

## Improving Adaptive Learning Models Using Prosodic Speech Features

Wilschut, Thomas; Sense, Florian; Scharenborg, Odette; van Rijn, Hedderik

**DOI**

[10.1007/978-3-031-36272-9\\_21](https://doi.org/10.1007/978-3-031-36272-9_21)

**Publication date**

2023

**Document Version**

Final published version

**Published in**

Artificial Intelligence in Education - 24th International Conference, AIED 2023, Proceedings

**Citation (APA)**

Wilschut, T., Sense, F., Scharenborg, O., & van Rijn, H. (2023). Improving Adaptive Learning Models Using Prosodic Speech Features. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial Intelligence in Education - 24th International Conference, AIED 2023, Proceedings* (pp. 255-266). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13916 LNAI). Springer. [https://doi.org/10.1007/978-3-031-36272-9\\_21](https://doi.org/10.1007/978-3-031-36272-9_21)

**Important note**

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

***Green Open Access added to TU Delft Institutional Repository***





***'You share, we take care!' - Taverne project***

**<https://www.openaccess.nl/en/you-share-we-take-care>**

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



# Improving Adaptive Learning Models Using Prosodic Speech Features

Thomas Wilschut<sup>1</sup> , Florian Sense<sup>2</sup> , Odette Scharenborg<sup>3</sup> ,  
and Hedderik van Rijn<sup>1</sup> 

<sup>1</sup> Department of Experimental Psychology, University of Groningen,  
Groningen, The Netherlands

{t.j.wilschut,d.h.van.rijn}@rug.nl

<sup>2</sup> InfiniteTactics, LLC, Beaver Creek, USA

florian.sense@infinite-tactics.com

<sup>3</sup> Department of Multimedia and Computing, Delft University of Technology,  
Delft, The Netherlands

o.e.scharenborg@tudelft.nl

**Abstract.** Cognitive models of memory retrieval aim to describe human learning and forgetting over time. Such models have been successfully applied in digital systems that aid in memorizing information by adapting to the needs of individual learners. The memory models used in these systems typically measure the accuracy and latency of *typed* retrieval attempts. However, recent advances in speech technology have led to the development of learning systems that allow for spoken inputs. Here, we explore the possibility of improving a cognitive model of memory retrieval by using information present in speech signals during spoken retrieval attempts. We asked 44 participants to study vocabulary items by spoken rehearsal, and automatically extracted high-level prosodic speech features—patterns of stress and intonation—such as pitch dynamics, speaking speed and intensity from over 7,000 utterances. We demonstrate that some prosodic speech features are associated with accuracy and response latency for retrieval attempts, and that speech feature informed memory models make better predictions of future performance relative to models that only use accuracy and response latency. Our results have theoretical relevance, as they show how memory strength is reflected in a specific speech signature. They also have important practical implications as they contribute to the development of memory models for spoken retrieval that have numerous real-world applications.

**Keywords:** Adaptive Learning · Cognitive Modeling · Automatic Speech Recognition · Machine learning · Speech prosody · Pitch · Speaking Speed · Intensity

## 1 Introduction

Model-based adaptive learning systems optimize learning by tailoring learning procedures to the needs of the individual learner [11, 14, 22]. To this end, such systems aim to estimate and predict the extent to which a learner has successfully memorized information. As memory strength cannot be observed directly, models of memory retrieval use

behavioral proxies, such as accuracy scores and response latencies to make informed predictions [1, 15, 26]. Here, we will build upon a number of recent studies showing that human speech contains information on a speaker's emotional state, confidence, and the accuracy of the spoken response [7, 10]. The aim of this study is to explore the theoretical and practical feasibility of using the information present in spoken retrieval attempts to improve cognitive models of memory retrieval that can be applied in adaptive learning systems. To that end, we will here examine (a) whether the results of earlier studies, which identified a specific prosodic speech signature associated with accuracy and speaker confidence, generalize to a learning paradigm specifically, and (b) whether the information present in spoken retrieval attempts can be used to improve predictions of memory retrieval success in an applied learning setting.

As the extent to which a learner has successfully memorized an item is a latent state that cannot be measured directly, cognitive models of learning and forgetting use behavioral proxies. Response accuracy is a logical candidate, as it indicates whether or not the learner could successfully retrieve the memorandum. Correspondingly, accuracy is used in many models that predict performance [15, 23]. However, using (only) accuracy as a behavioral index of memory strength results in a number of issues. First, accuracy-based models of forgetting have difficulties accounting for the passage of time between events (for example, in early Bayesian knowledge tracing models, information was never forgotten once an item flipped to the “known” state after an accurate response [13]). Second, using accuracy as a proxy of memory strength does not allow for meaningful discrimination within correct responses, and as a consequence, accurate performance predictions require many incorrect responses [19]. Because of these limitations, some models use the latency of a response in addition to its accuracy to predict performance. A core assumption these models rely on is the link between response latency and memory strength [4]. An abundance of experimental data supports this link: Faster responses are generally associated with more accurate responses and a stronger association between cue and response compared to slower responses [21]. Using response latency in addition to accuracy to predict learner performance has been proven to be successful in a range of adaptive learning applications [11, 15, 19, 21].

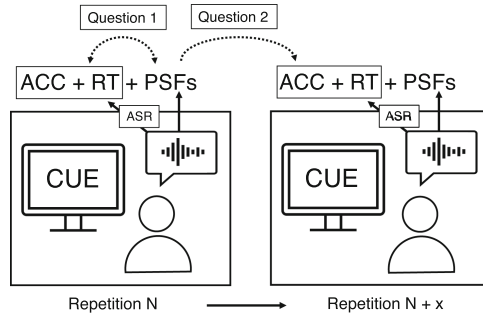
An example of a successful approach to modeling forgetting using accuracy and response latency is the MemoryLab adaptive scheduling system [www.memorylab.nl/en](http://www.memorylab.nl/en). The system is based on the ACT-R theory's declarative memory framework [1]. In this cognitive model, individual facts are represented as a memory chunk that has a certain activation, which corresponds to the fact's strength in memory. It gets boosted every time a fact is rehearsed and gradually decays over time if an item is not repeated. As some facts are more difficult to learn than other facts, and as some learners forget facts at a different rate than other learners, all facts are given a unique decay rate parameter for all learners, which allows the model to vary the rate at which items are forgotten between learners and between facts. The model relies on the above-mentioned link between retrieval speed and the memory strength—or activation—of that fact. Combining the link between response latency and memory activation with its activation decay functions, the model can calculate expected response times for retrieval attempts at any future point in time. Observed discrepancies between expected and observed response times and accuracy scores can consequently be used to update decay parameters, resulting in a system that can accurately capture and predict forgetting over time [20, 24].

Accuracy and latency comprise only a part of all potential sources of information that may be used to inform models that aim to predict forgetting. In this study, we will focus on spoken language, which contains prosodic speech features (PSFs). PSFs are high-level properties of units of speech longer than individual phonemic segments, such as syllables, words or sentences [17]. PSFs can be roughly divided into three categories. First, **intonation** is the melodic pattern of an utterance and refers to the dynamics in pitch over the duration of a speech segment. Second, **rhythm** is defined by dynamics in timing, or speaking speed, over the duration of the speech segment. Third, **stress** refers to intensity (loudness) given to a syllable of speech, resulting in changes in relative intensity. Prosodic information usually reflects information that is not necessarily present in grammar or choice of vocabulary, such as the emotional state of the speaker, emphasis, or the form of utterance (e.g., question versus statement versus command) [27, 29].

Of particular importance for the current study is the idea that speakers implicitly convey their levels of certainty about a response through speech prosody. For example, in many languages, speakers end a sentence or word with raising pitch and lower intensity when asking a question or to express uncertainty about a response [10]. Listeners are able to decode these prosodic speech signatures across languages and cultures [7]. Next to the idea that speech prosody conveys subjective speaker confidence, a recent study by Goupil and Aucouturier demonstrated that objective accuracy is distinctly reflected in the speech signal [6]. In their study, participants were instructed to complete a visual detection task: A word was briefly presented on the screen, followed by a visual mask. After a short interval, participants needed to verbally choose which word they saw from a number of alternatives and rate their confidence in the response. The results showed that the participants' speaking speed and intonation were associated with the objective accuracy of the response, and the intensity of the response was associated with subjective confidence. To date, it is unclear if the above mentioned effects generalize to the context of a learning paradigm.

Overall, earlier findings suggest that a speech signal for spoken retrieval attempts may contain information that can be used to improve models of memory retrieval in two ways. On one hand, as they are associated to objective accuracy, they may directly carry information about the latent memory strength of a response. At the same time, they may carry information about memory strength indirectly, as they are associated to subjective confidence in the response. In the current study, we aim to explore the hypotheses that these PSFs are indeed informative in a learning context and that they can be used to improve models of memory retrieval.

To pursue these research questions, we will rely on recent advancements in speech technology which have led to the development of adaptive learning systems that allow for spoken input. Such systems allow users to efficiently practice pronunciation, which is an important part of language acquisition that is largely omitted in traditional approaches [5]. Furthermore, they can be applied in situations where users do not have the ability (for example, young children) or opportunity (for example, while driving a car) to type. Finally, spoken learning allows for the extraction of PSFs in the spoken utterances. Compared to more traditional approaches to automatic speech processing, that often rely on deep learning-based classification of spectral components of the speech signal [12], extracting PSFs from the speech signal is computationally relatively



**Fig. 1.** Design and research questions. Participants saw a cue and responded using speech. Using automatic speech recognition (ASR), the accuracy (ACC) and response time (RT) of the response is determined. The first research question examines if PSFs derived from the speech signal are associated to accuracy and response time on the same trial. The second research question considers if current repetition (N) PSFs can be used to improve predictions for future repetition (N + X) accuracy and response time for the same item.

inexpensive [25]. In short, advances in speech technology have made the implementation of speech-based learning systems, in which PSFs are automatically extracted, practically feasible.

Given the growing popularity and practical feasibility of speech-controlled learning applications, we here explore the possibility of exploiting information present in the speech signal to improve cognitive models of memory retrieval used in personalized learning applications. To that end, we first aim to extend earlier studies that find a specific PSF signature associated with accuracy and with subjective confidence to the context of a memory retrieval paradigm specifically (see Fig. 1, Question 1). To foreshadow the results, we indeed found an association between PSFs and accuracy and response latency for spoken retrieval attempts. Second, in this study, we are the first to explore the possibility of improving predictions of future retrieval performance using a model that incorporates PSFs for previous attempts compared to a model uses accuracy and response latency only (see Fig. 1, Question 2). To foreshadow the results, we found that using previous repetition PSFs *in addition to accuracy and response latency* substantially improves predictions of future retrieval performance.

## 2 Methods

The analyses reported in this study are based on data from the experiment reported in [28], which demonstrated that latency-based adaptive learning algorithms can improve learning efficiency in speech-based learning systems. That study used four within-subjects conditions: two learning modality conditions (typing-based and speaking-based learning) and two item scheduling conditions (fully adaptive MemoryLab learning, based on the learners' response times and accuracy scores, and less adaptive flashcard-inspired learning, using the learners' accuracy only). All analyses reported here are based on the speech conditions of the experiment only. For the first research question (exploring the extent to which different PSFs can be used to explain performance on the same trial) data from both scheduling conditions was included. The

second research question explored whether using PSFs can result in increased performance predictions compared to adaptive scheduling systems that used response times and accuracy scores only. Therefore, only data from the MemoryLab adaptive scheduling condition was included. We will briefly reiterate the relevant details of the study here, but for more details, see [28]. Materials, analysis scripts, data and code to recreate the current experiment and analyses are available on <https://osf.io/dfexp/>.

## 2.1 Participants

In total, data from 44 participants was available for analysis. This sample size was chosen based on previous studies [28]. Participants were first-year Psychology students that received course credit for participants. Participants were 17–29 years of age, and 73% female. Participants were native Dutch speakers and indicated that they were fluent in English. Participants gave informed consent and the study was approved by the ethical committee of the department of Psychology at the University of Groningen (study approval code: PSY-2021-S-0025).

## 2.2 Design and Procedure

Participants were asked to study the English translation of Swahili vocabulary items. For the first presentation of an item, participants saw a Swahili word on the computer screen in text, together with the written English translation of this word. Additionally, the spoken English translation of the word was presented through headphones. In all subsequent presentations, only the Swahili word was shown, and participants were instructed to speak the correct English translation, after which they received corrective written and auditory feedback. Voice utterances were transcribed to text automatically and in real time using the Google Web Speech API to provide corrective feedback. To prevent that minor transcription, tense or number errors would result in scoring the response as incorrect, responses were considered correct if Levenshtein's edit distance from response to answer [30] was equal to or less than 2. Response times were defined as the time elapsed between the start of the presentation of the item and the time at which the participant started speaking (voice onset time). In the MemoryLab adaptive scheduling conditions, items were scheduled based on the accuracy and latency of previous responses (see Introduction). For a more detailed description of the model, see [18, 23].

## 2.3 Materials

For details on the materials used (word lists, exemplar pronunciations) as well as the setup (software and hardware) used in the experiment, see [28].

## 2.4 Speech Feature Extraction

PSFs were extracted after data collection using Praat 6.2.07 [3]. Before extracting PSFs, all silences were trimmed based on zero-crossings and short term intensity. We selected PSFs to include in the analyses based on previous literature (pitch slope, intensity,

speaking speed) and on an exploratory basis (average pitch, jitter, shimmer). Subsequently, we extracted the following PSFs for each utterance: (1) the fundamental frequency for each 0.01 s window in the speech signal; (2) the average frequency over the full duration of the speech signal; (3) the average intensity over the duration over the speech signal; (4) the speaking speed defined as the average number of syllables uttered per second; (5) mean local jitter and (6) mean local shimmer. Finally, (7) we used least squares linear regression on all pitch observations in an utterance to compute the pitch slope. All PSFs were standardized within participants.

## 2.5 Data and Statistical Analyses

Statistical analyses were conducted in R 3.4.1 [16]. Trials containing PSFs that were more than 5 standard deviations above or below the mean value, as well as trials with response latencies below 200ms, were considered outliers and were removed from the data set before analyses. In total, 1.02% of all trials were excluded based on these criteria. After preprocessing, the data set contained a total of 7,334 learning trials from 44 participants.

To examine the association between PSFs and response accuracy and latency on the same trial, we computed Pearson's correlations. In addition, we fitted four mixed effects regression models to explain accuracy (logistic mixed effects regression models) and response latency (linear mixed effects regression models) using PSFs. In these models, standardized PSFs were added as fixed effects, and we controlled for by-item and by-participant variation by adding these factors as random effects to the models [2].

To explore if PSFs can be used to improve the adaptive learning model predictions, we trained and tested two variations of two regression models to predict (a) current trial response latency and (b) accuracy. Because regression analyses with a large number of predictor variables have a tendency to over-fit the data (reducing out-of-sample predictive properties) we used elastic-net penalized regression [31]. We used 10-fold cross validation to choose the optimal elastic net hyperparameters for mixture (alpha) and penalty (lambda). Both sets of models contained predictions based on the memory activation estimated during learning by the MemoryLab model as predictor. The models differed by their inclusion of the PSFs: only one set of models contained PSFs on previous trials. We included PSFs for up to five repetitions back for the same item. All models were trained and tested using a second 10-fold cross validation procedure, in which the models were trained on 90% of the data and tested on the remaining 10% of data 10 times. Because of the unbalanced nature of the sample (more correct than incorrect responses), we report average test classification precision, recall and F1-score metrics for the models that predict accuracy [8]. We report average test R-squared values, as well as root mean square error (RMSE) values, for the models predicting response latencies.

## 3 Results

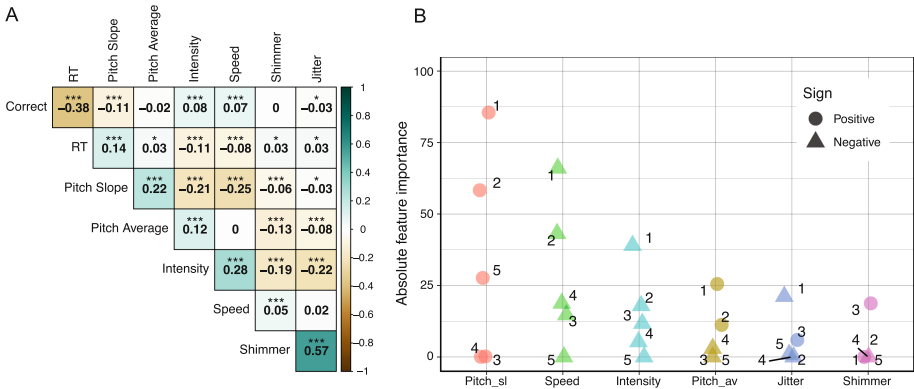
### 3.1 The Association Between Speech Prosody and Memory Retrieval Performance

The first main goal of this study is to test the reliability of earlier studies that find an association between some PSFs and speaker confidence and accuracy [6, 7, 9], and to



examine the exact way in which these effects can be found in a learning paradigm. Figure 2A shows Pearson’s correlations between accuracy scores, response times, and PSFs for spoken retrieval attempts during learning. The top two rows of Fig. 2A show that there are several PSFs that correlate significantly with accuracy and response latencies. First, in line with earlier literature [6], we find that standardized pitch slope positively correlates with response times, and negatively correlates with correctness, indicating that longer response times and lower accuracy are associated with rising pitch. Average standardized intensity is negatively associated with response times and positively associated with correctness, indicating more accurate and faster responses are, on average, louder. Finally, speaking speed is negatively associated to response times and positively associated to correctness indicating that faster and more accurate responses are associated with higher speaking speed.

To further corroborate the correlational analyses reported above, we fitted four mixed effects regression models. Table 1.1 shows the results of the model predicting current trial accuracy from standardized pitch slope, speaking speed, intensity, jitter, shimmer, and average pitch<sup>1</sup>. The results show that pitch slope, speaking speed, and average intensity significantly explain accuracy: The lower the pitch slope, the higher the speaking speed and the higher the intensity, the higher the accuracy of the same response. Shimmer, jitter, and average pitch do not explain accuracy. When response times are added to the model (Table 1.2), pitch slope and intensity still significantly explain accuracy.



**Fig. 2. A.** Pearson’s correlations between accuracy, response latencies, and PSFs for the same spoken retrieval attempt. Note: \*p<.05; \*\*p<.01; \*\*\*p<.001. **B.** Absolute PSF importance (penalized regression coefficient (ms)) in predicting current repetition recall performance (response latency) from previous repetition PSFs. Black numbers indicate how many trials back the PSF was recorded. Circles show positive regression coefficients, triangles show negative coefficients.

<sup>1</sup> The logistic regression coefficients in Table 1.1 and 1.2 can be converted to probabilities using an inverse logit transform. For example, in Table 1.1, a one standard deviation increase in pitch slope was associated with a decrease in accuracy from  $e^{(1.885)} / (1 + e^{(1.885)}) = 0.868$  to  $e^{(1.885-0.260)} / (1 + e^{(1.885-0.260)}) = 0.835$ .

**Table 1.** The association between PSFs and retrieval performance.

	Accuracy		Response latency (ms)	
	(1. without RT)	(2. with RT)	(3. without accuracy)	(4. with accuracy)
Response latency (ms)		−0.001***		
Accuracy (cor./incor.)				−1541.371***
Pitch slope (z)	−0.260***	−0.196***	148.827***	99.541***
Speaking speed (z)	0.099***	0.058	−90.516***	−70.819***
Average intensity (z)	0.120**	0.084*	−62.560**	−39.173*
Jitter (z)	−0.049	−0.065	16.411	4.902
Shimmer (z)	0.050	0.087	38.286	46.826*
Average pitch (z)	−0.030	−0.028	20.423	13.315
Constant	1.885***	3.226***	2264.926***	3569.319***
Observations	7,334	7,334	7,334	7,334

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$

Table 1.3 shows that pitch slope, speaking speed and intensity explain variation in response times (also see Fig. 2B): The higher the pitch slope, the higher the response times. The higher the speaking speed, the lower the response time. The louder the learner gave the answer, the higher correctness. Finally, the higher the mean local shimmer, the longer the response latency. Table 1.4 shows the model in which accuracy is also included. In this case, pitch slope, speaking speed and intensity remain significantly associated with response latency.

Overall, the mixed effects models corroborate what is apparent in the correlational analyses reported above: Three features show the strongest association to memory strength or ease-of-retrieval. First, rising pitch (positive pitch slope) utterances are associated to low accuracy and long response latencies. Second, high speaking speed is associated with high accuracy and low response times. Third, we found that loud responses (high intensity) are associated to high accuracy and low response times. These results indicate that in the context of a learning task, PSFs are informative of the extent to which a learner has successfully memorized an item.

### 3.2 Improving Predictions of Future Performance Using Speech Prosody

The results discussed in the previous section demonstrate that a number of PSFs recorded during spoken retrieval attempts are associated with retrieval accuracy and response times, and therefore carry information about the extent to which the learner has successfully memorized an item. Given the reliability of these effects in the context of a learning paradigm, we will now test whether they can be used to improve out-of-sample predictions of future retrieval success in applied learning settings. We used penalized regression analyses to predict retrieval performance (accuracy and response times) from speech prosody during earlier retrieval attempts for the same item.

Table 2 shows the average results of four penalized regression models, two predicting current trial accuracy and two predicting current trial response times. For each outcome variable, we fitted one model with the original MemoryLab algorithm's esti-

mations as predictor (right columns, ‘without PSFs’) and one model that contained both the original MemoryLab model estimations, as well as the six PSF recordings for the five preceding repetitions of the same item (left columns, ‘with PSFs’). The results show that using previous repetition PSFs in addition to the original model resulted in increased test classification accuracy, precision, recall, and F1-score values. The F1-score increased by 13.5%, indicating substantial improvements in the precision and robustness of the model by adding PSFs. Using previous repetition PSFs also resulted in an increase in explained variance in response times (6.9%), and a reduction in overall Root Mean Square Error (RMSE).

**Table 2.** Improving model predictions using prosodic speech features.

Accuracy	With PSFs	Without PSFs	RT	With PSFs	Without PSFs
Accuracy	0.882	0.864	Test R <sup>2</sup>	72.689	67.953
Precision	0.684	0.657	Test RMSE	700.484	749.035
Recall	0.339	0.286			
F1-score	0.453	0.399			

Note: values represent averages for 10-fold cross validated test predictions.

Figure 2B shows the importance of each of the PSFs included in the elastic net regression analyses. The average absolute feature importance (defined as the penalized regression coefficient) is shown on the x-axis. The black numbers indicate how many trials back the PSF was recorded. The best predictor of response times is pitch slope, followed by speaking speed and intensity. For most PSFs, the previous two repetitions were the most important predictors of current trial response time. PSFs recorded more than two repetitions in the past are generally less informative for current trial performance (more than two repetitions back, most speech features have very low coefficients). For average pitch, jitter and shimmer, we found low coefficients and inconsistent signs over preceding repetitions, underlining the lack of evidence for the informative value of these features. Overall, these results show that using pitch slope, speaking speed and intensity for the two repetitions preceding the current repetition most substantially improve predictions of future recall performance relative to accuracy and response latency.

## 4 Discussion

In this study, we (a) explored the association between high-level prosodic features in speech and recall performance in a learning paradigm, and (b) examined the possibility of using prosodic speech information to improve a cognitive model of memory retrieval that can be used to for item scheduling in an applied learning system. To this end, we analysed speech recordings from over 7,000 retrieval attempts for 44 participants and automatically extracted six high-level PSFs. We will reiterate and interpret the results for (a) and (b) in turn.

Correlational and mixed effects generalized linear regression analyses revealed that accuracy during spoken retrieval attempts is negatively associated to pitch slope, and positively associated to speaking speed and intensity. In other words, for incorrect retrieval attempts, participants were more likely to speak with rising pitch, lower speaking speeds, and lower intensity than for correct retrieval attempts. In addition, we found that response latency is positively associated to pitch slope, and negatively associated to speaking speed and intensity. These results replicate and extend the findings of Goupil and Acouturier [6], who recently reported that higher intensity was associated with higher retrieval accuracy. Intuitively, our results also align with earlier research on the association between *confidence* and prosody, in which rising pitch and lower speaking speeds were tied to uncertainty about a response [6, 10]. While subjective confidence in the response was not directly measured in this study, it is reasonable to assume that overall confidence in accurate and fast retrieval attempts was higher than confidence in inaccurate or slower responses. Finally, we found no clear effects of average pitch, average jitter and average shimmer in an utterance. These features were included in our analyses on an exploratory basis (we are not aware of any relevant studies that report an association between accuracy or subjective confidence in a response). We conclude that the latter features are not useful indices of memory strength in a learning context.

We found that using previous repetition PSFs, in addition to previous repetition response latencies and accuracy scores, to predict current repetition accuracy resulted in increased classification performance. The F1-score of the model increased with 13.5%, indicating that the balanced predictive performance of the model improved substantially. Similarly, we show that utilizing previous repetition PSFs increases the explained variance of the model predicting current repetition response latencies by 6.9%. Overall, these results show that using previous repetition PSFs in addition to response times and accuracy can result in a substantial improvement of overall adaptive learning model performance. These results demonstrate that PSFs provide information in addition to accuracy and response times that can be used as a behavioral proxy of latent memory strength in adaptive learning models.

Our results lead to various suggestions for future work. First, although we find that information gathered through prosodic speech analyses can improve model predictions, our results do not show *why* a specific speech signature is associated with better or worse recall performance. More specifically, pitch slope, speaking speed, and intensity could reflect the objective memory strength for a vocabulary item, much like the assumed relationship between memory strength and response times [4]. Alternatively, they could be a reflection of subjective confidence in the response. Future studies should shed light on this issue by directly assessing the speakers' subjective confidence in each response. Second, this study used single words as cues and response options. Arguably, more or stronger prosodic information could be extracted if combinations of multiple words, or short sentences are used, because longer utterances give the speaking more opportunity to vary intonation, rhythm and stress [17]. Finally, in the current study, PSFs were extracted offline after completion of the experiment. In future projects, PSFs should be extracted to update predictions of future performance in real time to optimize scheduling or feedback presentation in an adaptive learning session. A technical challenge associated with this approach is that for the current study, PSFs were standardized by each learner using the complete learning history for that learner, which would not be

possible in a real-time setting. Future studies should consider the practical feasibility of standardising PSFs based on the first couple of responses only.

In addition, it is important to consider issues related to the specificity of participant sample used in the current study. In an additional set of analyses, we found no significant effects of age or gender on the effects reported in this study. In addition, earlier research suggests that universal PSFs are found in speech throughout languages and cultures [7, 17, 29]. Nevertheless, we acknowledge that it is of key importance to further examine if the results found in this study, which can be construed as a proof-of-concept, generalize to other groups of learners throughout the world.

Despite these open questions, our results have important implications. First, they have fundamental relevance as they are the first to couple a specific prosodic speech signature (falling pitch, high speaking speed, high vocal intensity) to high accuracy and fast responses in a learning paradigm, suggesting that PSFs may be used as a measure of speaker confidence during memory recall. More generally, although more research into the exact nature of the relationship between PSFs and memory performance is necessary, 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 are educationally relevant because they can contribute to improving cognitive models of memory retrieval that are used in real world learning settings: The results of this study lead to specific recommendations on *how* to use PSFs in adaptive learning models.

In 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. As such, they are a promising candidate to be used in learning research and in educationally relevant speech-based learning applications.

## References

1. Anderson, J.R., Bothell, D., Lebiere, C., Matessa, M.: An integrated theory of list memory. *J. Mem. Lang.* **38**(4), 341–380 (1998)
2. Baayen, R.H., Davidson, D.J., Bates, D.M.: Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**(4), 390–412 (2008)
3. Boersma, P.: Praat: doing phonetics by computer (2006). <http://www.praat.org/>
4. Byrne, M.D., Anderson, J.R.: Perception and action. *Atomic Comp. Thought* **16**, 23–28 (1998)
5. Golonka, E.M., Bowles, A.R., Frank, V.M., Richardson, D.L., Freynik, S.: Technologies for foreign language learning: a review of technology types and their effectiveness. *Comput. Assist. Lang. Learn.* **27**(1), 70–105 (2014)
6. Goupil, L., Aucouturier, J.J.: Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition* **212**, 104661 (2021)
7. Goupil, L., Ponsot, E., Richardson, D., Reyes, G., Aucouturier, J.J.: Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nat. Commun.* **12**(1), 1–17 (2021)
8. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and *F*-score, with implication for evaluation. In: Losada, D.E., Fernández-Luna, J.M. (eds.) *ECIR 2005. LNCS*, vol. 3408, pp. 345–359. Springer, Heidelberg (2005). [https://doi.org/10.1007/978-3-540-31865-1\\_25](https://doi.org/10.1007/978-3-540-31865-1_25)

9. Jiang, N.: Lexical development and representation in a second language. *Appl. Linguis.* **21**(1), 47–77 (2000)
10. Jiang, X., Pell, M.D.: The sound of confidence and doubt. *Speech Commun.* **88**, 106–126 (2017)
11. Lindsey, R.V., Shroyer, J.D., Pashler, H., Mozer, M.C.: Improving students' long-term knowledge retention through personalized review. *Psychol. Sci.* **25**(3), 639–647 (2014)
12. Liu, Z.T., Rehman, A., Wu, M., Cao, W.H., Hao, M.: Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. *Inf. Sci.* **563**, 309–325 (2021)
13. Nedungadi, P., Remya, M.: Incorporating forgetting in the personalized, clustered, Bayesian knowledge tracing (pc-BKT) model. In: 2015 International Conference on Cognitive Computing and Information Processing (CCIP), pp. 1–5. IEEE (2015)
14. Papousek, J., Pelánek, R., Stanislav, V.: Adaptive practice of facts in domains with varied prior knowledge. In: Educational Data Mining 2014, pp. 6–13 (2014)
15. Pavlik, P.I., Anderson, J.R.: Using a model to compute the optimal schedule of practice. *J. Exp. Psychol. Appl.* **14**(2), 101 (2008)
16. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020)
17. Reed, B.S.: *Analysing Conversation: An Introduction to Prosody*. Macmillan International Higher Education (2010)
18. Sense, F., Behrens, F., Meijer, R.R., Van Rijn, H.: An individual's rate of forgetting is stable over time but differs across materials. *Top. Cogn. Sci.* **8**(1), 305–321 (2016)
19. Sense, F., Meijer, R.R., Van Rijn, H.: Exploration of the rate of forgetting as a domain-specific individual differences measure. *Front. Educ.* **3**(112) (2018)
20. Sense, F., van der Velde, M., Van Rijn, H.: Predicting university students' exam performance using a model-based adaptive fact-learning system. *J. Learn. Anal.* **8**, 1–15 (2021)
21. Settles, B., Brust, C., Gustafson, E., Hagiwara, M., Madnani, N.: Second language acquisition modeling. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 56–65 (2018)
22. Settles, B., Meeder, B.: A trainable spaced repetition model for language learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1848–1858 (2016)
23. Van Rijn, H., van Maanen, L., van Woudenberg, M.: 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 (2009)
24. Van Rossum, G., Drake, F.L.: *Introduction To Python 3: Python Documentation Manual Part 1*. CreateSpace (2009)
25. Ververidis, D., Kotropoulos, C.: Sequential forward feature selection with low computational cost. In: 2005 13th European Signal Processing Conference, pp. 1–4. IEEE (2005)
26. Walsh, M.M., et al.: Mechanisms underlying the spacing effect in learning: a comparison of three computational models. *J. Exp. Psychol. Gen.* **147**(9), 1325 (2018)
27. Wennerstrom, A.: *The Music of Everyday Speech: Prosody and Discourse Analysis*. Oxford University Press, Oxford (2001)
28. Wilschut, T., Sense, F., van Rijn, H.: Speaking to remember: model-based adaptive vocabulary learning using automatic speech recognition. Available at SSRN 4227060 (2022)
29. Xu, Y.: Speech prosody: a methodological review. *J. Speech Sci.* **1**(1), 85–115 (2011)
30. Yujian, L., Bo, L.: A normalized Levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1091–1095 (2007)
31. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**(2), 301–320 (2005)