Effects of consonantal context and acoustic proximity of response and distractor. *JASA Express Letters, 1*(1), Article 015204.
Gowen, E., & Poliakoff, E. (2012). How does visuomotor priming differ for biological and non-biological stimuli? A review of the evidence. *Psychological Research, 76*, 407–420.
Heyes, C. (2011). Automatic imitation. *Psychological Bulletin, 137*(3), 463–483.
Jarick, M., & Jones, J. A. (2009). Effects of seeing and hearing speech on speech production: A response time study. *Experimental Brain Research, 195*, 175–182.
Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving, 7*(1), Article 2.
Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 634–647.
Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America, 67*(3), 971–995.
Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*(6), 618–625.
Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*(1), 90–102.
Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.
Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*, 1171.
Longo, M. R., Kosobud, A., Berthenthal, B., & I. (2008). Automatic imitation of biomechanically possible and impossible actions: Effects of priming movements versus goals. *Journal of Experimental Psychology: Human Perception and Performance, 34*(2), 489–501.
Manandhar, B., & Nandram, B. (2021). Hierarchical Bayesian models for continuous and positively skewed data from small areas. *Communications in Statistics—Theory and Methods, 50*(4), 944–962.
McGinn, C., & Torre, I. (2019). Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. *2019 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, 211–221.
Mertens, P. (2004, March). *The prosogram: Semi-automatic transcription of prosody based on a tonal perception model.* Speech Prosody 2004, International Conference, Nara, Japan.
Mileva, M., & Lavan, N. (2022). *How quickly can we form a trait impression from voices?* PsyArXiv. https://doi.org/10.31234/osf.io/zd4un
Molina, I., & Martín, N. (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics, 46*(5), 1961–1993.
Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., & Adank, P. (2016). The effect of speech distortion on the excitability of articulatory motor cortex. *NeuroImage, 128*, 218–226.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347.
Pisoni, D., Nusbaum, H., & Greene, B. (1985). Perception of synthetic speech generated by rule. *Proceedings of IEEE, 73*, 1665–1676.
Press, C., Bird, G., Flach, R., & Heyes, C. (2005). Robotic movement elicits automatic imitation. *Cognitive Brain Research, 25*(3), 632–640.
Press, C., Gillmeister, H., & Heyes, C. (2006). Bottom-up, not top-down, modulation of imitation by human and robotic models. *European Journal of Neuroscience, 24*(8), 2415–2419.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.
Ralston, J. V., Pisoni, D. B., Lively, S. E., Greene, B. G., & Mullennix, J. W. (1991). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors, 33*(4), 471–491.
Roon, K. D., & Gafos, A. I. (2015). Perceptuo-motor effects of response-distractor compatibility in speech: Beyond phonemic identity. *Psychonomic Bulletin & Review, 22*(1), 242–250.
Schramm, P., & Rouder, J. (2019). *Are reaction time transformations really beneficial?* PsyArXiv. https://doi.org/10.31234/osf.io/9ksa6
Stürmer, B., Aschersleben, G., & Prinz, W. (2000). Correspondence effects with manual gestures and postures: A study of imitation. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), 1746–1759.
Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, Č., & Tånnander, C. (2019, September). *Speech synthesis evaluation—State-of-the-art assessment and suggestion for a novel research program*. Proceedings of the 10th Speech Synthesis Workshop (SSW10), Vienna, Austria.
Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*(8), 989–994.
Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2
Wu, Y., Evans, B., & Adank, P. (2019). Sensorimotor Training Modulates Automatic Imitation of Visual Speech. *Psychonomic Bulletin & Review, 26*, 1711–1718. https://doi.org/10.3758/s13423-019-01623-8

Open access statement The aims, predictions, design, and proposed analysis of the experiment were preregistered online (https://aspredicted.org/dr552.pdf) under number 67457 "*Automatic imitation of synthetic speech.*" All stimulus materials, Klatt and R scripts, and raw (text) data can be found on the Open Science Framework (https://osf.io/k6q28).