Behavior Research Methods
https://doi.org/10.3758/s13428-018-1037-4

Is automatic speech-to-text transcription ready for use in psychological experiments?

Kirsten Ziman[1] · Andrew C. Heusser[1] · Paxton C. Fitzpatrick[1] · Campbell E. Field[1] · Jeremy R. Manning[1]

Abstract

Verbal responses are a convenient and naturalistic way for participants to provide data in psychological experiments (Salzinger, The Journal of General Psychology, 61(1),65–94:1959). However, audio recordings of verbal responses typically require additional processing, such as transcribing the recordings into text, as compared with other behavioral response modalities (e.g., typed responses, button presses, etc.). Further, the transcription process is often tedious and time-intensive, requiring human listeners to manually examine each moment of recorded speech. Here we evaluate the performance of a state-of-the-art speech recognition algorithm (Halpern et al., 2016) in transcribing audio data into text during a list-learning experiment. We compare transcripts made by human annotators to the computer-generated transcripts. Both sets of transcripts matched to a high degree and exhibited similar statistical properties, in terms of the participants' recall performance and recall dynamics that the transcripts captured. This proof-of-concept study suggests that speech-to-text engines could provide a cheap, reliable, and rapid means of automatically transcribing speech data in psychological experiments. Further, our findings open the door for verbal response experiments that scale to thousands of participants (e.g., administered online), as well as a new generation of experiments that decode speech on the fly and adapt experimental parameters based on participants' prior responses.

Keywords Annotation · Free recall · Mechanical Turk · Memory · Speech-to-text · Verbal response

Introduction

Speech-to-text engines became popular in the 1990s (Kurzweil et al., 1990) when the performance of speech recognition algorithms (primarily based on Hidden Markov Models; Rabiner (1989)) reached sufficient levels to provide plausible, though still often inaccurate, transcripts (Bamberg et al., 1990). Recent advances in deep learning have ushered in a new era of substantially more accurate speech recognition (Hinton et al., 2012). Today, speech-to-text engines are ubiquitous, and are embedded into applications running on myriad devices ranging from phones to watches to thermostats to cars and beyond.

While automated speech decoding is now widespread in mainstream society, the technology has not yet been widely adopted by the psychological research community to facilitate analyses of verbal responses. However, speech decoding has the potential to save researchers an enormous amount of time when analyzing verbal response data, and to enable new experimental designs that adapt based on parameters derived from decoded speech data. Further, whatever their current limitations, as speech-to-text algorithms continue to mature, their utility in psychological research should improve as well.

We sought to explore the feasibility of embedding a modern speech-to-text translation engine into a psychological experiment that relies on verbal responses as its primary data source. As a proof of concept, we had participants study and verbally recall a series of random word lists. We had human annotators manually transcribe the recorded audio data (UPenn Computational Memory Lab, 2015), and we also transcribed the data automatically using the Google Cloud Speech API (Halpern et al., 2016). We then carried out a series of analyses to compare the human-generated and computer-generated transcripts.

Overall, we found that the human-generated and computer-generated transcripts matched to a high degree.

✉ Jeremy R. Manning
jeremy.r.manning@dartmouth.edu

[1] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA

Published online: 23 April 2018

Our main interest was in assessing the extent to which the computer-generated transcripts recovered (with high fidelity, treating the human-generated transcripts as the "ground truth") the major patterns in free recall dynamics that have been well reported in the literature (Murdock 1962; Kahana 1996, 2012, 2017; Manning et al. 2015). We also identified points of disagreement between the two transcription methods, particularly in how they handled non-recall vocalizations. Our results suggest that automated speech-to-text transcription tools are mature enough to provide (within limits) a viable alternative to human annotation. This provides a potential means of carrying out verbal response experiments on thousands of participants on online platforms such as Amazon's Mechanical Turk (Crump et al., 2013). Furthermore, the possibility of incorporating this technology into experiments that adapt on the fly according to prior verbal responses (where rapid ongoing manual transcription would be infeasible) is particularly exciting.

Methods

Participants

Thirty Dartmouth undergraduate students (22 female, 8 male, aged 18–21) participated in our study. All participants had (by self-report) normal or corrected-to-normal vision, reading, memory, and attentional abilities. Each participant gave written, informed consent to volunteer for our study. They received course credit for their participation. Our experimental protocol was approved by the Committee for the Protection of Human Subjects at Dartmouth College.

Materials

We collected data in a sound-attenuated testing room, using a 27-inch 2016 iMac desktop computer. All audio was recorded using the iMac's built-in microphone. The experiment was implemented in jsPsych (de Leeuw, 2015) and psiTurk (Gureckis et al., 2015), along with custom code for sending audio data to the Google Cloud Speech API.

Our stimulus set comprised a pool of 256 words chosen from an online repository of themed word lists (Col, 2017). To create the word pool, we (manually) chose 8, 12, 16, or 20 common words from each of 15 semantic categories: body parts, building-related, cities, clothing, countries, flowers, fruits, insects, instruments, kitchen-related, mammals, states, tools, trees, and vegetables.

Our experiment code and data may be downloaded here. We also created an open-source Python toolbox for analyzing and plotting free recall data, and for automatically transcribing audio data (Heusser et al., 2017).

Experimental paradigm

Each participant studied a total of eight lists comprising 16 words each (128 words total). The lists were structured such that each contained four exemplars from each of four non-overlapping (but otherwise randomly selected from the pool) semantic categories, and each word appeared in (at most) one list. The specific set of 128 to-be-studied words were chosen anew for each participant, and the lists were generated randomly (with the words on each list shuffled randomly) for each participant. All text was displayed in black Courier New font, centered vertically and horizontally on a white background, and each letter was sized to occupy 5% of the screen width.

During each experimental *trial* (Fig. 1), the participant studied and recalled words from a single 16-word list. Each trial began with 2 s of blank white screen, followed by a 2-s presentation of the first word on the list, followed by two more seconds of blank white screen. Each subsequent word was presented for 2 s, with a 2-s inter-stimulus interval (blank screen) before the next word's presentation. Two seconds after the last word was cleared from the screen, a red microphone icon appeared in the center of the screen, which prompted the participant to verbally recall as many words as they were able, from the just-presented list. The participant was given 60 s to recall the words "in the order they [came] to mind." Participants were instructed (at the beginning of the experiment) to speak "slowly and clearly" in order to facilitate analyses of their verbal response data. After 60 s, the microphone icon disappeared, the trial concluded, and the participant was given the opportunity to take a brief break before initiating the next trial.

Speech-to-text transcription

Each participant contributed a total of eight 60-s recordings of their verbal recalls of the studied word lists (one recording per list). We transcribed each recall recording into text manually using human transcribers (i.e., *human-generated*) and automatically using the Google Cloud Speech API (i.e., *computer-generated*).

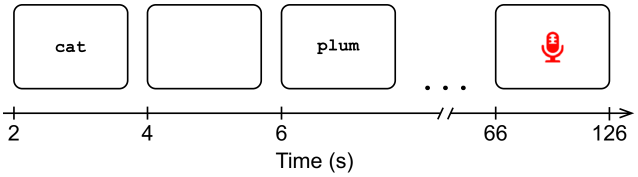

Fig. 1 Experimental paradigm. The timeline displays the sequence of events during a single experimental trial, during which the participant studies and recalls words from a single list

Human-generated transcripts

Two co-authors of this paper (PCF and CEF) manually transcribed the audio data using a transcription software tool, Penn TotalRecall (UPenn Computational Memory Lab, 2015). The transcribers listened to each trial's audio file in turn, played back at 1x speed, pausing or repeating playback as often as needed for them to be confident in their transcripts. Using the full 256-item word pool as a reference, any clear mispronunciations (e.g., "marimbo" instead of "marimba") or plurality errors (e.g., "hips" instead of "hip") were corrected to match the words in the word pool. In addition, any utterances that were judged by the transcribers to be non-recall vocalizations (e.g., "um," "wait, let me think...," etc.) were excluded from the transcript. These transcription decisions (to make pronunciation and plurality corrections, and to exclude non-recall vocalizations) were intended to highlight aspects of speech-to-text transcribing that human listeners might be especially well suited to, relative to automated methods.

Computer-generated transcripts

We used the Google Cloud Speech API to produce a computer-generated text transcript of each participant's verbal responses. A total of 240 audio files, totaling four hours of recordings, were transcribed (eight 1-min recordings per participant, for each of the 30 participants). We passed the 256-item word pool to the automatic transcriber as a *speech context*, which provides "hints" to the speech recognizer about which words to expect. Note that we did not pass any information about which specific words were reflected in any specific audio file, and only half of the total word pool was presented to any given participant. The speech recognizer returned, for each audio file, a list of automatically transcribed words and vocalization onset times. In addition, for each decoded utterance, the speech recognizer returned a confidence rating ranging from 0 (not confident) to 1 (highly confident); these confidence ratings roughly correspond to the estimated probability that the given word label matched the given speech utterance. The implementation details of the Google Cloud Speech API are proprietary, but the API is made publicly available here.

Results

We sought to evaluate the transcription accuracy of a modern speech-to-text engine applied to recordings of verbal responses from a list-learning experiment. We used the annotations of human transcribers as a benchmark. We carried out a preliminary analysis to assess the degree of absolute agreement between the human-generated and computer-generated transcripts. We then carried out a series of post hoc analyses to evaluate how well the computer-generated transcripts recovered the detailed recall dynamics (Kahana 2012, 2017; Manning et al. 2015) reflected in the human-generated transcripts.

Transcription accuracy

An accurate computer-generated transcript should satisfy three basic criteria. First, it should have a high *hit rate*, in that the computer-generated transcript should contain each of the words also contained in the human-generated transcript. We defined the hit rate as the average (across lists) proportion of words in the human-generated transcripts that were also contained in the computer-generated transcripts. Second, it should have a low *false alarm rate*, in that the computer-generated transcript should *not* contain words that were not also contained in the human-generated transcript. We defined the false alarm rate as the average (across lists) proportion of words in the computer-generated transcripts that were *not* contained in the human-generated transcripts. Third, for words in both sets of transcripts, the speech onset times should match well.

Because the speech-to-text engine we evaluated is probabilistic, each outputted response in the computer-generated transcripts is associated with a confidence rating. For this analysis, we used the receiver operating characteristic (ROC) to evaluate how the hit rate and false alarm rate varied as a function of the speech-to-text engine's confidence ratings (Fig. 2; area under the ROC: 0.907). Our analysis revealed that the human-generated transcripts matched the computer-generated transcripts well, in terms

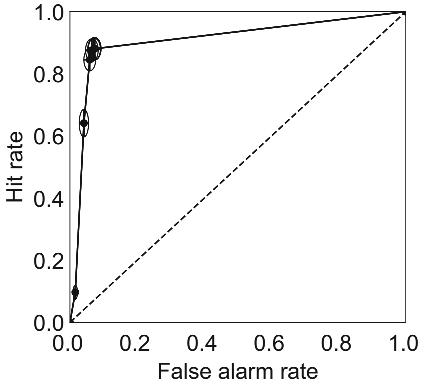

Fig. 2 Receiver operating characteristic (ROC) curve. False alarm rate and hit rate as a function of the speech-to-text engine's confidence ratings (evaluated on the interval [0, 1] in increments of 0.1). The ROC curve reflects an average across a total of 240 lists studied by 30 participants. *Error ellipses* denote 95% confidence intervals (across subjects)

of the set of words each contained. This finding, that the two sets of transcripts matched well, indicates that the verbal responses transcribed by the speech-to-text engine were an accurate reflection of what the participants actually said (as judged by human observers). Note that in subsequent analyses we ignored the speech-to-text engine's confidence ratings (i.e., we included every transcribed word, regardless of rated confidence, in our analyses below).

In addition to evaluating the degree of match between the words identified in the human-generated and computer-generated transcripts, we also compared the speech onset times of words that appeared in both transcripts. We first correlated the onset times within list, whereby we obtained a total of eight correlations for each participant (one per list). The correlations on every list exceeded 0.99 (Fig. 3a). We next correlated the manually and automatically tagged onset times within subject, aggregating across all of the lists they encountered. We designed this analysis to catch potential failures of the speech-to-text engine to accurately identify differences in speech onset times across lists. Both sets of transcripts again displayed highly correlated onset times; all correlations exceeded 0.995 (Fig. 3b). Last, we correlated the manually and automatically tagged onset times of all recalls, aggregated across all lists and participants. We designed this analysis to catch potential failures of the speech-to-text engine to accurately identify differences in speech onset times across subjects. Again, the two sets of onsets times matched closely ($r = 0.99$, $p < 0.001$; Fig. 3c). Taken together, these onset time analyses indicate that the speech-to-text engine accurately identified speech onset times as identified by human annotators.

The above analyses show that, to a first approximation, the human-generated and computer-generated transcripts agreed well in terms of the words they contained and the times at which those words were vocalized. We next sought to evaluate the degree to which the computer-generated transcripts captured the detailed recall dynamics reflected in the human-generated transcripts.

Recall dynamics

Participants in our free-recall experiment studied and recalled random word lists. In general, the free-recall literature has characterized participants' recall dynamics along four dimensions (for review see Kahana (2012)). First, given a just-studied random word list, which word do participants tend to recall first? Second, in which order(s) do participants transition from recalling one word to the next? Third, which words do participants recall overall? And fourth, what sorts of errors (recalls of words that they had *not* studied) do participants make? We evaluated the degree to which the computer-generated transcripts captured each of these dimensions as compared with the human-generated transcripts.

Probability of first recall

The *probability of first recall* curves in Fig. 4a display the proportion of trials in which participants began their recall sequences with words at each study position. In other words, which words on the just-studied lists did participants tend to recall first? The probability of first recall curves derived from the human-generated and computer-generated transcripts overlapped to a high degree. These curves indicate that participants in this experiment exhibited a strong *primacy effect*, whereby they most often began their recall sequence by recalling the first word presented on the just-studied list.

To better characterize the degree of match between the human-generated and computer-generated transcripts, and following our prior work (Manning et al., 2011), we defined a *primacy ratio* as the average probability of initiating recall with any of the first three studied words, divided by the average probability of initiating recall with any of the middle six studied words from the most recent list. This yielded a pair of numbers for each participant; the first described the strength of the primacy effect as measured from the human-generated transcripts, and the

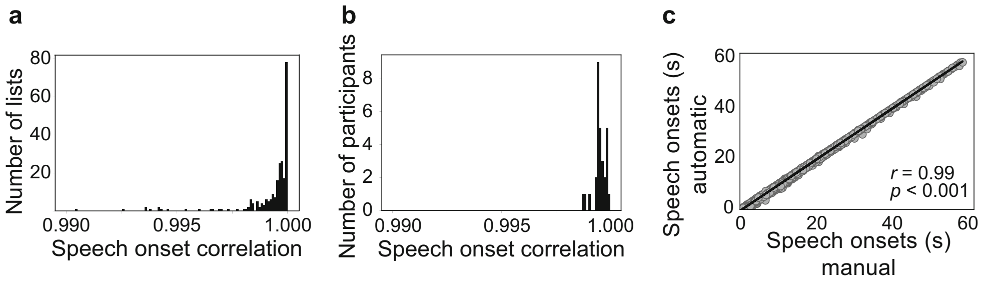

Fig. 3 Speech onset times during recall. **a** Within-list correlations between human-generated and computer-generated speech onset times during recall. Each participant contributed data for eight lists. **b** Within-subject correlations between human-generated and computer-generated onset times. **c** Onset times for individual recalls, as identified manually and automatically. Each recall appears as a single *dot*

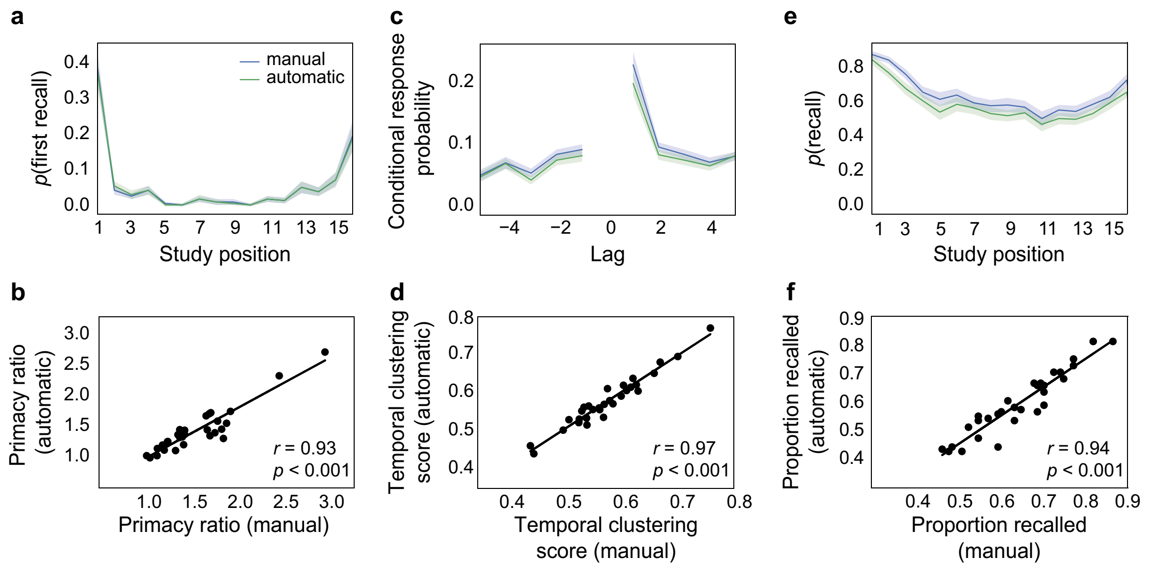

Fig. 4 Recall dynamics. **a, b** Initiating the recall sequence. **a** Probability of recalling each word first, as a function of its presentation position. Participants most often began their recall sequences with the first-presented word from the just-studied list. **b** The "primacy ratio" (see text) reflects participants' tendency to initiate recall with words presented early (versus in the middle of the list). We assessed the agreement between the strength of this primacy effect as measured using the human-generated (*manual*) and computer-generated (*automatic*) transcripts. Each *dot* reflects the average primacy ratios for one participant. **c–d** Recall transitions. **c** The conditional probability of recalling each word as a function of its presentation position relative to the previously recalled word (lag). Participants often temporally cluster their recalls by successively recalling words that were presented at nearby positions on the list (Kahana, 1996). **d** We assessed the agreement between the degree of temporal clustering as measured manually and automatically. Each *dot* reflects the average temporal clustering scores (see text) for one participant. **e, f** Overall recall probabilities. **e** Probability of recalling each word as a function of its presentation position. **f** We assessed the agreement between the average proportion of words recalled as measured manually and automatically. *Error ribbons* in **a**, **c**, and **e** denote 95% confidence intervals (across subjects), estimated via 5000 bootstrap iterations

second described the strength of the primacy effect as measured from the computer-generated transcripts. These two measures were highly correlated across participants ($r = 0.93$, $p < 0.001$; Fig. 4b), reflecting the high degree of agreement between the human-generated and computer-generated transcripts.

Recall transition probabilities

Given that a participant has just recalled a word from the just-studied list, which word are they likely to recall next? The lag conditional response probability curves (Kahana, 1996) displayed in Fig. 4c reflect the conditional probability of recalling each word on the just-studied list as a function of its study position relative to the previously recalled word (*lag*). The curves show that participants tend to successively recall words that came from nearby study positions on the studied lists, a phenomenon referred to as *temporal clustering*.

Following (Polyn & Kahana, 2008), we defined a *temporal clustering score* for each participant, reflecting their average tendency to successively recall words that came from nearby study positions. For each recall transition, we create a distribution of the absolute values of the differences (lags) between the study position of the just-recalled word and the set of words that had not yet been recalled. We then computed the percentile rank (in the distribution of absolute lags) of the next word the participant recalled. When we observed a tie, we assigned that recall the average percentile rank of all similarly ranked potential recalls. We defined the temporal clustering score as the average percentile rank across all recalls, from all lists, from that participant (we first averaged the ranks of recalls from each list, and then averaged across lists). If the participant always recalled the closest yet-to-be-recalled word, they would be assigned a temporal clustering score of 1. If the participant recalled the words in a random order (with respect to the words' study positions) this would yield a temporal clustering score of 0.5. We computed temporal clustering scores for each participant using both the human-generated transcripts and the computer-generated transcripts; the two transcripts yielded highly similar temporal clustering scores (Fig. 4d; $r = 0.97$, $p < 0.001$).

Overall recall probabilities

Prior work on free recall has established that participants are more likely to remember words that they studied at the beginning or end of a list, relative to middle words (these are often referred to as the *primacy effect* and *recency effect*, respectively; Murdock (1962)). We plotted the proportion of words that participants recalled as a function of their study position and found that the human-generated and computer-generated transcripts agreed well and exhibited similar primacy and recency effects (Fig. 4e). We also considered the overall proportion of studied words that participants remembered, as measured using the human-generated and computer-generated transcripts. The two types of transcripts agreed well (Fig. 4f; $r = 0.94$, $p < 0.001$).

Taken together, the above analyses show that the specific words and onset times identified in the human-generated and computer-generated transcripts agreed well in terms of identifying the specific sequences of words participants remembered from the lists they studied, and the precise timing of each utterance. We next turn to a series of analyses aimed at characterizing the errors participants made, as identified using the human-generated and computer-generated transcripts.

Recall errors

We first examined *prior list intrusion errors*, whereby participants mistakenly recalled a word from an earlier list in the experiment, rather than from the most recently studied list. Previous work has established that prior list intrusions are made more often from recently studied lists (e.g., the list before the most recent one) than from lists from much earlier in the experiment (e.g., five lists back in the experiment; for review see Kahana (2017)). Both the human-generated and computer-generated transcripts reflected this pattern, and agreed closely (Fig. 5a). For each participant, we also computed the average proportion of prior list intrusions that involved words from one list back, two lists back, and so on (up to six lists back). We computed these proportions using the human-generated and computer-generated transcripts and compared the results (Fig. 5b–f). We found that the numbers of prior list intrusions identified using both methods matched reliably (one back: $r = 0.88$, $p < 0.001$; two back: $r = 0.93$, $p < 0.001$; three back: $r = 1.00$, $p < 0.001$, five back: $r = 0.56$, $p < 0.005$; six back: $r = 1.00$, $p < 0.001$; both manual and automatic transcripts yielded zero prior list intrusions from any participant from four lists back).

In addition to prior list intrusions, participants occasionally make *extra-list intrusion errors* by recalling words that had never been presented. Whereas the human transcribers intentionally filtered out non-recall vocalizations, the speech-to-text software effectively treated all vocalizations as recalls. For example, a human transcriber would treat the vocalization "the other two were also cities" as a non-recall vocalization, whereas the speech-to-text engine treats this as a series of six successive recalls. Similarly, a human transcriber would treat the vocalization "marimba, oh, did I already say harmonica?" as reflecting two recalls (of 'marimba' and 'harmonica'), whereas the speech-to-text software labels the utterance as a series of seven successive recalls. In other words, whereas human transcribers

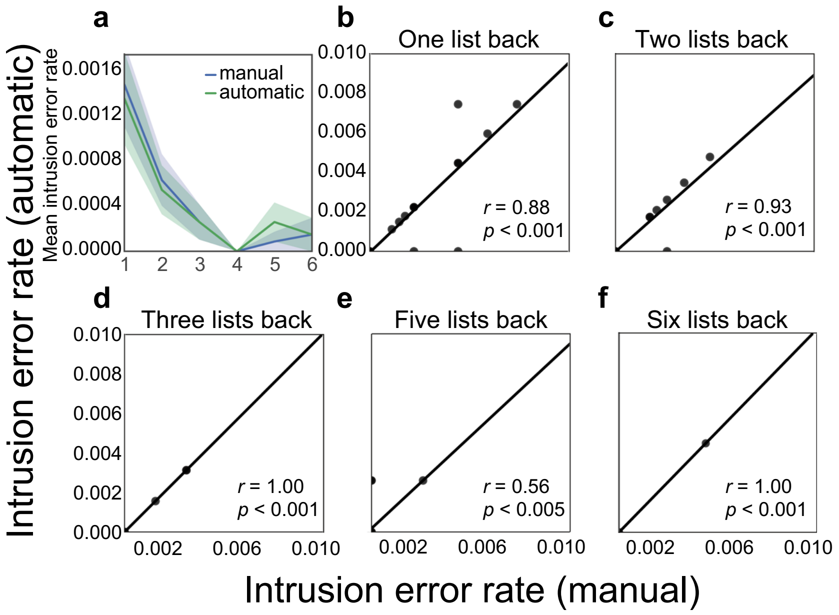

Fig. 5 Prior list intrusions. **a** Average intrusion error rates for each list back (1-3, 5, and 6) as measured using the human-generated (*blue*) and computer-generated (*green*) transcripts. *Error ribbons* denote 95% confidence intervals (across subjects), estimated via 5000 bootstrap iterations. **b-f** Each *dot* reflects the average proportion of prior list intrusion errors made by a single participant, as measured using the human-generated (x-axis) and computer-generated (y-axis) transcripts. The error rates are separated into five panels according to which list the errors came from, relative to the just-studied list

easily identify instances of participants "thinking aloud," automated methods do not distinguish recall from non-recall utterances.

Given this distinction between the manual and automatic transcriptions, we expected that there would (artificially) be a greater number of extra-list intrusions identified in the computer-generated, versus human-generated, transcripts (indeed, this pattern was reflected in our data; $t(29) = 7.15$, $p < 0.001$). However, although the computer-generated transcripts overestimated the numbers of extra-list intrusions, the numbers of extra-list intrusions identified in the human-generated and computer-generated transcripts were reliably correlated ($r = 0.53$, $p < 0.005$; Fig. 6). This indicates that the computer-generated transcripts do not accurately reflect the *absolute* proportions of extra-list errors, but they do accurately reflect the *relative* proportions of extra-list errors.

Discussion

To gain insight into the extent to which modern speech-to-text engines might replace human annotators, we carried out a series of analyses on verbal responses recorded during a list-learning experiment. We found that the human-generated and computer-generated transcripts were largely in agreement. The computer-generated transcripts also accurately reflected most of the detailed statistical patterns that we identified in participants' recall behaviors using human-generated transcripts. The major point of disagreement between the human- and computer-generated transcripts concerned how errors were reflected in the two types of transcripts. Whereas human annotators filtered out non-recall vocalizations, the computer-generated transcripts treated all vocalizations as recalls. This inflated the number of extra-list intrusion errors present in the computer-generated transcripts. Despite this, the computer-generated transcripts still accurately reflected individual variations in the relative numbers of errors generated by different participants. Overall, our results indicate that modern speech-to-text engines can accurately transcribe participants' verbal responses. To the extent to which direct transcripts are sufficient for capturing the behavioral phenomena of interest, our findings suggest that computer-generated transcripts can be used to capture and characterize verbal response patterns. This may carry substantial savings (of time and money) compared with human-generated transcripts. When large amounts of response data are collected (e.g., investigations into the effects of overt rehearsal on free and serial recall; Rundus (1971), Tan and Ward (2000), and Tan and Ward (2008)) these savings may be particularly beneficial.

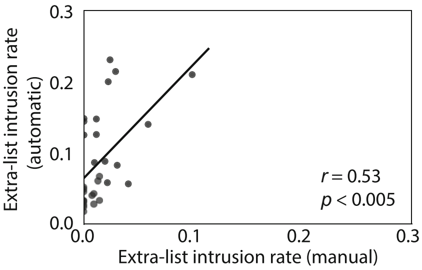

Fig. 6 Extra-list intrusions. Each *dot* reflects the average proportion of extra-list intrusions made by a single participant, as measured using the human-generated (x-axis) and computer-generated (y-axis) transcripts

A note on our choice of speech-to-text engine

Our analyses in this manuscript leveraged a single speech-to-text engine (Halpern et al., 2016). We chose the Google Cloud Speech API due to its ease of use, the ability to provide a "speech context" (which played an important role in improving the transcription accuracy), the ability to obtain confidence ratings for each transcribed utterance, and the ability to automatically identify vocalization onset times. We have intentionally avoided detailed comparisons between this speech-to-text engine and the other promising speech-to-text engines available today that may have other advantages or disadvantages (e.g., Pocketsphinx; Huggins-Daines et al. (2006)). Rather, the focus of our current analyses is to provide a proof-of-concept example of how modern speech-to-text engines can transcribe verbal response data. Our results highlight the immediate promise of existing speech-to-text technologies, and we expect that the quality of computer-generated transcripts will improve as the methods continue to mature.

Speech-to-text engines as a driver for scalable online verbal response experiments

Amazon's Mechanical Turk, launched in 2005, is an online marketplace that enables individual *requesters* to post small jobs that are carried out (usually in return for a small payment) by *workers* throughout the world. Over the past several years, psychological researchers and social scientists have begun to use Mechanical Turk as a convenient and low-cost platform for quickly collecting large amounts of experimental data (Paolacci et al., 2010). Despite the decreased level of control over the experimental environment relative to in-laboratory experiments, Mechanical Turk workers yield (for many, though not all, experiments) high-quality behavioral data that are similarly reliable to data collected in the laboratory (Paolacci et al., 2010; Buhrmester et al., 2011; Crump et al., 2013). Recently, developed tools like jsPsych (de Leeuw, 2015) and psiTurk

(Gureckis et al., 2015) facilitate the transition of laboratory experiments to the Mechanical Turk marketplace, substantially lowering the barriers to entry for research psychologists.

Our study suggests the feasibility of collecting verbal response data through Mechanical Turk. Whereas verbal responses are traditionally transcribed by human listeners (an approach that cannot easily scale to thousands of hours of recordings collected from thousands of participants via Mechanical Turk), automatic parsing via speech-to-text engines provides a potential avenue for quickly and cheaply transcribing vast quantities of data.

One potential downside to automated speech-to-text parsing is that these engines can be vulnerable to adversarial attacks, whereby malicious users intentionally generate recordings that the speech-to-text engine will reliably transcribe incorrectly (e.g., Carlini and Wagner (2018)). While we would not expect this to be a widespread problem in typical online experimental settings, researchers are also beginning to devise strategies for counteracting adversarial examples (Madry et al., 2017). Nevertheless, the potential existence of adversarial examples (which a human observer would likely have transcribed correctly) should be considered on an as-needed basis when applying these methods to massive online experiments that cannot easily be manually checked in detail.

Speech-to-text engines as a driver for adaptive verbal response experiments

Adaptive tests and experiments can dramatically reduce the time needed to assess knowledge and measure psychophysical and neuropsychological parameters. For example, computer adaptive testing is now widespread on standardized tests including the Graduate Record Examination (van der Linden & Glas, 2000), and the staircase method is commonly used to rapidly estimate participants' psychophysical thresholds (Cornsweet, 1962). Modern variants of this technique, such as Bayesian active learning, use adaptive experiments to quickly map complex multivariate receptive fields based on neural data (Park & Pillow, 2012). Over the past several decades, researchers have also developed adaptive psychological experiments that leverage real-time processing of physiological signals, such as functional magnetic resonance imaging (Cox et al., 1995; Cox & Jesmanowicz, 1999; Cohen, 2001; deCharms, 2008; deBettencourt et al., 2015) and electroencephalography (e.g., Angelakis et al. (2007)).

Adaptive experiments driven by vocal responses have been limited, presumably because sufficiently accurate speech-to-text engines have only recently been broadly available. However, the computer-generated transcripts in our study captured many of the key patterns in participants' recall sequences. These transcripts may be generated on the fly during an experiment, for use in adapting future experimental trials (e.g., to optimize learning, more quickly converge on an estimate of participants' abilities or strategies, etc.).

Conclusions

The above findings show that automatic speech-to-text transcription, though imperfect, recovers many of the fundamental behavioral phenomena in free recall data. Our results provide a proof of concept that automatic speech-to-text transcription is sufficiently accurate to serve as an effective substitute for human annotators in list-learning experiments. Additional study is needed to understand how broadly the level of performance we observed might generalize to other verbal response experiments, noisy recording environments, etc. Nevertheless, as improved speech-to-text algorithms are discovered and developed, we expect this to alleviate the need for human annotators.

Acknowledgements We are grateful for useful discussions with Justin Hulbert and Talia Manning. Our work was supported in part by NSF EPSCoR Award Number 1632738. The content is solely the responsibility of the authors and does not necessarily represent the official views of our supporting organizations.

References

Angelakis, E., Stathopoulou, S., Frymiare, J. L., Green, D. L., Lubar, J. F., & Kounios, J. (2007). EEG neurofeedback: A brief overview and an example of peak alpha frequency training for cognitive enhancement in the elderly. *The Clinical Neuropsychologist*, *21*(1), 110–129.

Bamberg, P., Chow, Y.-L., Gillick, L., Roth, R., & Sturtevant, D. (1990). The Dragon continuous speech recognition system: a real-time implementation. In *Proceedings of DARPA Speech and Natural Language Workshop*, (pp. 78–81).

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data *Perspectives on Psychological Science*, *6*(1), 3–5.

Carlini, N., & Wagner, D. (2018). Audio adversarial examples: targeted attacks on speech-to-text. arXiv:1801.01944

Cohen, M. S. (2001). Real-time functional magnetic resonance imaging. *Methods*, *25*, 201–220.

Col, J. (2017). Enchanted learning. Retrieved from http://www.enchantedlearning.com

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *The American Journal of Psychology*, *75*(3), 485–491.

Cox, R. W., & Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magnetic Resonance in Medicine*, *42*, 1014–1018.

Cox, R. W., Jesmanowicz, A., & Hyde, J. S. (1995). Real-time functional magnetic resonance imaging. *Magnetic Resonance in Medicine*, *33*, 230–236.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One*, *8*(3), e57410.

deBettencourt, M. T., Cohen, J. D., Lee, R. F., Norman, K. A., & Turk-Browne, N. B. (2015). Closed-loop training of attention with real-time brain imaging. *Nature Neuroscience*, *18*(3), 470–475.

deCharms, R. C. (2008). Applications of real-time fMRI. *Nat Rev Neurosci*, *9*(9), 720–729.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., & Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., & Bäuml, M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, (pp. 2338–2342).

Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., & Manning, J. R. (2017). Quail: a Python toolbox for analyzing and plotting free recall data. The Journal of Open Source Software, https://doi.org/10.21105/joss.00424

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., & Rudnicky, A. I. (2006). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Vol. 1, pp. 185–188).

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory Cognition*, *24*, 103–109.

Kahana, M. J. (2012). *Foundations of human memory*. New York: Oxford University Press.

Kahana, M. J. (2017). Memory search. In Byrne, J. H. (Ed.) *Learning and memory: A comprehensive reference, second edition (pp. 181–200)*. Oxford: Academic Press.

Kurzweil, R., Richter, R., Kurzweil, R., & Schneider, M. L. (1990). *The age of intelligent machines*. Cambridge: MIT Press.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083

Manning, J. R., Norman, K. A., & Kahana, M. J. (2015). The role of context in episodic memory. In Gazzaniga, M. (Ed.) *The cognitive neurosciences, 5th edition*, (pp. 557–566). Cambridge: MIT Press.

Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, *108*(31), 12893–12897.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgement and Decision Making*, *5*(5), 411–419.

Park, M., & Pillow, J. W. (2012). Bayesian active learning with localized priors for fast receptive field characterization. In *Advances in Neural Information Processing Systems*, (pp. 2348–2356).

Polyn, S. M., & Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences*, *12*(1), 24–30.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*(1), 63–77.

Salzinger, K. (1959). Experimental manipulation of verbal behavior: A review. *The Journal of General Psychology*, *61*(1), 65–94.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *26*, 1589–1626. https://doi.org/10.1037/0278-7393.26.6.1589

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review*, *15*(3), 535–542.

UPenn Computational Memory Lab (2015). Penn TotalRecall. Computer Software.

van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Berlin: Springer.