Beyond Responding Fast or Slow: Improving Cognitive Models of Memory Retrieval using Prosodic Speech Features

Thomas Wilschut (t.j.wilschut@rug.nl) Department of Experimental Psychology, University of Groningen

Florian Sense (florian.sense@infinitetactics.com) InfiniteTactics, LLC

Odette Scharenborg (o.e.scharenborg@tudelft.nl) Multimedia Computing Group, Delft University of Technology

Hedderik van Rijn (d.h.van.rijn@rug.nl) Department of Experimental Psychology, University of Groningen

Keywords: Adaptive Learning, ACT-R, Cognitive Modeling, Speech Features, Prosody

Background

Technology plays an increasingly important role in education. Digital adaptive learning (AL) systems have successfully improved the efficiency of fact and word learning by tailoring learning procedures to the needs of individual learners (Lindsey, Shroyer, Pashler, & Mozer, 2014; Papousek, Pelánek, & Stanislav, 2014; Van Rijn, Van Maanen, & Van Woudenberg, 2009). AL systems typically track learning performance (measured using typed responses to practice problems) in real time and use this information to provide personalized feedback, select appropriate practice materials or optimize item repetition schedules. The effectiveness of AL systems critically depends on their ability to estimate the extent to which learners have successfully memorized study materials. Some AL systems employ a cognitive model of memory retrieval to estimate the strength of item representations in the learners' memory. For example, the SlimStampen system (Van Rijn et al., 2009) is based on the ACT-R architecture's model of human declarative memory (Anderson, Bothell, Lebiere, & Matessa, 1998) and functions by measuring response times (RTs) and accuracy scores to determine optimal item repetition schedules. The system relies on the assumption that RTs are a good proxy for the strength of fact representations in memory: The quicker the learner produces a correct response, the stronger the memory representation for that item is assumed to be (Anderson & Schooler, 1991; Jescheniak & Levelt, 1994; Levelt, 1999; Van Rijn et al., 2009) The above-described approach uses the limited information available (RTs and accuracy) to estimate memory strength for typed retrieval attempts and uses this information to optimize item repetition schedules.

Recent advances in speech technology have allowed for the transition from typing-based AL systems to speech-based AL systems (Wilschut et al., 2021). In speech-based learning, there is additional information that AL models can use to estimate the strength of item representations in memory: Spoken language contains prosodic speech features (PSFs), which are supra-segmental properties of speech (Xu, 2011). PSFs are commonly used by speakers to convey information beyond the literal meaning of the utterance, and can be roughly divided into three categories: Intonation, the melodic pattern of an utterance, defined by the dynamics in pitch over the duration of a speech segment; rhythm, the dynamics in timing and speaking speed of a speech segment; and stress, which refers to the intensity that is given to a syllable of speech, resulting in changes in relative loudness.

Here, we aim to examine if spoken retrieval attempts contain information that goes beyond what is already encapsulated in the RT and accuracy scores for that retrieval attempt. We hypothesize (1) that prosodic speech features are associated with retrieval accuracy and (2) that PSFs carry information that can be used - in addition to RTs and accuracy scores - to more accurately estimate the extent to which a learner has successfully memorized an item, and predict later retrieval success.

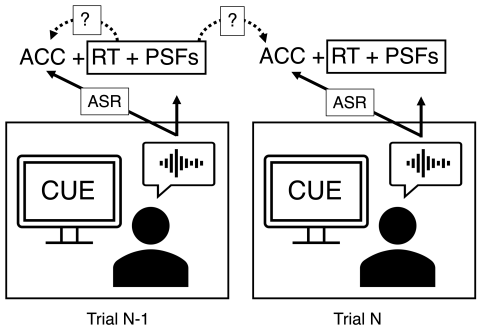

Figure 1: Design and research questions. Participants saw a cue (see Methods) and responded using speech. Using ASR, the accuracy of the response is determined. The first research question examines if PSFs derived from the speech signal are associated to accuracy (ACC) on the same repetition (left question mark). The second research question considers if previous-repetition PSFs can be used to explain current-repetition accuracy (right question mark).

Methods

A graphical description of the design and research questions is shown in Figure 1. Fifty participants studied Swahili-English vocabulary items using the SlimStampen adaptive scheduling system. Swahili items were presented on a computer screen, and participants were asked to respond by pronouncing the English translation of the item. Participants'

utterances were transcribed to text in real time using Google Cloud Speech-to-Text (see https://cloud.google.com/speech-to-text) automatic speech recognition (ASR). Speech features were extracted afterwards using Praat 6.2.07 (Boersma, 2006).

Results

The results of this study are twofold. The first part concerns the relationship between speech features and retrieval accuracy. As hypothesized, the accuracy of retrieval attempts was associated with specific speech feature characteristics. More specifically, higher retrieval accuracy was associated with falling pitch (negative pitch slope), higher loudness and higher speaking speed ($r(7847)$ = -0.10, $p < .001$; $r(7847)$ = 0.05, $p < .001$; $r(7847) = 0.07$, $p < .001$, respectively).
Second, we explored the possibility of using PSFs to explain next-repetition retrieval accuracy. We conducted a logistic mixed-effects regression model to explain current-repetition accuracy using (1) model-based activation estimations, (2) previous-repetition pitch slope, (3) previous-repetition speaking speed and (4) previous-repetition loudness, see Table 1. As expected, model-based activation, estimated using past-repetition RTs and accuracy scores, significantly explained accuracy (z = 4.60, $p < .001$, see Figure 2A and Table 1). Importantly, previous-repetition pitch slope and previous-repetition speaking speed also explained variance in current-repetition retrieval accuracy (z = -2.60, p = .008; z = 3.37, $p <$ .001, see Table 1 and Figure 2B and Figure 2C, respectively). Loudness did not significantly explain next-trial accuracy (z = 0.82, p = .412, see Table 1). Adding the previous-repetition PSFs to a model with only model-estimated activation as independent variable resulted in an 16% increase in explained variance in current-repetition retrieval accuracy (R^2 = 0.162 and R^2 = 0.135 for the model with and without PSFs, respectively). Together, our results show that we can use pitch dynamics and speaking speed in addition to RTs and accuracy scores to improve explanations of next-trial retrieval accuracy.

Table 1: Logistic mixed-effects regression model explaining current trial accuracy from model-estimated activation and previous-trial PSFs.

	β	*SE*	*z*	*p*
Intercept	1.63	0.12	13.56	<0.001
Activation_n	−0.45	0.10	4.60	<0.001
Pitch slope_{n-1}	−0.15	0.06	−2.66	0.008
Loudness_{n-1}	0.04	0.05	0.82	0.412
$\text{Speaking speed}_{n-1}$	0.18	0.05	3.37	<0.001

Conclusion

We show that spoken retrieval attempts contain information about the extent to which a learner has memorized an item, and that PSFs can be used to improve model predictions for learner performance on future trials. Our results are important in two ways. First, they have theoretical implications, as they elucidate how speaker accuracy is reflected in speech prosody: to our knowledge, we are the first to demonstrate that inaccurate and slow responses are associated with a rising pitch, low vocal loudness and low speaking speed, suggesting that PSFs can be used as a measure of speaker certainty or confidence. More generally, PSFs may prove to be a valuable new tool in the further exploration of important open research questions (e.g., about speaker certainty/confidence or feeling-of-knowing and a range of other meta-memory judgements). Second, our results have practical implications, as they can contribute to the further development of speech-based AL systems. We show that PSFs can be used to improve AL model accuracy predictions. Importantly, compared to more traditional (deep-learning-based) approaches to automatic speech processing, extracting PSFs from the speech signal is computationally inexpensive, making them especially suitable to be used in real-time AL applications. In short, we show that PSFs are a promising candidate to be used in educationally relevant speech-based learning applications.

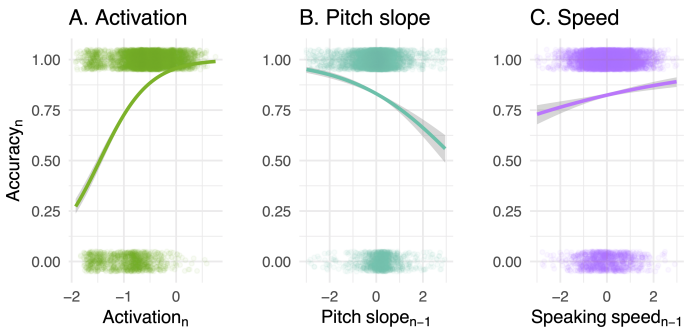

Figure 2: Explaining retrieval accuracy on the current trial (n) using (A) memory activation estimated by a response time-based ACT-R model of memory retrieval, and (B-C) high level PSFs on the previous retrieval attempt for the same item (n-1). Dots show empirical accuracy, lines show accuracy estimations.

References

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Boersma, P. (2006). Praat: doing phonetics by computer. *http://www. praat. org/*.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4),

824.

Levelt, W. J. (1999). Models of word production. *Trends in cognitive sciences*, *3*(6), 223–232.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science*, *25*(3), 639–647.

Papousek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In *Educational data mining 2014* (pp. 6–13).

Van Rijn, H., Van Maanen, L., & Van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th international conference of cognitive modeling* (Vol. 2, pp. 7–6).

Wilschut, T., Sense, F., Van der Velde, M., Fountas, Z., Maass, S., & Van Rijn, H. (2021). Benefits of adaptive learning transfer from typing-based learning to speech-based learning. *Frontiers in AI and Big Data*, *4*.

Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences*, *1*(1), 85–115.