Memory performance and text-to-speech functionality

A case for evidence-driven customization?

JENS F. BINDER, Nottingham Trent University, UK
RICHARD BOWEN, Nottingham Trent University, UK

CCS Concepts: • **Human-centered computing → Natural language interfaces**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics → User characteristics**.

Additional Key Words and Phrases: speech-based technology, interdisciplinarity, social science, humanities, conceptualization, methodology

1 WHAT DO WE KNOW ALREADY

We present a psychological perspective on memory performance and text-to-speech narration (TSN). Advances in TSN mean that by now speech can be generated at varying levels of naturalness (e.g., human-like or distorted), with different perceived demographic markers (e.g., signaling gender, age, level of education), and at different levels of familiarity (e.g., resembling own voice or modelled on the voice of known others). These advances mean that TSN is now rich in social signals, with potentially wide-ranging effects on human cognition that have rarely been considered to date. Here, we focus on assistive technologies and an approach that is intended to improve their customization.

TSN functionality is an integral part of assistive technologies, often in contexts where memory, learning and retention of information are of concern. This is, unsurprisingly, most prominent where the comprehension of textual material is to be improved, for learning disabilities [21] and visual impairment [16]. Studies have further been extended to less specific disabilities related to interaction and conversation, for example autism [4]. Assistive technologies have also, sometimes independent of TSN functionality, been used to target cognitive impairment and memory directly [6], e.g., for the purpose of rehabilitation. An area of development, therefore, where issues of memory and TSN become most closely intertwined concerns users who have ongoing and complex support needs in their daily lives due to cognitive impairment and would ideally benefit from continuous human-human interaction, such as users with dementia [13, 14, 17].

Memory has been shown to be dependent on social cues at various stages of the encoding and retrieval process, but research on memory performance to date offers little that allows for considering the specific effects of TSN. We therefore draw on psychological models that imply malleable and context-dependent memory performance, to outline further avenues for theory-informed research. Socially motivated information processing and a dynamic memory system are the two fundamentals that suggest that TSN, as it enables human-machine interaction to emulate more and more human-human interaction, can give rise to memory effects that are best documented in the extensive literature on social cognition, group dynamics and social identity.

First of all, humans show different levels of memory performance for stimuli perceived as animate and inanimate [18]. Animacy has been to shown to have an effect on language comprehension and the organization of knowledge [12]. This may be due to richer encoding of animate stimuli [10] and to increased allocation of attention in processing [8]. While the experimental work on animacy and adaptive memory has not considered speech characteristics to date, the theory implies, first of all, TSN effects that rest on the naturalness of language. This encompasses all aspects of TSN that help to create the illusion of natural speech: cadence and flow of speech and more specific prosodic aspects that affect intonation, stress, and rhythm. Such effects should be independent of the time that is needed to get used to

artificial-sounding speech. Animacy effects run more deeply and should persist even when users are given time to get acquainted with artificial-sounding speech.

The effects of naturalness are not necessarily novel in the domain of interactive technologies, and are captured by design principles going back as far as the uncanny valley hypothesis [11]. What is novel is that speech generation is now at a point where we can expect characteristics of this particular feature to have measurable effects on user cognition. What is more, we can now draw on theories of memory dynamics in a social context, assuming that speech characteristics can come with any number of social cues, most notably those that suggest membership in a particular social category.

Self-relevance and social categorization in terms of ingroup and outgroup membership are drivers for selective attention and selective memory. Some of these effects are due to ingroup favoritism and refer to, for example, better memory for positive behaviors observed in ingroup members and negative behaviors observed in outgroup members [7]. There is, however, evidence for a general increase in memory performance when the focus is on information relevant to the ingroup rather than the outgroup [3]. In team settings, memory performance has been shown to depend on the team composition (in terms of, e.g., social closeness, but also gender) [1, 2]. These effects are explained by an adaptive process. Individuals adjust their encoding of information, and the organization of knowledge in memory, according to their assumptions about others' knowledge structure [9, 19, 20].

2 HOW WE STUDY THE PHENOMENON AND WHAT WE WOULD LIKE TO KNOW

Research on memory has used a range of tasks in laboratory-based experimentation, and these methods are easily adapted for studying effects of TSN. Our current study, as work in progress, focuses on the factor of naturalness of voice in TSN and its effects on short-term memory. Adapting a standard list learning procedure, participants are presented with four lists of short phrases, each with a different narrator: high and low cadence text-to speech narrators as well as high and low cadence human narrators. Narrators read out the items on these lists, which consist of short three-word phrases containing a noun and a verb i.e. 'warn of politics'. Each list is presented twice, with either a male or female voice, after which the participant is asked to verbally recall any phrases or words they can remember. For TSN phrases, easily accessible standard software is used: the Microsoft Windows Narrator App [5] and Amazon Polly [15]. Responses are scored for their accuracy to the original lists, with full phrases and single words being scored separately. Accuracy scores are then used to calculate to what extent the type of narrator predicts performance in recall.

The experimental set-up allows to test the general hypothesis that short-term recall improves to the extent that narrator's voice becomes more similar to that of a normal human voice. At the same time, the set-up inevitably leaves gaps in our understanding. Among the more immediate lingering points are issues surrounding the type of memory under investigation (e.g., long-term, short-term, episodic, semantic), the specific nature of the stimulus materials (e.g., naturalistic, standardized, rich in content, and so forth), and the definition of what counts as more or less naturalistic voice.

3 WHAT WE WANT TO LEARN FROM OTHER DISCIPLINES AND WHAT WE WANT TO TEACH THEM

The shortcomings of focused experiments, such as the comparatively slow and incremental progress in understanding, is likely to be off-set, we argue, by the promise of arriving at robust recommendations for the customization of TSN, and potentially other technologies that feature voice and speech generation.

Next to a compelling experimental design, however, the overall outcome of such research depends on the quality and relevance of the materials and stimuli used. Other disciplines can provide crucial input for this: current standards in

speech generation, the adaptability of generated speech to that of specific individuals, relevant speech characteristics that lend themselves to systematic variation in experiments, and systems of customization as they can present themselves to end users are all aspects of such input. In particular, we feel the need to go beyond freely available software solutions as they are used here and as they are frequently used in psychological research. Further, next to isolated laboratory-based research, more interdisciplinary studies on implementation and evaluation of TSN technologies in the field are also much needed opportunities that will help to identify other relevant factors to take back into controlled experimentation.

An experimental approach is not unique to psychology and can be found in many disciplines with an interest in behavioral indicators. A more unique aspect of the psychological perspective promoted here consists in the theoretical underpinnings. Theories and models of memory performance have been shaped in close co-evolution with experimental methods, but they also provide frames for orientation in applied research. Put differently, recommendations for customization will be most enduring when they are backed by evidence and a compelling theoretical explanation. Recommendations could focus on the appropriate level of liveliness, on the similarity or dissimilarity with others close to the user, on auditory ingroup and outgroup markers etc. Empirical evidence can tell us, for example, which of the following notions is more likely to be the case: that memory performance is better when speech is modelled on a close other (due to familiarity, levels of attention and motivation, activation of existing memories in the system, social compliance, and so forth), or that memory is better when speech is modelled on an unknown other (due to novelty, the need to focus more for understanding, the deeper level of elaboration, conversational norms, and so forth). Our approach can help to further specify and refine such notions, thereby turning them into more robust design recommendations. This is where psychology is likely to be of wider value to other disciplines concerned with TSN and memory.

As technologies enable exchanges that increasingly resemble human-human interaction the boundary between human-machine and human-human settings will become increasingly blurred, at least in the eyes of users. Theories and findings regarding users' mental models of human interaction partners will therefore become more important. The social aspects of memory performance provide a first testing ground for these predictions.

REFERENCES

[1] J. Andersson. 2001. Net effect of memory collaboration: how is collaboration affected by factors such as friendship, gender and age? *Scand J Psychol.* 42, 4 (2001), 367–75. https://doi.org/1467-9450.00248

[2] J. Andersson and J. Ronnberg. 1997. Cued Memory Collaboration: Effects of Friendship and Type of Retrieval Cue. *European Journal of Cognitive Psychology* 9, 3 (1997), 273–287. https://doi.org/10.1080/713752558

[3] J. J. Van Bavel and W. A. Cunningham. 2012. A Social Identity Approach to Person Memory: Group Membership, Collective Identification, and Social Role Shape Attention and Memory. *Personality and Social Psychology Bulletin* 38, 12 (2012), 1566–1578. https://doi.org/0146167212455829

[4] SA Cassidy, B Stenger, L Van Dongen, K Yanagisawa, R Anderson, V Wan, S Baron-Cohen, and Cipolla R. 2016. Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions. *Comput Vis Image Underst* 148 (2016), 193–200.

[5] Microsoft Corporation. 2020. Complete guide to Narrator. Retrieved July 16, 2021 from https://support.microsoft.com/en-us/windows/complete-guide-to-narrator-e4397a0d-ef4f-b386-d8ae-c172f109bdb1

[6] B.-K. Dewar, M. Kopelman, N. Kapur, and B. A. Wilson. 2014. Assistive technology for memory. In *Assistive technology for cognition: A handbook for clinicians and developers*, B. O'Neill and A. Gillespie (Eds.). Routledge, New York, 31–46.

[7] J. W. Howard and M. Rothbart. 1980. Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology* 38, 2 (1980), 301–310.

[8] JK Leding. 2019. Adaptive memory: Animacy, threat, and attention in free recall. *Mem Cognit* 47, 3 (2019), 383–394. https://doi.org/10.3758/s13421-018-0873-x

[9] S. B. Marion and C. Thorley. 2016. A meta-analytic review of collaborative inhibition and postcollaborative memory: Testing the predictions of the retrieval strategy disruption hypothesis. *European Journal of Cognitive Psychology* 142, 11 (2016), 1141–1164. https://doi.org/10.1037/bul0000071

[10] M. J Meinhardt, R. Bell, A. Buchner, and J. P. Röer. 2020. Adaptive memory: Is the animacy effect on memory due to richness of encoding? *J Exp Psychol Learn Mem Cogn* 46, 3 (2020), 416–426. https://doi.org/10.1037/xlm0000733

[11] M. Mori. 1970. Bukimi no tani ["The uncanny valley"]. *Energy* 7, 4 (1970), 33–35.

[12] J. S. Nairne, J. E. VanArsdall, and M. Cogdill. 2017. Remembering the Living: Episodic Memory Is Tuned to Animacy. *Current Directions in Psychological Science* 26, 1 (2017), 22–27. https://doi.org/10.1177/0963721416667711

[13] S. Nakatani, S. Saiki, M. Nakamura, and K. Yasuda. 2018. Generating personalized virtual agent in speech dialogue system for people with dementia. In *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Springer, Cham, 326–337.

[14] A. Russo, G. D'Onofrio, A. Gangemi, F. Giuliani, M. Mongiovi, F. Ricciardi, F. Greco, F. Cavallo, P. Dario, D. Sancarlo, V. Presutti, and A. Greco. 2019. Dialogue Systems and Conversational Agents for Patients with Dementia: The Human-Robot Interaction. *Rejuvenation Research* 22, 2 (2019), 109–120. https://doi.org/10.1089/rej.2018.2075

[15] Amazon Web Services. 2021. Amazon Polly. Retrieved July 16, 2021 from https://aws.amazon.com/polly/

[16] R. F. Sharma and S. G. Wasson. 2012. Speech recognition and synthesis tool: assistive technology for physically disabled persons. *International Journal of Computer Science and Telecommunications* 3, 4 (2012), 86–91.

[17] F. Sposaro, J. Danielson, and G Tyson. 2010. iWander: An Android application for dementia patients. *Annu Int Conf IEEE Eng Med Biol Soc* 2010 (2010), 3875–8. https://doi.org/10.1109/IEMBS.2010.5627669

[18] J. E. VanArsdall, J. S. Nairne, J. N. S. Pandeirada, and J. R. Blunt. 2013. Adaptive memory: animacy processing produces mnemonic advantages. *Exp Psychol.* 60, 3 (2013), 172–8. https://doi.org/10.1027/1618-3169/a000186

[19] D. M. Wegner. 1987. *Transactive memory: A contemporary analysis of the group mind. In Theories of group behavior.* Springer, New York, NY.

[20] D. M. Wegner, R. Erber, and P. Raymond. 1991. Transactive memory in close relationships. *Journal of Personality and Social Psychology* 61, 6 (1991), 923–929. https://doi.org/doi/10.1037/0022-3514.61.6.923

[21] S. G. Wood, J. H. Moxley, E. L. Tighe, and R. K. Wagner. 2018. Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of Learning Disabilities* 51, 1 (2018), 73–84. https://doi.org/10.1177/0022219416688170