IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 March 2023, accepted 9 April 2023, date of publication 13 April 2023, date of current version 25 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266731

RESEARCH ARTICLE

Artificial Neural Network Based Audio Reinforcement for Computer Assisted Rote Learning

PARISA SUPITAYAKUL, ZEYNEP YÜCEL, (Member, IEEE), AND AKITO MONDEN, (Member, IEEE)
Graduate School of Natural Science and Technology, Okayama University, Okayama-shi 700-8530, Japan

Corresponding author: Parisa Supitayakul (pgw45ydd@s.okayama-u.ac.jp)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Okayama University, Faculty of Engineering, Department of Computer Science, Theory of Programming and Artificial Intelligence Research Group.

ABSTRACT The dual-channel assumption of the cognitive theory of multimedia learning suggests that importing a large amount of information through a single (visual or audio) channel overloads that channel, causing partial loss of information, while importing it simultaneously through multiple channels relieves the burden on them and leads to the registration of a larger amount of information. In light of such knowledge, this study investigates the possibility of reinforcing visual stimuli with audio for supporting e-learners in memorization tasks. Specifically, we consider three kinds of learning material and two kinds of audio stimuli and partially reinforce each kind of material with each kind of stimuli in an arbitrary way. In a series of experiments, we determine the particular type of audio, which offers the highest improvement for each kind of material. Our work stands out as being the first study investigating the differences in memory performance in relation to different combinations of learning content and stimulus. Our key findings from the experiments are: (i) E-learning is more effective in refreshing memory rather than studying from scratch, (ii) Non-informative audio is more suited to verbal content, whereas informative audio is better for numerical content, (iii) Constant audio triggering degrades learning performance and thus audio triggering should be handled with care. Based on these findings, we develop an ANN-based estimator to determine the proper moment for triggering audio (i.e. when memory performance is estimated to be declining) and carry out follow-up experiments for testing the integrated framework. Our contributions involve (i) determination of the most effective audio for each content type, (ii) estimation of memory deterioration based on learners' interaction logs, and (iii) the proposal of improvement of memory registration through auditory reinforcement. We believe that such findings constitute encouraging evidence the memory registration of e-learners can be enhanced with content-aware audio incorporation.

INDEX TERMS E-learning, neural networks, artificial intelligence, cognitive theory of multimedia learning, cognitive load, distinctiveness account, perceptual decoupling, adaptability, educational data mining.

I. INTRODUCTION AND MOTIVATION

Humans' information registration framework has a *three-store structure* composed of sensory memory, working memory, and long-term memory [1]. Sensory memory reacts to perception through the five senses, which are sight, hearing, taste, smell and touch. Subsequently, the information received by sensory memory is stored for some time, before being transmitted to short-term and/or working memory and being registered in long-term memory (or discarded) [2], [3].

In educational psychology, the theory, which explains the roles of these processes in learning, is called "Cognitive

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

theory of multimedia learning'' (CTML) [4]. This theory has three basic elements, namely dual-channel assumption, limited-capacity assumption, and active-processing assumption [4]. In this study, we consider in particular the roles of the dual-channel and limited-capacity phenomena in learning and investigate how such principles of educational psychology (developed principally for conventional settings like a classroom) can be deployed for improving memory registration in technology-based learning.

To that end, we focus on a particular kind of learning system, namely, virtual flashcard software, which is used for *rote learning*, i.e. a memorization technique based on repetition. Due to its simplicity, rote learning constitutes an important element in learning foreign language vocabulary, mathematical formulas, and chemical facts and the like [5], [6], and [7]. However, rote learning also bears several challenges, such as being monotonous, which makes it difficult to sustain focus and motivation.

In that respect, the objective of this study is to build a mechanism for supporting learners carrying out computer-assisted rote learning, such that their memory registration improves.

Specifically, we first point out the distinctions in memory registration, which arise due to variations in the type of learning content and the type of stimulus that delivers it. Based on empirical evidence, we identify the type of audio stimulus, which leads to the largest improvement in memory registration concerning each content type under investigation.

We then exploit such findings together with the dual-channel phenomenon and distinctiveness account. In particular, we propose reinforcing visual stimuli with audio incorporation, particularly for those pieces of content, that the learner is likely to forget. To estimate the instants (i.e. particular piece of the content) at which the learner needs such kind of reinforcement, we train an artificial neural network (ANN) with the data on learners' interaction (i.e. activity log) with the e-learning system. We incorporate this with memory test results and determine the proper piece of learning material to trigger the most appropriate kind of audio reinforcement. The estimator is embedded into the e-learning system so as to make real-time estimations and support the information delivered by visual material.

By presenting several rote learning tasks varying in content (i.e. verbal or numerical) and difficulty (i.e. easy or hard), we carried out a comprehensive assessment. The proposed ANN-based audio reinforcement scheme is shown to offer (on average) an improvement over the baseline (i.e. visual-only case) as well as extreme (i.e. constant triggering) cases concerning all three kinds of learning content. In addition, constant triggering is seen to lead to a decrease in performance, which suggests that the triggering should be handled with care.

Our work stands out as the first study addressing the distinctions in memory registration for different combinations of learning content and stimulus. Our contributions involve the determination of the most effective audio for each content type, the estimation of memory deterioration based on learners' interaction logs, and the proposal for improvement of memory registration through ANN-based auditory reinforcement.

II. RELATED WORK

E-learning has numerous advantages such as being economical, customizable, etc. However, it also has several challenges [8], mainly due to a lack of social interaction with the instructor or other learners. Namely, the learners, who study alone in their own schedule and room, often find it difficult to establish self-discipline and lose motivation easily [9].

To tackle such disadvantages, it is necessary first to understand the basics of human cognition and memory concerning learning and then to investigate what kind of remedies can be offered by deploying the embedded capabilities of the digital host platform (e.g. recognition of disengagement, failure, dropout or invoking reengaging/reinforcing stimuli, etc.). Thus, in what follows we first explain relevant concepts in human perception, the way information is registered, retained/discarded, and the effect of concurrent stimuli and then explain about existing works on supporting e-learners with artificial intelligence (AI) based adjustments and interferences.

Although humans perceive information through the five senses, visual and auditory channels are considered to be the main channels for receiving information. According to [10], visual and auditory modalities are processed in separate streams with different properties and limited capacity. Thus, importing a large amount of information through a single channel may overload that channel and some part of the information may get lost. On the contrary, if the information is imported by multiple channels simultaneously, it is possible to avoid causing a burden on any channel and, thus, a larger amount of information can be processed [11].

There are numerous studies in literature examining the relationship between visual and audio stimuli and learning. For instance, Thompson and Paivio examine the independence of auditory and visual nonverbal stimuli (pictures, corresponding environmental sounds, or picture-sound pairs) in free recall and report the best recall rate for the picture-sound pairs [12]. They repeat a similar procedure by introducing a distracter task, which confirms the efficacy of using two modalities at once as compared to single modalities separately. Moreover, Tindall-Ford et al. investigate further the effect of modality by comparing the learning materials between various dual-mode presentations (e.g. auditory text and visual diagrams) and single-modality formats (e.g. visual text and visual diagrams) [13]. They show that dual-mode presentation (audio and visual) has a higher performance than single-modality formats (visual) and explain this observation as an expansion of working memory due to the perception of information through multiple channels. Also, Heikkilä et al. show that memory efficiency is better when the content (e.g. pictures, words, sounds) is accompanied by semantically congruent sound [14]. Although these studies provide valuable information about humans' perception and processing of information, they target *recollection* (i.e. remembering of a past event or experience) rather than *learning* (i.e. acquiring of a new piece of information), which limits a generalization to real-life learner activity.

Nevertheless, there are also works on dual-modality presentation specifically in learning settings. For instance, Moreno and Mayer study the redundancy in multimedia learning and examine the difference between using the auditory alone and auditory+visual modalities [15]. They conclude that redundant stimuli help the students to comprehend the learning material better and that using dual modality is more effective than using a single modality. In addition, Kim and Godfroid focus specifically on second language learning and show that learners gain information from all modalities [16]. They report an interesting observation that implicit knowledge is developed with better efficiency when the content is presented visually and claim visual input to be beneficial especially for beginners. In addition, Kaplan-Rakowski and Loranc-Paszylk focus on the effect of auditory stimuli on learning foreign vocabulary presented together with (i) sound effects, (ii) pronunciation, (iii) sound effects+pronunciation, (iv) no audio [17]. The results indicate that vocabulary presented with sound effects achieves significantly higher scores. In light of these studies, proper integration of audio stimuli into e-learning systems is suggested to boast a big potential of improving learning efficiency.

Interpersonal variations pose a big challenge in the proper integration of multi-channel stimuli into e-learning systems. Therefore, the utilization of artificially intelligent tools, which can adapt to the behavioral patterns of different learners, is considered to be very beneficial [18]. Nevertheless, personalization is highly multi-dimensional and it is quite difficult to find the right balance between the numerous factors for every single individual.

In that respect, the seminal works of Gardner, Kolb et al. and Felder & Silverman on learning style models (LSM) emerge as structured bases, on which personalization of learning can be contrived [19], [20], [21]. There is an abundant number of works targeting the identification of LSMs [22], [23], mostly adopting the approach of Felder & Silverman. In addition, a substantial part of those deploy machine learning methods [24] and in this section, we will summarize several such studies, which are relevant for us, particularly from the point of view of estimation approach (i.e. ANN-based).

Gambo and Shakir [25] generate a data set by simulating students' learning behavior and then developing an ANN to map the meta-cognitive skills to LSMs. Their main limitation is the deployment of simulated data instead of data collected from human participants. In addition, Zhang et al. [26] predict students' learning preferences along the four dimensions of Felder & Silverman LSM using deep belief networks (DBN) on online learning session data. They show that the proposed DBN provides more accurate estimates than several other alternative architectures, but the advantage of such estimates in improving learning efficacy is not demonstrated experimentally. On the other hand, Zhang et al. [27] use a real-world data set to predict students' course grades through a multi-source sparse attention convolutional ANN architecture, which also provides an explanation/interpretation of failures. The advantages of their method lie principally in enabling personalized course recommendation and association mapping between courses. The main limitation is that it is defined at coarse-level (e.g. at semester or exam level) and generalization to finer levels (i.e. individual assignments or questions) is not trivial. Many other studies estimating student performance (e.g. withdrawal, pass/fail, drop-out) in a coarse-grained fashion with such tools as ANN, support vector machines, logistic regression share this disadvantage [28], [29], [30], [31], [32]. In that respect, this study aims to obtain similar predictions, but at a much finer scale (within seconds or minutes). Moreover, it also deploys those predictions in taking actions for supporting learners in a timely fashion.

III. OVERVIEW

Our methodology can be broken into three main stages as shown in Figure 1.

- **Exploration stage**: We consider three types of rote learning tasks (with different contents) and three kinds of stimuli to deliver those tasks. We deliver each type of content with each kind of stimulus (see also Figure 2-(a)) to a group of 9 participants (3 females and 6 males with an average age of 22.5 ± 2.3).[1] In this first set of experiments, we collect learners' activity logs and performance in memory tests. In the subsequent data analysis stage, we determine the stimulus type, which leads to the highest learning rate concerning each content type, by evaluating learning rates.
- **Design stage**: To estimate learners' improvement/deterioration of memory performance concerning each piece of information (i.e. each item to be memorized), we develop an ANN. Specifically, we deploy the activity log files collected in the exploration stage, which are used to derive the features to track learners' interaction with the e-learning software as well as the evolution of their perception of difficulty concerning each piece of task. Based on such information, we train our ANN and construct our estimation model. Finally, we integrate the estimation model into the e-learning system, such that it provides appropriate audio reinforcement for various content types, while also adapting to individual learners on-the-fly.

[1] In particular, our participants are undergraduate or graduate university students or workers, who are not involved with this area of research, are all mother tongue Japanese speakers and reported no serious health issues related to their visual or auditory senses. In addition, based on a dummy session carried out before (actual) data collection, it is confirmed that they can read the text on the screen and hear the audio without problems (see also Appendix D).

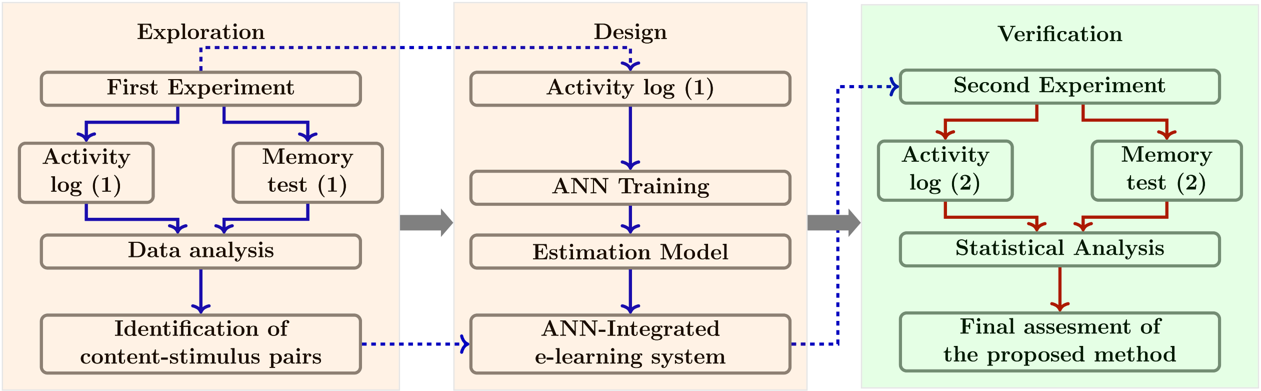


FIGURE 1. Overview and components of each stage of our study.

- **Verification stage**: We integrate the ANN into our e-learning platform and test it on 11 participants (1 female and 10 males with an average age of 23.2 ± 1.8). In particular, when memory performance is estimated to be deteriorating concerning a certain piece of information, the reinforced e-learning system triggers the appropriate kind of audio stimulus at the next viewing of that item. In this second set of experiments, we collect additional activity logs and memory tests, which are used in the statistical analysis to prove the efficacy of the proposed approach.

IV. EXPERIMENTATION

In experiments, participants carried out rote learning tasks on Anki, a free and open-source virtual flashcard program [33]. Analogous to physical flashcards, each virtual flashcard has a "front" and a "back" face (referred also as Q-face and A-face, see also Appendix A). When the learner is exposed to the front face of a card, he/she takes some time to recall the information on the A-face. Then, he/she "flips" the card to disclose the A-face and register (i.e. memorize) it. Once ready, he/she proceeds by evaluating the difficulty of the card choosing "Again", "Good" or "Easy". We term watching the front and then back face as a *viewing* of that card. Concerning each task, let the participants study a group of 12 cards, called a *deck* for 4 minutes. Anki registers learners' activity logs in terms of UNIX time of button clicks, card and deck IDs, and their subjective evaluations of difficulty (see Appendix A for details).

To discover the effect of audio incorporation on memory retention of different learning materials, we consider various couplings of content type and stimulus type. The content types are common for exploration and verification stages and are denoted with C_E, C_H, C_N. Here, C_E stands for easy verbal, C_H is hard verbal and C_N is numerical content (see Appendix B). The stimulus types differ between the exploration and verification stages. At the exploration stage, we use S_V, S_A, S_B, where S_V is visual (baseline), S_A is a combination of visual and informative audio, which refers to the human-readout of the visual information, and S_B is a combination of visual and non-informative audio, which is a bell sound (see Appendix C). At the verification stage, we deploy S_E and S_F, where S_E is the proposed estimation-based scheme and S_F is full audio reinforcement (triggering audio stimuli for each item to be memorized).

V. EXPLORATION STAGE

At the exploration stage, each participant performed nine sessions corresponding to the nine couplings of C_E, C_H, C_N and S_V, S_A, S_B (see Figure 2-(a)).

(i) Three sessions are implemented with three kinds of content displayed *only visually* (i.e. S_V).

(ii) Three sessions are implemented with three kinds of content such that half of each task is delivered with only visual information, whereas the other half is delivered with a *combination of visual and informative audio* (i.e. S_A).

(iii) Three other sessions are implemented three kinds of contents such that half of each task is delivered visually, whereas the other half is delivered with a *combination of visual and non-informative audio* (i.e. S_B).

This configuration allows us to identify the difference in recollection performance due to content type as well as to decide which kind of audio reinforcement is best for each content type (See Appendix E).

Note also that the partial couplings in (ii) and (iii) are inspired by the *distinctiveness account* [34], [35]. Specifically, delivering part of the learning session with only visual stimuli and the other part with visual and audio is expected to highlight the supplementary stimuli, which will in turn enhance the subsequent memorability of the related pieces of information.[2]

[2] If the supplemental stimuli are attached to all pieces of information, there is the danger that distinctiveness is undermined and the reinforcement effect vanishes [35]. Such a configuration is implemented and tested in Section VII, whose results confirm that full triggering has lower performance than a mixed configuration.

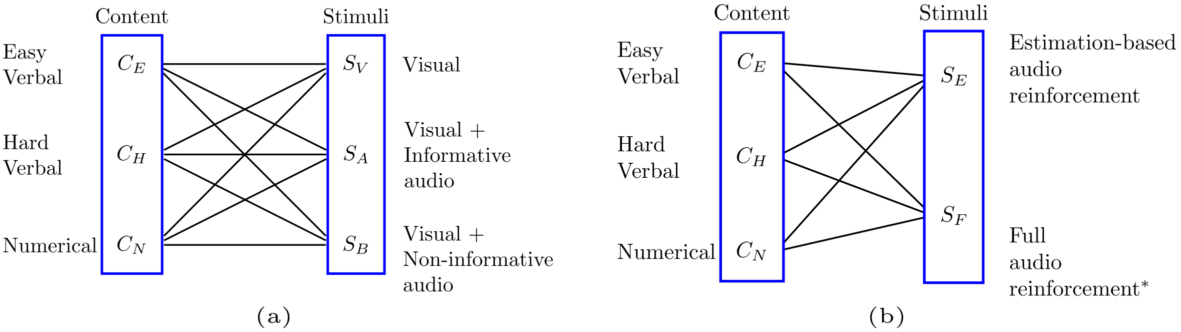


FIGURE 2. Each participant performed a total of (a) nine learning sessions in the exploration stage and (b) six learning sessions in the verification stage.

TABLE 1. (a) Statistics and (b) ANOVA and effect size concerning information gained I_G for varying content types at exploration stage.

(a)

Content	#	μ	ϵ
C_E	238	0.91	0.02
C_H	319	0.54	0.03
C_N	309	0.75	0.02

(b)

Content	#	p	d
$C_E - C_H$	557	$< \mathbf{10^{-2}}$	0.88
$C_E - C_N$	547	$< \mathbf{10^{-2}}$	0.41
$C_H - C_N$	628	$< \mathbf{10^{-2}}$	0.46

TABLE 2. (a) Statistics and (b) ANOVA and effect size concerning information gained I_G for varying stimulus types at exploration stage.

(a)

Stimulus	#	μ	ϵ
S_A	299	0.74	0.03
S_B	277	0.75	0.03
S_V	290	0.66	0.03

(b)

Stimulus	#	p	d
$S_A - S_B$	576	0.68	0.08
$S_A - S_V$	589	**0.04**	0.17
$S_B - S_V$	567	**0.02**	0.20

To evaluate performance, we study the amount of *information gained* I_G, which we define as the change in learner's knowledge succeeding the learning session as compared to his/her preceding state (see Appendix E for details). In the following tables, the number of data points is denoted by #, and mean and standard error are denoted by μ and ϵ, respectively, whereas p-value and effect size are represented with p and d.

First, we pay regard to variation in content types (omitting variations in stimulus type) and compute the statistics concerning I_G as presented in Table 1-(a). We see that C_E attains a much higher I_G on average (i.e. 0.91) than C_H and C_N. This indicates that the e-learning platform is useful for studying *easy* content; namely, *refreshing* a material, which is possibly already known (studied) in the past but then forgotten. Moreover, I_G is considerably higher for C_N than for C_H, even though the two can be equally unfamiliar to the learner as explained in Appendix B. In that respect, it is possibly more beneficial to use traditional techniques (e.g. pen and paper) in studying C_H.

From Table 1-(b), we can see that the three content types present a significantly different relation concerning I_G. It is interesting that although we do not pay any regard to the type of stimulus delivering the content, the significance of the difference is still obvious. This indicates the substantiveness of content type as compared to the stimuli that deliver it. Moreover, the related values of effect size d can be regarded as medium to large.

Next, we examine the relationship between stimulus type and I_G paying regard to variations in stimulus type (and omitting variations in content type, see Table 2-(a)). We can see that on average I_G is lower for contents presented with visual-only stimuli S_V (i.e. 0.66) than informative audio S_A and non-informative audio S_B. In other words, having an extra stimulus, irrespective of being informative or non-informative, is beneficial over having only a visual stimulus. Note also that all cases in Table 2-(a) attain a quite low standard error, which indicates that the mean of the given sample is considerably accurate.

In Table 2-(b), it is seen that S_V is different from audio incorporated cases S_A and S_B in a statistically significant way ($p = 0.04$ and $p = 0.02$). Although the difference between S_A and S_B is not found to be significant ($p = 0.68$), this may be due to the fact that we disregard content types in Table 2-(b). In addition, the related values of effect size d are quite small.

Subsequently, we present information gained I_G for each pair of content type and stimulus type in Table 3. Concerning C_E, we observe that the biggest improvement over S_V is provided by the incorporation of S_A (i.e. 0.84 → 0.96). Also, the ANOVA presented in Table 3-(b) proves that the improvement provided by S_A over S_V is statistically significant ($p = 0.01$) with small to medium effect size. However, learning rates concerning C_E are already quite high with S_V,

TABLE 3. (a) Statistics and (b) ANOVA concerning information gained I_G for each pair of content type and stimulus type at exploration stage.

(a)

	C_E			C_H			C_N		
Stimulus	#	μ	ϵ	#	μ	ϵ	#	μ	ϵ
S_A	91	**0.96**	0.02	106	0.50	0.05	102	**0.78**	0.04
S_B	66	0.92	0.03	107	**0.64**	0.05	104	0.76	0.04
S_V	81	0.84	0.04	106	0.47	0.05	103	0.71	0.04

(b)

	C_E			C_H			C_N		
Stimulus	#	p	d	#	p	d	#	p	d
$S_A - S_B$	157	0.40	0.14	213	**0.05**	0.27	206	0.67	0.06
$S_A - S_V$	172	**0.01**	0.40	212	0.68	0.06	205	0.22	0.17
$S_B - S_V$	147	0.12	0.26	213	**0.02**	0.33	207	0.41	0.11

TABLE 4. ANOVA and effect size concerning information gained I_G for each pair of content type and stimulus type at exploration stage.

	(a)			(b)			(c)		
	C_E			C_H			C_N		
Stimulus	#	p	d	#	p	d	#	p	d
$S_A - S_B$	157	0.40	0.14	213	**0.05**	0.27	206	0.67	0.06
$S_A - S_V$	172	**0.01**	0.40	212	0.68	0.06	205	0.22	0.17
$S_B - S_V$	147	0.12	0.26	213	**0.02**	0.33	207	0.41	0.11

as compared to those of C_H and C_N. In that respect, it may also be considered that it is not possible to expect a significant improvement in C_E through audio reinforcement. In other words, there is a more imminent need for improvement concerning the other content types C_H and C_N.

Concerning C_H, S_A is observed to yield virtually the same I_G as visual-only stimuli S_V (i.e. 0.47 → 0.50), despite what the dual-coding theory suggests [36] (see Table 3-(a)). However, very interestingly, non-informative audio S_B is observed to provide a bigger improvement (i.e. 0.47 → 0.64). To unveil the reasons for such unexpected observations, we examined the interaction of the learners with the e-learning platform and noticed that the rate of card views is higher for S_B than S_V and S_A, which may be explained by the fact that the non-informative audio sets a certain faster pace in viewing. This increased rate of viewing may help in recalling the information even more than the human-readout. Note also that the improvement provided by S_B over S_V and S_A is statistically significant ($p = 0.02$ and $p = 0.05$, see Table 3-(b)). In that respect, we propose using non-informative audio S_B to support the learners studying hard verbal contents C_H, i.e. $C_H + S_B$.

On the other hand, C_N is observed to benefit from informative audio S_A and non-informative audio S_B in a similar way (i.e. $0.71 \rightarrow 0.76, 0.78$, see Table 3-(a)). Also, the improvement due to S_A over S_V and S_B is not statistically significant ($p = 0.22$ and $p = 0.67$, see Table 3-(b)). However, since S_A is found to be slightly better than S_V and S_B, as seen in Table 3, we decided to take a chance and support the learners studying numerical contents C_N with informative audio S_A, i.e. $C_N + S_A$.

Note that even though all cases in Table 3 attain a quite low standard error indicating an accurate representation of the true mean, the effect sizes given in Tables 1∼3 are small to medium. Moreover, to understand in general how *persistent* the new knowledge is, concerning each case presented in Table 3, we also checked *Information retained* I_R, which we define as the increase in learner's knowledge from the beginning of the learning session to 5 minutes after the finish of the learning session (see Appendix E for details). The tendencies between information gained I_G and information retained I_R are found to be quite parallel between different couplings of content type and stimulus type (see Appendix F for details).

VI. DESIGN STAGE

The data collected in the exploration stage is deployed in designing an estimator. Specifically, we aim at estimating learners' memory performance concerning *each item to be memorized*,[3] as either improving or deteriorating. As an estimator, we prefer an ANN due to its ability to capture complex characteristics of data.

A. INPUT AND OUTPUT VARIABLES

Certain variables, which potentially contain characterizing information about learners' behavior or state, are determined to be the inputs of our estimator model. Table 5 provides a list of the variables that we choose to deploy as inputs of the ANN.

Content type and stimulus type are considered to be part of the inputs, as the learners may present a different reaction to each pair (rows 1,2 of Table 5). Moreover, since audio triggering is carried out for half of the flashcards in any deck

[3] Note that in Section V the purpose was to determine the *on average* most efficient coupling of content type and stimulus type. In that respect, the items to be reinforced with audio are chosen arbitrarily. On the other hand, in the design stage the purpose is to trigger the reinforcement at the *proper moment* and for the *proper piece of information*.

TABLE 5. Inputs of the ANN estimator.

	Category	Relating Variables	Encoding/ Possible values
1	Content type	'Easy', 'Hard', 'Numerical'	One-hot
2	Stimulus type	'Visual', 'Informative', 'Non-informative'	One-hot
3	Audio attachment	Attached (or not)	Binary
4	Number of viewing	'Number of viewing'	$\mathbb{R}$
5	Temporal variables	't_q', 't_a'	$\mathbb{R}$
6	Difficulty	Subjective evaluation	$\mathbb{R}$

at the exploration step, the absence/presence of audio clips are also fed as inputs (row 3 of Table 5).

The "number of current viewing" at a given time is defined as the number of times that the learner watched a flashcard from the beginning of the learning session until that moment (row 4 of Table 5). Since learners' memory performance is expected to improve gradually (i.e. with the growing number of viewings), it is considered to be a valid indicator.

The time spent on Q-face (t_q) and time spent on A-face (t_a) of a flashcard may present a nontrivial relation with memory performance (row 5 of Table 5). A very short t_q can be either due to the confidence of the learner in his/her recollection or due to a complete lack of recall, whereas intermediate values may indicate an effort to remember. A very short t_a may indicate confidence in evaluation, whereas a long t_a may suggest either an effort to register the answer or a hesitation between two evaluation choices (e.g. Good or Easy). Thus, although t_q and t_a cannot directly be associated with a certain state, they can provide valuable information once they are considered together with the other inputs.

Subjective evaluation may perhaps be considered to be an indicator good enough on its own (i.e. without incorporating with the above-mentioned inputs). But we observed the learners to be quite conservative and pessimistic with their evaluations and to opt for the lower score, should they hesitate between two choices. In that respect, we prefer to make estimations based on not only their subjective evaluations (row 6 of Table 5), but also by taking into account additional indicators.

Since we represent content type and stimulus type with one-hot-encoding [37], a total of 11 variables are used as inputs of the ANN (see Table 5). All variables are integers, where the number of current viewing is dimensionless, and t_q and t_a are in millisecond resolution. Note also that although certain variables are registered as integers into the log file, they are fed as real numbers into the ANN, after being preprocessed.

The output layer is composed of a single neuron, which can take a binary value (0/1) and indicates whether the audio reinforcement needs to be triggered (as a 1) or not (as a 0). In the training stage of the ANN, the ground truth for output is judged based on the improvement or deterioration of the learner in terms of his/her subjective evaluations in two consecutive interactions with a certain flashcard. Specifically, if the learner evaluates a flashcard with increasing confidence (e.g. first "Again" and then "Good"), we assume that his/her memory performance has increased and no audio reinforcement is necessary and the ground truth is set to 0. On the other hand, if he/she gives a lower rating following a higher rating, memory performance is assumed to deteriorate and ground truth is set to 1.

Obviously, the learner may also repeat his/her evaluations, for which we adopt the following strategy. Namely, if the evaluations are low (i.e. two consecutive "Again"s), then the ground truth is set to 1. In this case, although there is no improvement or deterioration, the evaluation of the learner is the lowest possible, and thus, audio reinforcement is considered to be beneficial.

If the evaluations are medium (i.e. two consecutive "Good"s), there is again no change in state, but after several steps there is a possibility of improvement or deterioration. Thus, we adopt a Markovian approach based on transition probabilities, which are estimated empirically. For instance, based on the data collected in the exploration stage, we compute the probability of evaluating a certain card as "Good" and then as "Again" in two consecutive interactions as

$$P_{GA} = \frac{\#(e_i = G \wedge e_{i+1} = A)}{\sum\limits_{X \in \{A,G,E\}} \#(e_i = G \wedge e_{i+1} = X)} \tag{1}$$

where #() denotes the number of instances satisfying a condition, e_i is the evaluation of a card at i^{th} interaction with it, X is an evaluation label, which can be "Again" (A), "Good" (G) or "Easy" (E). Based on such probabilities, we estimate the following subjective evaluations until one that is different than "Good" is achieved. To that end, we draw a random number y from the standard Normal distribution $y \sim \mathcal{N}(0, 1)$ and compare it with the empirical probabilities.

$$e'_{i+1} = \begin{cases} A & 0 \leq y \leq P_{GA} \\ G & P_{GA} < y \leq P_{GA} + P_{GG} \\ E & P_{GA} + P_{GG} < y \leq 1 \end{cases} \tag{2}$$

If the estimated subsequent evaluation e'_{i+1} is once again "Good", we continue drawing random numbers until an e'_{i+1} which is different than "Good" is achieved. Based on that evaluation, we trigger or not the audio.

Note that if the learner evaluates a flashcard as "Easy", then that flashcard is removed from the queue and there will be no subsequent interaction with it, which makes estimation unnecessary.

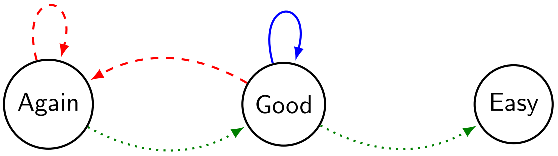


FIGURE 3. Assessment of improvement and deterioration of memory registration. The transitions depicted in green (dotted lines) and red (dashed lines) are considered to correspond to improvement and deterioration, respectively. For evaluating the transitions depicted in blue (solid line), we estimate future evaluations (states) based on empirical probabilities of the corresponding Markov model.

B. PREPROCESSING OF INPUT VARIABLES

For building an efficient estimator, data preprocessing is an essential step. Similar to most other estimation/recognition problems, two main issues with our data set arise as (i) outliers and (ii) significantly different ranges of inputs. In that respect, before building the estimation model, the variables are preprocessed so as to remove the outliers, separate class distributions, and mitigate class imbalance [38].

Especially, the temporal variables are likely to have a problem with outliers. For each temporal variable, we retain the data points, which belong to the interval $\mu \pm 2\sigma$ of the relating distribution and discard all other data points. This empirical approach, which is also known as the 68-95-99.7 rule of thumb, is expected to retain roughly 95% of the data and discard the other 5%, which is considered to be reasonable [39].

As explained in Section VI, most inputs are binary (or one-hot encoded), but there are also certain inputs, which can take values from a large range in $\mathbb{R}$ (see Table 5). Such a difference in scale is likely to cause problems in the training stage. To solve this issue, non-binary input variables are standardized using the `StandardScaler` tool provided by the `sklearn` machine learning library for the Python programming language. Namely, each data point x is represented with its standard score z as

$$z = \frac{(x-\mu)}{\sigma}, \tag{3}$$

where μ and σ are the mean and standard deviations of all observations $\{x\}$ of that variable. Also in the verification stage, the data collected in real-time pass through the same preprocessing operations before being fed into the estimator (i) to judge the necessity of audio reinforcement and (ii) to update the model for achieving an incremental learning process.

C. BASIS ESTIMATOR AND ADAPTIVE UPDATES

Initially, a *basis estimator* is built in an offline manner using the data collected in the exploration stage. This pre-developed model is integrated with the e-learning system and is invoked at the verification stage experiments. To suit the system to every learner in an on-the-fly manner, the basis estimator is updated and adapted automatically to his/her behavior, which is known as *incremental learning* in machine learning.

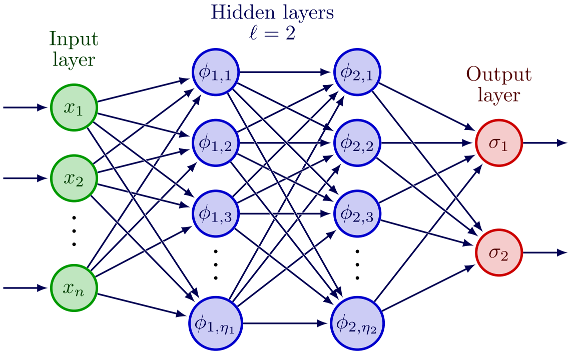


FIGURE 4. The architecture of the ANN. The number of hidden layers (ℓ) is 2 and the number of neurons in the first and second layers (η_1, η_2) are 16 and 8, respectively. The activation function in the hidden layers (ϕ) and at the output layer (σ) are sigmoid and softmax, respectively.

Therefore, in the verification stage the estimator continuously adapts itself to the individual, who uses the e-learning system. Although such a dynamic scheme enables exploiting the potential of all available data (i.e. including those collected within milliseconds), it also requires the model to be as simple and easily trained as possible for enabling on-the-fly adaptation. Taking this into consideration, we decided to use MLP, which is one of the most common neural network architectures and is known to be lightweight and fast.[4]

The tuning of hyper-parameters of MLP involves adjustment of the number of layers, hidden nodes in each layer, activation functions, optimizers, number of training epochs (early stopping), and dropout rate (see Figure 4). To achieve the optimal configuration, we adopted an exhaustive grid-search strategy, experimented with varying combinations of hyper-parameters, and determined the configuration exceeding the edge of accuracy.

It is common to use fully-connected networks (i.e. dense layers) in MLP such that every node in a layer is connected to all nodes in the previous and next layers (see Figure 4). Then, the output of a neuron at, for instance, the first hidden layer will be

$$y_{1,j} = \phi_{1,j}\left(\sum_{i\in[1,n]} w_{1,i} \cdot x_i\right) \tag{4}$$

where $\phi_{1,j}$ is the activation function of the jth neuron at the first layer, $w_{1,i}$ are the weights associated with the inputs of that neuron, and $1 \leq j \leq \eta_1$.

As for the activation function, we decided to use sigmoid function ϕ at the hidden layers and softmax function σ at the output layer (see Equation 5), since they help in reducing the effect of small changes on the outputs, enable normalization

[4] We also considered using feedback connections (Long short-term memory), but decided to stick to feed-forward structure due to its computational efficiency and the simplicity of our time series data.

of the output and probabilistic estimations.[5]

$$\begin{aligned}\phi(x) &= \frac{1}{1+\exp(-x)},\\ \sigma(x_i) &= \frac{\exp(-x_i)}{\sum\limits_{j=1,2} \exp(-x_j)}.\end{aligned} \tag{5}$$

As an optimizer, we considered three popular alternatives as Adam, Stochastic gradient descent (SGD), and Root Mean Square Propagation (RMFprop) and chose to use Adam, since we saw that it provides satisfactory results with a reasonable computational load.

There is no standard rule in machine learning for choosing the optimal number of hidden layers/nodes of a neural network. The number of nodes has to be large enough so that the model has sufficient degrees of freedom to learn, but a model too big is likely to over-fit the training data set and not generalize on new unseen data. The size of the training data should be used as an indicator of the size of the model. Considering the amount of training data, we consider the number of hidden layers $\ell \in \{1, 2, 3\}$ and the number of hidden nodes $\eta \in \{4, 8, 16, 32, 64, 128\}$ (between 20 and 500 parameters). For all possible combinations of ℓ and η, we examined model accuracy and found out that $\ell = 2$ with $\eta_1 = 16$ and $\eta_2 = 8$ (i.e. the number of hidden nodes in layers 1 and 2) is the best combination for accurate and stable results. As for the output layer, we use two neurons, since we have two possibilities of memory performance as improving and deteriorating.

Using dense layers may pose a danger of over-fitting, unless the training set has sufficient samples. Otherwise, the neural network learns the details as well as the noise on the training data. The dropout technique is used to prevent such over-fitting problems and refers to the omission of some arbitrary nodes during training. The portion of hidden nodes ignored at each epoch is called the *dropout rate*, which we represent with α. It is necessary to carefully adjust α, since high dropout rates may cause instability in accuracy and loss values. In this study, we considered $\alpha \in \{0.1, 0.2, 0.3\}$ and examined the learning curve of accuracy and loss concerning training and test sets (see Figure 5). For the dropout rate of $\alpha = 0.1$, when train and validation loss decrease and stabilize, accuracy is observed to be around 70%, which is considered to be a good fit nature, whereas dropout rates of $\alpha = 0.2$ and $\alpha = 0.3$ depict volatile characteristics.

Another method used to prevent over-fitting is the *early stopping* of training. Namely, excessive learning makes the model more complicated and causes over-fitting. To avoid that, it is suggested to abort training when validation error reaches a minimum. Specifically, the performance of the model on the validation set is monitored at every epoch, and training is automatically terminated as soon as the loss reaches its lowest point. We have empirically observed that early stopping kicks in around the update cycle of 35.

VII. VERIFICATION STAGE AND RESULTS

We integrated the trained model with the e-learning platform and performed follow-up experiments for verifying the efficacy of the design. The main difference between exploration stage and verification stage experiments is the integration of ANN-based audio reinforcement (in addition to a fresh set of decks). In the verification stage, the learners are supported with non-informative audio in verbal contents and with informative audio in numerical content. Since such audio is triggered based on the ANN estimator, this configuration is called *estimation based* scheme S_E (see also Figure 2-(b)).

In addition, we consider an alternative audio configuration, namely *full* support S_F. Note that this refers to the extreme triggering, where the learner receives audio reinforcement every time he/she interacts with a card. The reason for considering S_F is for removing any doubt about the *scarcity* of audio reinforcement. Namely, since audio reinforcement has a positive effect on learning rate, one may think that triggering audio at all interactions (i.e. for each card, irrespective of the anticipation of memory improvement/deterioration) is a safe choice and guarantees that no failures are omitted. The inefficiency of such extreme triggering is proven through empirical evidence.

Since S_V is the standard configuration of the e-learning system, it can be considered as a baseline and, thus, results concerning S_V are also presented.

A. EASY VERBAL CONTENTS

In Table 6-(a), the automatic estimation scheme S_E is seen to stand out as the best, followed by visual-only scheme S_V and full support S_F. Namely, with ANN-based audio reinforcement, the learners gain the most (i.e. I_G is largest for S_E).[6] One can also see that the learners forget the least in the case of S_E, since the decrease from I_G to I_R is smaller for S_E than for S_V and S_F. Also, the full-support scheme S_F is seen to lead to the lowest performance among the three cases. In addition, the number of data points is found to be sufficient to identify reliably the position of the means ($\epsilon \ll \mu$).

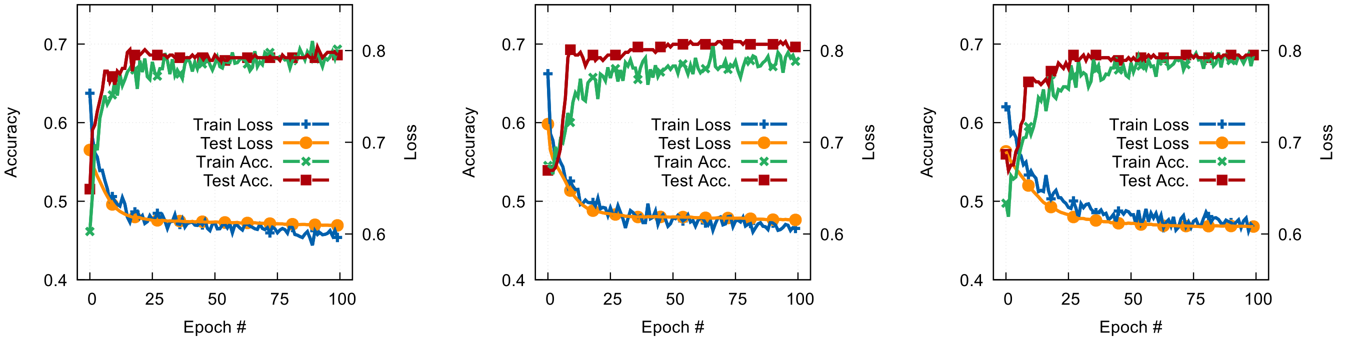
Table 6-(b) presents the relating ANOVA. Examining the values relating S_E–S_F pair, we see that I_G achieves $p = 0.01$ and I_R achieves $p < 10^{-2}$, meaning that S_E is significantly better than S_F. Watching $S_E - S_V$ pair, we see that the improvement of S_E over S_V is not significant for I_G ($p = 0.59$) and I_R ($p = 0.16$). Moreover, the effect size is found to be small to medium.

B. HARD VERBAL CONTENTS

From Table 7-(a), S_E is seen to achieve the best results both in information gained I_G and information retained I_R for C_H. However, the ANOVA scores presented in Table 7-(b) indicate that the difference is not significant ($p = 0.10$ and $p = 0.17$), and yet not very far from the universally accepted

[5] Specifically, $\phi_{1,j} = \phi_{2,k} = \phi$, where $1 \le j \le \eta_1$, $1 \le k \le \eta_2$ and $\sigma_1 = \sigma_2 = \sigma$.

[6] Since I_R is the largest for S_E, the learner is seen also to retain the largest amount of information with S_E. See Appendix F for the statistics on retained knowledge I_R.



(a) (b) (c)

FIGURE 5. Accuracy and loss against epoch for dropout rates of (a) $\alpha = 0.1$, (b) $\alpha = 0.2$, and (c) $\alpha = 0.3$. Train and test loss are illustrated in blue lines with pluses and orange lines with circles, respectively. Train and test accuracy are illustrated in green lines with crosses and red lines with squares, respectively.

TABLE 6. (a) Statistics and (b) ANOVA and effect size concerning information gained I_G and information retained I_R relating to easy verbal contents C_E at verification stage.

(a)

Stimulus	#	I_G		I_R	
		μ	ϵ	μ	ϵ
S_E	127	**0.87**	0.03	**0.81**	0.03
S_F	128	0.75	0.04	0.66	0.04
S_V	81	0.84	0.04	0.73	0.05

(b)

Stimulus	#	I_G		I_R	
		p	d	p	d
S_E - S_F	262	**0.01**	0.30	$\mathbf{< 10^{-2}}$	0.34
S_E - S_V	240	0.59	0.08	0.16	0.20
S_F - S_V	238	0.12	0.22	0.33	0.14

TABLE 7. (a) Statistics (b) ANOVA and effect size concerning information gained I_G and information retained I_R relating to hard verbal contents C_H at the verification stage.

(a)

Stimulus	#	I_G		I_R	
		μ	ϵ	μ	ϵ
S_E	130	**0.56**	0.04	**0.43**	0.04
S_F	126	0.46	0.04	0.37	0.04
S_V	106	0.47	0.05	0.37	0.05

(b)

Stimulus	#	I_G		I_R	
		p	d	p	d
S_E - S_F	256	0.10	0.20	0.28	0.13
S_E - S_V	236	0.17	0.18	0.32	0.13
S_F - S_V	232	0.86	0.02	0.96	0.01

threshold of 0.5. Similar to values concerning C_E presented in Table 6, the standard error values concerning C_H are quite low (see Table 7-(a)) and the effect size is small (see Table 7-(b)). These results are possible to explain with the perceptual decoupling phenomenon and the distinctiveness account [40].

According to the perceptual decoupling phenomenon, one may opt for turning off certain senses to be able to focus fully on the others, which may happen more often for difficult learning materials [41], [42]. In particular, beginners are known to benefit more from visual input (written materials) than other modes [16], which explains why S_E or S_F did not lead to any improvement over S_V.

In addition, when humans are exposed to generated stimuli (e.g. audio attached) and not-generated stimuli (e.g. silent) in the same session, the subsequent memorability of the generated stimuli is shown to improve [34], since distinct stimuli are better encoded in memory. When distinctiveness is undermined (e.g. by constant audio attachment), this effect vanishes [34], which explains why S_F results in very similar scores to S_V.

C. NUMERICAL CONTENTS

Memory performance relating to C_N is presented in Table 8. According to the statistics of I_G, average scores for S_E are found to be slightly better than those for S_F and S_V. Note that this effect diminishes as time lapses and the values of I_R get closer in all three schemes. In addition, there is no significant difference in any of the cases as shown in Table 8-(b). Note that as we mentioned in Section V incorporation of audio was already not expected to contribute to the registration of C_N in a considerable way based on the observations at the exploration stage and this expectancy is confirmed at the verification stage. Note also that level significance decreases with time similar to Section VII-B. In addition, the standard error values of C_N are quite lower than the concerning mean and similar to those of C_E and C_H (see Table 8-(a)), whereas effect size is again small (see Table 8-(b)).

VIII. LIMITATIONS

Although the number of participants is not very low (20 in total) and the number of data points is at the level of hundreds, it is better to use a larger set for a more accurate modeling and evaluation.[7]

In addition, the input of the estimator is the activity logs of the learners, which limits the number of variables that

[7] The experiments are carried out around the end of 2020 and the beginning of 2021. In that period, our institute had a strict policy for the prevention of Covid19 pandemic, and classes and research guidance were held online. Thus, we had difficulty in recruiting participants.

TABLE 8. Statistics concerning (a) information gained I_G and (b) information retained I_R relating to the numerical contents C_N at verification stage.

(a)

		I_G		I_R	
Stimulus	#	μ	ϵ	μ	ϵ
S_E	131	**0.77**	0.04	**0.69**	0.04
S_F	124	0.70	0.04	0.66	0.04
S_V	103	0.71	0.04	**0.69**	0.05

(b)

		I_G		I_R	
Stimulus	#	p	d	p	d
S_E - S_F	255	0.20	0.16	0.66	0.05
S_E - S_V	234	0.28	0.14	0.97	0.00
S_F - S_V	235	0.90	0.02	0.65	0.06

can be derived (i.e. from button clicks) [43]. If the system is incorporated with other kinds of sensory information (e.g. gaze tracker, galvanic skin response), learners' state can be observed better.

Another limitation relates to the inherent characteristic of the e-learning system to serve useful in reviewing previously learned content but to be inefficient in learning from scratch. Namely, some of the participants commented that they do not feel comfortable with studying unfamiliar content *for the first time* on a computer. They expressed their preference for studying such material initially in a more conventional way (reading a textbook, taking notes etc.), which deploys the visual channel rather than the auditory one. We consider such reports to be in agreement with the findings of [16], which emphasize the permanence of visual channel in developing implicit knowledge. Nevertheless, we consider this to be a general drawback of technology-based learning systems rather than a shortcoming specific to the proposed system. Based on such observation, in addition to memory test scores, it may be beneficial to account for learners' evaluation of experience and satisfaction in assessing the overall effectiveness of the e-learning platform [44] through surveys, which are not collected in the reported experiments.

IX. CONCLUSION AND FUTURE WORK

This study investigates the possibility of increasing e-learning efficiency by integrating an automatic audio reinforcement mechanism, which estimates the improvement or deterioration of learners' memory performance. The estimator relies on an ANN and is trained with data collected in a series of experiments, where three types of content are coupled with three kinds of audio stimulus in a pre-determined arbitrary way. By this means, the most efficient content-stimulus pairs are determined. Our results indicate that there is no single best audio for reinforcing all contents. Namely, informative audio is found to be better to reinforce numerical content, whereas non-informative audio is superior for reinforcing verbal contents.

We train a basis estimator with ANN architecture for estimating the particular piece of content, which needs to be reinforced. The ANN is also designed in such a way that it adapts itself on-the-fly to each different learner using the real-time data from him/her. Our results from follow-up experiments with a reinforced e-learning system show that ANN-based audio interference is beneficial in the short term (i.e. subsequent to the learning session) for all three types of contents. In addition, triggering audio constantly (i.e. at all cards) is found to cause a decrease in performance, rather than an increase, which indicates audio reinforcement should be handled with care.

As future work, we would like to increase the data in amount, variation and modality. In addition, we plan to carry out user surveys with open-ended questions and ratings after the experiments. Moreover, there is also a significant potential for improvement in facial video images [45] and biological data (e.g. electrodermal activity). A more fundamental future work relates to the strategy of how the estimation result is blended into the e-learning system. Namely, the current method is based on passive audio, which can be replaced with other sorts of instructions or interventions. As for the former, it would be interesting to modify the e-learning system to incorporate active audio (i.e. the learner produces the written text) [34]. Regarding the former, various alternatives including avatars, prizes, levels etc. can be considered.

APPENDIX A
E-LEARNING SOFTWARE

In experiments, we used the free and open-source software "Anki" [33], which is a spaced-repetition virtual flashcard program. Anki has a large user community, which maintains a variety of shared decks aimed at rote learning of numerous areas/subjects such as foreign language vocabulary, humanities (e.g. historical dates, geographical facts, law) and science (e.g. anatomical facts, formulas, equations) [46].

Analogous to physical flashcards, each virtual flashcard has a "front" face and a "back" face. The front face contains a query and the back face contains its correspondence. Figure 6 gives a sample sequence of snapshots from Anki, where the task is to memorize country-capital associations. In this example, the query is a country (see Figure 6-(a)) and its correspondence is the capital of this country (see Figure 6-(b)). The query on the front can be considered as a "question" and its correspondence on the back can be considered as the "answer". Thus, we call the front as Q-face and the back as A-face.

Note that, one can present flashcards in different languages and also customize the language of the graphical user interface (GUI, e.g. menu, buttons etc.). In our experiments, in order to make the learning environment easy to use for our participants, the flashcards and the GUI were presented in their mother tongue (i.e. Japanese). However, in Figure 6, in order to increase accessibility, we translate the interface and the learning material (i.e. Q- and A-faces) into English.

The learner interacts with the e-learning software using the buttons appearing at the bottom of the screen (see Figure 6) and the time course of this interaction is as follows. First, the

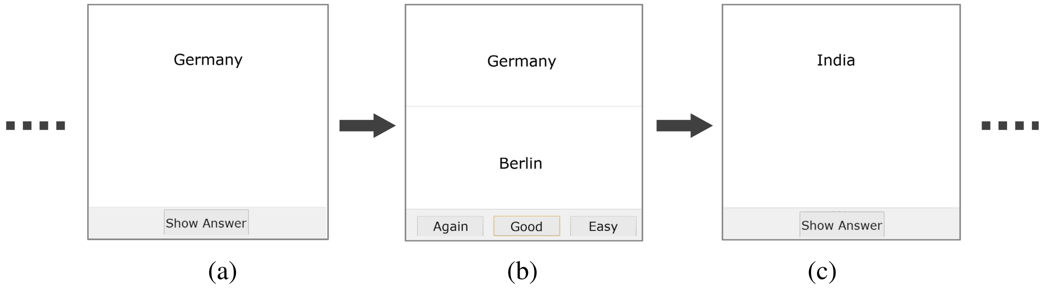


FIGURE 6. A sample sequence of snapshots from the e-learning software. (a) Front and (b) back faces of a sample flashcard. (c) Front face of the next flashcard.

learner is exposed to the Q-face of a flashcard and takes some time to recall the information on the A-face. When he/she feels ready, he/she can "flip" the flashcard by pressing the "Show Answer" button. After disclosing the A-face, the learner can confirm whether he/she recalled correctly or not. If he/she failed to recall the answer correctly or did not recall it at all, he/she can take some time to watch the A-face for registering (i.e. memorizing) this information. When the learner feels ready, he/she can proceed to the next flashcard by evaluating the difficulty of the current flashcard by pressing one of three buttons of "Again", "Good" or "Easy". Note that we did not specify any criteria or rules to the participants regarding which button to choose and let them use their own judgment.

We term watching the Q-face and then the A-face of a flashcard as a *viewing* of that flashcard (e.g. Figures 6-(a) and (b)). In the experiments, we did not set a time limit for a single viewing or watching Q- or A-faces per se. However, we decided to let the participants study a group of 12 flashcards, called a *deck*, for a (predetermined) duration of 4 minutes.

The duration for studying such a deck and the number of flashcards in it are decided in the light of several relevant works in the literature on memory span and capacity. According to [47] and [48], humans have a limited capacity to memorize, which is only about 7 ± 2 items at once. Since this study targets at improving the learning rate, we opt for a reasonably challenging task, which would result neither in perfect success nor in total frustration. In other words, we chose the number of flashcards in a deck in such a way that there would be a fair amount of recollection but also room for improvement. In that respect, 12 flashcards in 4 minutes is considered to be adequate.

As activity log, the e-learning software registers the interaction of the learner with the GUI in terms of *temporal*, *identifier* and *evaluation* information. Temporal information is registered in UNIX time at millisecond resolution and includes the time of prompt (i.e. the instant when the Q-face of a flashcard appears), time of flip (i.e. the instant when the learner presses the "Show Answer" button and discloses the A-face), and time of evaluation (i.e. the instant when the learner assesses his/her opinion on difficulty of a flashcard). On the other hand, identifier information refers to the 13-digit integer codes that designate the particular deck or flashcard that is being studied (i.e. displayed) at a given time instant. Finally, the evaluation information is again an integer code (1 $\sim$ 3) of the chosen button among "Again", "Good" or "Easy".

Note that each "deck" is associated with a single "content" type. In other words, we do not mix different kinds of learning material in the same deck. Moreover, an assignment given to the participant at once is termed as "task". In that respect, the three terms of "deck", "content" and "task" are sometimes used interchangeably, if one or the other fits better to the context.

APPENDIX B
CONTENT TYPES

For investigating the effect of audio reinforcement relating to a variety of learning materials, we consider three kinds of contents, which contain different types of information as verbal and numerical. Note that the information on Q-faces of the flashcards relating to all 3 contents (tasks) are verbal, but the information on A-faces are different, i.e. either verbal or numerical. Moreover, the verbal ones have two levels of difficulty (i.e. easy and hard). As mentioned in Section IV, we denote the content with easy verbal answers with C_E, the content with hard verbal answers with C_H and the content with numerical answers with C_N.

The two verbal contents C_E and C_H contain country-capital associations (the country appears on the Q-face and its capital appears on the A-face, see Figure 6). In particular, we consider difficulty to be based on *de facto* properties admissible to generic e-learning system users. In that respect, we judge the difficulty of a flashcard based on the level of "expected familiarity" of a common learner with the information on its A-face (see [49], [50] for details.)

The numerical content C_N involves chemical element-atomic number association, where the chemical element appears on the Q-face and its atomic number appears on the A-face. Since the content belongs to a specific field, in which none of the participants reported any dedicated skill, education or experience, we assume the learners to be unfamiliar with C_N. In addition, we assume that all chemical element-atomic number associations have the same difficulty level (i.e. there are no easier or harder pairs) and that they are all as unfamiliar to the participants as any item in C_H (i.e. the information is equally fresh/unaccustomed).

However, in C_N they need to recall a number, whereas in C_H they need to recall a word. Although the participants have no prior information or familiarity with any of the items in those, numbers are expected to be easier to register than words, since they have a more rigid structure (i.e. in our case, at most 3 digits), whereas the words are quite different in the number of characters or pronunciation difficulty. Thus, C_H is considered to have a higher degree of freedom than C_N. Thus, we recognize that familiarity is not the only element determining difficulty etc. It is also worth mentioning that numerical information is suggested to be governed by particular schemes in the cognitive registration process as pointed out by [51] and [52]. In that respect, we examine participants' performance concerning various combinations of content types and stimuli. In the next section, we elaborate on the different sorts of stimuli deployed in the experiments.

APPENDIX C
STIMULUS TYPES

"Stimulus" refers to the modality of the signal, which is used to deliver the learning task, and is known to make a distinguishing effect on the registration rate [53]. In our experiments, we used two basic kinds, which are visual and auditory. In particular, we delivered Q-faces solely using visual information for all contents and all e-learning system configurations (baseline and reinforced). However, after flipping the flashcard, depending on system configuration, the participants received the information on A-face either only visually or audio-visually. Note that after flipping the flashcard, the entire information of the A-face are delivered to the participant, which embraces also the information on the Q-face (see Figure 6-(b)).

At exploration stage, when there is no audio incorporation, the e-learning system delivers solely visual stimuli. This baseline configuration is denoted with S_V. As audio stimuli, we considered two kinds as informative or non-informative. In particular, informative audio refers to the human readout of the information, which is already displayed visually. The configuration involving informative audio is denoted with S_A. On the other hand, non-informative audio is simply a bell sound. The configuration involving non-informative audio is denoted with S_B. Informative audio is inherently congruous with the visually displayed information, whereas non-informative is neutral (i.e. neither congruous nor incongruous).

The reason for using informative audio is the promising evidence reported by [13], [15], and [16] in improving memory registration (see also Section II). On the other hand, the reason for opting for human audio rather than generated speech is the *voice principle* [4], which suggests that the information provided through recording of human voice is recalled with a higher rate as compared to the information provided through synthesized speech (e.g. audio generated by a text-to-speech system). In that respect, we compiled a set of audio stimuli such that it contains actual human speech (see Appendix G for the details).

The reason for using non-informative audio is to enable the measurement of learners' reaction to the absence/presence of audio rather than registering the enclosed information. Namely, in case the learner recovers a possible disengagement due to breaking of the silence without actually registering any information through the audio channel, this can be detected by comparing learning rates corresponding to S_V and S_B. Note also the non-informative audio is played after a certain brief period upon disclosing the A-face, which may help the learner to be aware of his/her time use.

Note that in exploration stage, when a deck is delivered through stimulus types S_A and S_B, half of the flashcards in that deck are delivered visually, whereas the other half are presented audio-visually. The flashcards to be delivered visually or audio-visually are determined before the session in a completely arbitrary way and all participants are given the same composition.

The reason for half-and-half mixing of visual-only and audio reinforced presentation (as opposed to total audio-reinforcement) is an inspiration from the *distinctiveness account* relating to the *production phenomenon* [34], [54]. In the context of memory registration, *production* refers to repeating of the perceived information through various ways, such as uttering, mouthing, whispering, spelling, hearing, writing, typing, or even singing. One of the first studies on the relation between recollection of material that is simply visually presented as compared to material *produced*, was carried out by [54]. They observed that the material produced is better recalled than the material merely viewed. They called this phenomenon the "generation effect", which led to a vast amount of studies on priming strategies or manipulations of subjects' behavior or stimuli in relation to memory retention.

On the other hand, a specific case of the generation effect is the *vocal production effect* (i.e. due to uttering). Here, production refers to uttering a word aloud during study (rather than to simply reading it silently) and is known to improve explicit memory [35]. In addition, common vocal production is shown to be superior against all of the above-mentioned different ways of production [55], [56], although also some alternatives such as mouthing and whispering are found to have a positive effect, though not as strong [57]. The advantage of vocal production over those other means is considered to be due to the presence of both articulation and audition components in speech, whereas audition is absent in mouthing and extremely limited in whispering. In addition to the context-free studies exploring the basics of the production effect, there were also studies specifically situtated in education and learning [58], which showed that production is a viable encoding strategy for educational material, due to a lasting effect and extension beyond isolated words (i.e. it applies also to word pairs and sentences) [59].

The mechanisms, that are hypothesized to be underlying the vocal production effect are termed in the literature as *accounts*. Several popular accounts include decision-based account, memory-based account [60], strength account [61], distinctiveness account [34] and attributional account [62]. In this study, we focus on the distinctiveness aspect. According to [35], when subjects are exposed to both generated stimuli (e.g. audio attached) and not-generated stimuli (e.g. silent) in the same session, the subsequent memorability of the generated stimuli is seen to improve, since distinct stimuli are likely to be better encoded in memory. However, when distinctiveness is undermined (e.g. by constant audio attachment), this effect vanishes.

Our study does not involve the active involvement of the participant, i.e. uttering of the visually displayed information by himself/herself. Instead we address a passive involvement, i.e. simply being exposed to a readout of the visually displayed information (by another person). Nevertheless, we still believe that distinctiveness may help in improving recollection. In other words, we consider routine (constant) triggering of audio may cause the participants to neglect the additional stimulus. Therefore, instead of attaching audio to all flashcards in a deck, we arbitrarily choose half of the flashcards in that deck and attach them audio clips (audio-visual), whereas the other half is delivered in a visual-only manner. Note that the validity of this contemplation is eventually confirmed in Section VII.

APPENDIX D
COUPLING OF CONTENT TYPES AND STIMULUS TYPES

At both exploration stage and verification stage, the participants were first given a dummy session such that they familiarize themselves with the e-learning software and ask questions, if they have any. The data collected during this first session is not subject to any analysis.

As mentioned in Appendix B, the three content types of easy verbal C_E, hard verbal C_H, and numerical C_N are used in the experiments. Moreover, these content types are common at exploration and verification stages. However, the stimulus types and their triggering schemes are different between exploration and verification stages.

Namely, at exploration stage three stimulus types are investigated as (i) visual only S_V, (ii) a combination of visual and informative audio S_A, and (iii) a combination of visual and non-informative audio S_B, whereas at the verification stage two stimulus types are studied as automatic estimation scheme S_E and full support S_F.

Moreover, as mentioned in Appendix C, at exploration stage, when a deck is delivered through stimulus types S_A and S_B, half of the flashcards in that deck are delivered visually, whereas the other half are presented audio-visually. In addition, at verification stage, the cards to be delivered audio-visually (i.e. with audio reinforcement, denoted with S_E) are determined based on the estimations of the ANN, whereas full support S_F triggers audio reinforcement every time an A-face is disclosed.

In experiments, each participant carried out one session for each coupling of content type and stimulus type investigated at that stage (e.g. $C_H + S_V$ at exploration stage or $C_E + S_F$ at verification stage). At the exploration stage, all possible couplings amount to nine learning sessions (see Figure 2-(a)), whereas at the verification stage they amount to six learning sessions (see Figure 2-(b)).

The order, in which the participants are exposed to those couplings, is adjusted carefully. Namely, if the exact same order were used for each participant, there could have been a correlation between memory performance and that order (e.g. suffering from fatigue at couplings which are presented later than others). For removing any doubt of such bias, we randomized the order of couplings of content type and stimulus type. Namely, we presented the couplings in an such order that there will be no bias in memory performance due to fatigue.

As mentioned in Appendix A, we did not have a time limit for viewing a single flashcard or maximum number of views per flashcard. But we set a time limit of 4 minutes for studying a single deck. In addition, the participants were allowed to abort in advance (i.e. earlier than 4 minutes), should they think that they are ready to take a memory test. Moreover, in answering memory tests, we did not set any time limit but it mostly took 1 or 2 minutes for the participants to finish these tests.

APPENDIX E
MEMORY TESTS AND ASSESSMENT OF LEARNING PERFORMANCE

In literature, the assessment of memory retention is mostly based on free recall, explicit recognition test, source identification or speeded reading test. Free recall refers to the subject's listing of the items in the presented task. Explicit recognition refers to recognition of the items among a set involving also distracters. Source identification refers to attributing the items to one of the several (usually two) sets of items (tasks). Speeded reading is the subjects' reading into a microphone a mixed list of items, which is then analyzed to detect the changes in his/her reading pattern.

For our experiment scenario, we consider a similar strategy to paired associate recall to be the most adequate. Specifically, we present the information on the Q-face and require the participant to write down the corresponding information on the A-face. In evaluating the memory tests, we register the performance relating to a certain flashcard as a 1, if the participant recalls the information on its A-face successfully, and otherwise as a 0.

Concerning a certain learning session, we made three tests, namely *once before* and *two times after* the session. We call these memory tests *prior* test, *short-term* test and *mid-term* test, respectively. Note that in all three tests concerning a deck, the Q-faces of the flashcards in that deck are provided but their sequence (order) is altered to avoid any bias due to visual memory (i.e. remembering locations of words).

The purpose of prior test is to determine the pieces of the learning material, the learners *readily possess*, so as to omit the interactions with those flashcards. Namely, if a learner correctly answers a question in the test performed before the e-learning session started, no gained/retained or lost knowledge (i.e. no learning and no forgetting) is expected relating to that flashcard. Therefore, in reporting the results on memory performance as well as designing the estimator model, we ignored the interactions with such flashcards.

The short-term test is given immediately after the learning session and its purpose is to assess the amount of *gained* information during that session. With the memory test scores concerning the prior test and short-term test, the amount of information *gained* I_G in a given session can be computed as the average of the difference between these two. Let d' be the set of all flashcards f from deck d, which are found to be *not readily known* in the prior test,

$$d' = \left\{f, \quad t_p(f) = 0\right\}, \tag{6}$$

where $t_p(f)$ is the score of flashcard f in the prior test. Then, the average amount of gained information concerning deck d, $I_G(d)$ can be computed as

$$I_G(d) = \frac{\sum\limits_{\forall f \in d'} t_s(f)}{\#(d')}, \tag{7}$$

where $t_s(f)$ is the score of flashcard f in the short-term test.

Mid-term test is given approximately 5 minutes after the learning session and its purpose is to measure the amount of *retained* information I_R. The time lag between the short-term and mid-term tests is determined based on the studies on the Ebbinghaus Forgetting Curve. In particular, Murre and Dros confirm that people start forgetting what they learned immediately after the lesson [63]. In addition, the ability to recall is empirically shown to decrease rapidly from the end of the lesson until some point within a day, and the rate of forgetting downscales with time. In that respect, the short time window following the lesson is considered to be the most essential period in understanding the human memory process. Clearly, there is a complex interplay between the duration of the learning session, the amount of learning material and the rate of forgetting. Taking these into consideration, we contemplated that 5 minutes is a fair duration for observing a change in memory retention.

Specifically, the average amount of *retained* information I_R is computed as

$$I_R(d) = \frac{\sum\limits_{\forall f \in d'} t_m(f)}{\#(d')}, \tag{8}$$

where $t_m(f)$ is the score of flashcard f in the mid-term test.

Note that I_G and I_R are expected to be non-negative, since participants' knowledge is expected to increase or remain the same in relation to their knowledge before the learning session. Also, I_G is expected to be larger than or equal to I_R. Note also that in presenting the memory performance results in Sections V and VII, we consider I_G, and I_R as computed *over all participants*.

TABLE 9. (a) Statistics and (b) ANOVA and effect size concerning information retained I_R for varying content types at the exploration step.

(a)

Task types	#	μ	ϵ
C_E	238	0.83	0.02
C_H	319	0.43	0.03
C_N	309	0.71	0.03

(b)

Task types	#	p	d
$C_E - C_H$	557	$< \mathbf{10^{-2}}$	0.90
$C_E - C_N$	547	$< \mathbf{10^{-2}}$	0.30
$C_H - C_N$	628	$< \mathbf{10^{-2}}$	0.58

TABLE 10. (a) Statistics and (b) ANOVA and effect size concerning information retained I_R for varying stimulus types at the exploration step.

(a)

Stimulus types	#	μ	ϵ
S_A	299	0.68	0.03
S_B	277	0.66	0.03
S_V	290	0.58	0.03

(b)

Stimulus types	#	p	d
$S_A - S_B$	576	0.64	0.04
$S_A - S_V$	589	**0.02**	0.19
$S_B - S_V$	567	0.07	0.15

In addition to descriptive statistics, we present also the standard error ϵ and effect size d for giving an insight into the variation between different users. The standard error ϵ helps us to understand whether the number of data points is sufficient to identify in a reliable way the position of the mean μ. Specifically, provided that $\epsilon \ll \mu$, the dispersion of sample means around the population mean is rather small, and thus the number of data points can be considered to be sufficient for identifying it. On the other hand, the effect size d compares the difference between the mean values $\mu_{1,2}$ and variances $\sigma^2_{1,2}$ of two distributions. Cohen defines effect size d as

$$d = \frac{\mu_1 - \mu_2}{s}, \tag{9}$$

where the term in the denominator is

$$s = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}. \tag{10}$$

Here, $n_{1,2}$ are the sizes of the two populations. In that sense, d points out to how different the mean values are regardless of the amount of data.

TABLE 11. (a) Statistics and (b) ANOVA and effect size concerning information retained I_R for each pair of task type and stimulus type at the exploration step.

(a)

	C_E			C_H			C_N		
	#	μ	ϵ	#	μ	ϵ	#	μ	ϵ
S_A	91	**0.90**	0.03	106	0.42	0.05	102	**0.75**	0.04
S_B	66	0.86	0.04	107	**0.50**	0.05	104	0.68	0.05
S_V	81	0.73	0.05	106	0.37	0.05	103	0.69	0.05

(b)

	C_E			C_H			C_N		
	#	p	d	#	p	d	#	p	d
$S_A - S_B$	157	0.47	0.12	213	0.19	0.18	206	0.32	0.14
$S_A - S_V$	172	$< \mathbf{10^{-2}}$	0.46	212	0.48	0.10	205	0.38	0.12
$S_B - S_V$	147	**0.05**	0.33	213	**0.04**	0.28	207	0.92	0.01

APPENDIX F
RESULTS RELATING TO INFORMATION RETAINED I_R AT THE EXPLORATION STEP

As mentioned in Section V, the tendencies in information gained I_G and information retained I_R are quite parallel.

Comparing Table 1-(a) and Table 9-(a) below, one may see that the visual-only S_V case is surpassed by informative audio S_A and non-informative audio S_B for C_E. Moreover, comparing Table 1-(b) and Table 9-(b), it is understood that the significance relation observed in the short-term test is valid in the midterm-test with similar effect sizes.

Comparing Table 2-(a) and Table 10-(a) below, it is seen that the higher memory retention rates of S_A and S_B than S_V persist in the mid-term test. Regarding significance, $S_A - S_B$ pair is found not to be significantly different, whereas $S_A - S_V$ has again a statistically significant difference. On the other hand, $S_B - S_V$ pair is no more significant in the mid-term test, although the relating p value is quite close the generally accepted limit ($p = 0.7 > 0.5$).

Comparing Table 4-(a) and Table 11-(a) below, it is seen that the memory retention rates concerning all pairs decrease slightly from the short-term test to the midterm-test. Nevertheless, the rates of decrease are quite similar, thus the pairwise relations are almost always sustained. Regarding the significance of the differences, $S_A - S_B$ pair concerning C_E is still associated with significantly different memory scores. Also, for the $S_B - S_V$ pair concerning C_H we achieve significance, meaning that non-informative audio S_B is more beneficial in enhancing I_R in addition to I_G. However, $S_A - S_B$ pair concerning C_H, which was borderline significant at short-term test ($p = 0.5$), is no more significantly different at mid-term test ($p = 0.19$). For C_N, although informative audio S_A leads to the highest performance in I_R, the difference is not statistically significant ($p = 0.32$).

APPENDIX G
COMPILING OF INFORMATIVE AUDIO

According to the *voice principle* [4], human audio (i.e. recording of human voice) is more effective in recalling information than synthesized speech (e.g. audio generated by a text-to-speech system). In that respect, for the informative audio stimuli used in our experiments, we asked a Japanese mother-tongue speaker (henceforth simply referred to as the *speaker*) to utter 504 words and 118 numbers (relating countries, capitals, chemical elements and atomic numbers).

To avoid inducing fatigue in the speaker, we divided the recording process into several sessions. Specifically, we implemented 6 sessions of 15 minutes such that the speaker does not suffer from considerable fatigue. In addition, we designed a framework, which takes charge of displaying images (of text) and recording their utterances. Namely, we displayed to the speaker a sequence of images on the screen of a notebook PC [64]. Each image contained a text, which can be either a word or a number, and a single sequence contained no more than 100 images. The number of images is adjusted depending on the number of words/numbers to be uttered/recorded and varied between 59 and 100. The speaker did not need to press a button or a key to proceed. In other words, the images (i.e. words or numbers) advanced automatically. In addition, the text was displayed on the screen for a sufficient amount of time (i.e. 4 seconds) such that the speaker could read it aloud without hurry and he could also rest briefly between successive images.

Since we wanted to assure that utterances of the numbers are "flat", they are not displayed in any specific order (i.e. incrementing or decrementing), but rather in a random manner, Namely, if the speaker expects the next text to utter to be the next number (i.e. integer), he could have displayed a recurrent voice pattern formed by a series of regular rises and falls in intensity. Although such a sequence would not sound unnatural in context, once the recordings of each number are extracted, one might perceive that it is not a standalone recording, but it rather belongs to a longer (counting) sequence. Since we wanted to avoid this effect, we displayed the numbers in random order.

Together with the sequence of images, an audio recording was initiated. A single audio clip is recorded for each image (i.e. text). Note that both the sequence of images and the audio recording are executed through programs implemented in Python. In that respect, we could precisely register the time correspondence between the display period of an image and the utterance of the corresponding word, which facilitates a temporally accurate post-processing of the audio.

Note that the afore-mentioned display duration of 4 seconds is abundant to read the text aloud, and thus, there are silent periods preceding and succeeding the utterances. In that respect, we post-processed the recorded audio clips by splitting them into individual utterances and removing the silent segments. To that end, we defined a minimum length for silent segments (in milliseconds) and an upper bound for *how quiet is silent* (in decibels relative to full scale) and any section of the audio, which satisfies these two criteria, is considered to be a silent period and discarded. After the audio clips were segmented and post-processed, the speaker listened and confirmed most of them. However, he determined several utterances (i.e. pronunciations) as unnatural or problematic, and thus, we held an extra recording session to re-record those utterances.

As mentioned in Appendix A, the audio clips are attached to A-faces and the settings of the e-learning software are adjusted such that once the audio stimuli are triggered, each audio clip is played twice. For instance, if the learner is studying of country-capital associations, once audio stimuli is triggered, he/she hears the *country name, capital name, country name, capital name*.

REFERENCES

[1] S. D. Sorden, "The cognitive theory of multimedia learning," in *Handbook of 782 Educational Theories*, vol. 1. NC, USA: IAP Information Age Publishing, 2012, pp. 1–22.

[2] R. C. Atkinson and R. M. Shiffrin, "Human memory: A proposed system and its control processes," in *Psychology of Learning and Motivation*, vol. 2. Amsterdam, The Netherlands: Elsevier, 1968, pp. 89–195.

[3] R. E. Mayer, "Using multimedia for e-learning," *J. Comput. Assist. Learn.*, vol. 33, no. 5, pp. 403–423, 2017.

[4] R. E. Mayer, "Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles," in *The Cambridge Handbook of Multimedia Learning*, vol. 16. Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 345–370.

[5] N. Refat, M. A. Rahman, A. T. Asyhari, I. F. Kurniawan, M. Z. A. Bhuiyan, and H. Kassim, "Interactive learning experience-driven smart communications networks for cognitive load management in grammar learning context," *IEEE Access*, vol. 7, pp. 64545–64557, 2019.

[6] K. H. Cheong, J. M. Koh, D. J. Yeo, Z. X. Tan, B. O. E. Boo, and G. Y. Lee, "Paradoxical simulations to enhance education in mathematics," *IEEE Access*, vol. 7, pp. 17941–17950, 2019.

[7] R. Khoii and S. Sharififar, "Memorization versus semantic mapping in L2 vocabulary acquisition," *ELT J.*, vol. 67, no. 2, pp. 199–209, Apr. 2013.

[8] N. Hara, "Student distress in a web-based distance education course," *Inf., Commun. Soc.*, vol. 3, no. 4, pp. 557–579, 2000.

[9] D. Zhang, J. L. Zhao, L. Zhou, and J. F. Nunamaker Jr., "Can e-learning replace classroom learning?" *Commun. ACM*, vol. 47, no. 5, pp. 75–79, 2004.

[10] C. G. Penney, "Modality effects and the structure of short-term verbal memory," *Memory Cognition*, vol. 17, no. 4, pp. 398–422, Jul. 1989.

[11] F. A. Inan, S. M. Crooks, J. Cheon, F. Ari, R. Flores, M. Kurucay, and D. Paniukov, "The reverse modality effect: Examining student learning from interactive computer-based instruction," *Brit. J. Educ. Technol.*, vol. 46, no. 1, pp. 123–130, Jan. 2015.

[12] V. A. Thompson and A. Paivio, "Memory for pictures and sounds: Independence of auditory and visual codes," *Can. J. Experim. Psychol./Revue Canadienne De Psychologie Expérimentale*, vol. 48, no. 3, pp. 380–398, Sep. 1994.

[13] S. Tindall-Ford, P. Chandler, and J. Sweller, "When two sensory modes are better than one," *J. Experim. Psychol., Appl.*, vol. 3, no. 4, pp. 257–287, Dec. 1997.

[14] J. Heikkilä, K. Alho, H. Hyvönen, and K. Tiippana, "Audiovisual semantic congruency during encoding enhances memory performance," *Experim. Psychol.*, vol. 62, no. 2, pp. 123–130, Jan. 2015.

[15] R. Moreno and R. E. Mayer, "Verbal redundancy in multimedia learning: When reading helps listening," *J. Educ. Psychol.*, vol. 94, no. 1, pp. 156–163, Mar. 2002.

[16] K. M. Kim and A. Godfroid, "Should we listen or read? Modality effects in implicit and explicit knowledge," *Modern Lang. J.*, vol. 103, no. 3, pp. 648–664, Aug. 2019.

[17] R. Kaplan-Rakowski and B. Loranc, "The impact of verbal and nonverbal auditory resources on explicit foreign language vocabulary learning," *System*, vol. 85, Oct. 2019, Art. no. 102114.

[18] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020.

[19] H. Gardner, "A reply to Perry D. Klein's 'multiplying the problems of intelligence by eight,'" *Can. J. Educ./Revue Canadienne De L'éducation*, vol. 23, no. 1, pp. 96–102, 1998.

[20] D. A. Kolb, R. E. Boyatzis, and C. Mainemelis, "Experiential learning theory: Previous research and new directions," in *Perspectives on Thinking, Learning, and Cognitive Styles*. Evanston, IL, USA: Routledge, 2014, pp. 227–248.

[21] R. Felder and L. Silverman, "Learning and teaching styles in engineering education," *Education*, vol. 78, pp. 674–681, Jan. 2002.

[22] M. Raleiras, A. H. Nabizadeh, and F. A. Costa, "Automatic learning styles prediction: A survey of the state-of-the-art (2006–2021)," *J. Comput. Educ.*, vol. 9, no. 4, pp. 587–679, Dec. 2022.

[23] B. Pardamean, T. Suparyanto, T. W. Cenggoro, D. Sudigyo, and A. Anugrahana, "AI-based learning style prediction in online learning for primary education," *IEEE Access*, vol. 10, pp. 35725–35735, 2022.

[24] F. Rasheed and A. Wahid, "Learning style detection in e-learning systems using machine learning techniques," *Exp. Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114774.

[25] Y. Gambo and M. Shakir, "An artificial neural network (ANN)-based learning agent for classifying learning styles in self-regulated smart learning environment," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 16, no. 18, pp. 185–199, 2021.

[26] H. Zhang, T. Huang, S. Liu, H. Yin, J. Li, H. Yang, and Y. Xia, "A learning style classification approach based on deep belief network for large-scale online education," *J. Cloud Comput.*, vol. 9, no. 1, pp. 1–17, Dec. 2020.

[27] Y. Zhang, R. An, S. Liu, J. Cui, and X. Shang, "Predicting and understanding student learning performance using multi-source sparse attention convolutional neural networks," *IEEE Trans. Big Data*, vol. 9, no. 1, pp. 118–132, Feb. 2023.

[28] A. Polyzou and G. Karypis, "Feature extraction for next-term prediction of poor student performance," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 237–248, Apr. 2019.

[29] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103676.

[30] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Hum. Behav.*, vol. 104, Mar. 2020, Art. no. 106189.

[31] S. Aydogdu, "Predicting student final performance using artificial neural networks in online learning environments," *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 1913–1927, May 2020.

[32] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm," *Comput. Hum. Behav.*, vol. 107, Jun. 2020, Art. no. 105584.

[33] D. Elmes. (2021). *ANKI—Friendly, Intelligent Flashcards*. Accessed: May 8, 2022. [Online]. Available: https://ankiweb.net/about

[34] J. D. Ozubko and C. M. MacLeod, "The production effect in memory: Evidence that distinctiveness underlies the benefit," *J. Experim. Psychol., Learn., Memory, Cognition*, vol. 36, no. 6, pp. 1543–1547, 2010.

[35] C. M. MacLeod, N. Gopie, K. L. Hourihan, K. R. Neary, and J. D. Ozubko, "The production effect: Delineation of a phenomenon," *J. Experim. Psychol., Learn., Memory, Cognition*, vol. 36, no. 3, pp. 671–685, 2010.

[36] Y. Bi, "Dual coding of knowledge in the human brain," *Trends Cognit. Sci.*, vol. 25, no. 10, pp. 883–895, Oct. 2021.

[37] D. Harris and S. Harris, *Digital Design and Computer Architecture*. San Mateo, CA, USA: Morgan Kaufmann, 2010.

[38] P. Supitayakul, "Online resource: Data set and source code for for ANN-based audio reinforcement for computer assisted rote learning," Okayama Univ., Okayama, Japan, Tech. Rep., 2022, doi: 10.6084/m9.figshare.19429796.

[39] F. Pukelsheim, "The three sigma rule," *Amer. Statist.*, vol. 48, no. 2, pp. 88–91, 1994.

[40] S. Bertsch, B. J. Pesta, R. Wiscott, and M. A. McDaniel, "The generation effect: A meta-analytic review," *Memory Cognition*, vol. 35, no. 2, pp. 201–210, Mar. 2007.

[41] J. W. Schooler, J. Smallwood, K. Christoff, T. C. Handy, E. D. Reichle, and M. A. Sayette, "Meta-awareness, perceptual decoupling and the wandering mind," *Trends Cognit. Sci.*, vol. 15, pp. 319–326, Jun. 2011.

[42] D. Smilek, J. S. A. Carriere, and J. A. Cheyne, "Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering," *Psychol. Sci.*, vol. 21, no. 6, pp. 786–789, Jun. 2010.

[43] Z. Yucel, P. Supitayakul, A. Monden, and P. Leelaprute, "Identification of behavioral variables for efficient representation of difficulty in vocabulary learning systems," *Int. J. Learn. Technol. Learn. Environ.*, vol. 3, no. 1, pp. 51–60, 2020.

[44] B. Alojaiman, "Toward selection of trustworthy and efficient e-learning platform," *IEEE Access*, vol. 9, pp. 133889–133901, 2021.

[45] Z. Yucel, S. Koyama, A. Monden, and M. Sasakura, "Estimating level of engagement from ocular landmarks," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 16, pp. 1527–1539, Oct. 2020.

[46] Anki User Community. (2021). *Shared Decks*. Accessed: May 27, 2022. [Online]. Available: https://ankiweb.net/shared/decks/

[47] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, no. 2, p. 81, 1956.

[48] M. Manoochehri, "Up to the magical number seven: An evolutionary perspective on the capacity of short term memory," *Heliyon*, vol. 7, no. 5, May 2021, Art. no. e06955.

[49] Z. Yucel, P. Supitayakul, A. Monden, and P. Leelaprute, "An algorithm for automatic collation of vocabulary decks based on word frequency," *IEICE Trans. Inf. Syst.*, vol. 103, no. 8, pp. 1865–1874, 2020.

[50] K. Furuya, Z. Yucel, P. Supitayakul, A. Monden, and P. Leelaprute, "Exploring the limits of an RBSC-based approach in solving the subset selection problem," in *Proc. EPiC Ser. Comput.*, 2021, pp. 1–11.

[51] S. Dehaene, *The Number Sense: How the Mind Creates Mathematics*. Oxford, U.K.: OUP, 2011.

[52] A. J. Baroody, N. P. Bajwa, and M. Eiland, "Why can't Johnny remember the basic facts?" *Develop. Disabilities Res. Rev.*, vol. 15, no. 1, pp. 69–79, 2009.

[53] R. Pillai and A. Yathiraj, "Auditory, visual and auditory-visual memory and sequencing performance in typically developing children," *Int. J. Pediatric Otorhinolaryngol.*, vol. 100, pp. 23–34, Sep. 2017.

[54] N. J. Slamecka and P. Graf, "The generation effect: Delineation of a phenomenon," *J. Experim. Psychol., Hum. Learn. Memory*, vol. 4, no. 6, p. 592, 1978.

[55] S. E. Gathercole and M. A. Conway, "Exploring long-term modality effects: Vocalization leads to best retention," *Memory Cognition*, vol. 16, no. 2, pp. 110–119, Mar. 1988.

[56] C. K. Quinlan and T. L. Taylor, "Mechanisms underlying the production effect for singing," *Can. J. Experim. Psychol./Revue Canadienne De Psychologie Expérimentale*, vol. 73, no. 4, pp. 254–264, Dec. 2019.

[57] N. D. Forrin, C. M. MacLeod, and J. D. Ozubko, "Widening the boundaries of the production effect," *Memory Cognition*, vol. 40, no. 7, pp. 1046–1055, Oct. 2012.

[58] J. D. Ozubko, K. L. Hourihan, and C. M. MacLeod, "Production benefits learning: The production effect endures and improves memory for text," *Memory*, vol. 20, no. 7, pp. 717–727, Oct. 2012.

[59] M. Icht and Y. Mama, "The effect of vocal production on vocabulary learning in a second language," *Lang. Teach. Res.*, vol. 26, no. 1, pp. 79–98, 2019.

[60] D. P. Mccabe, A. G. Presmanes, C. L. Robertson, and A. D. Smith, "Item-specific processing reduces false memories," *Psychonomic Bull. Rev.*, vol. 11, no. 6, pp. 1074–1079, Dec. 2004.

[61] J. D. Ozubko, J. Major, and C. M. MacLeod, "Remembered study mode: Support for the distinctiveness account of the production effect," *Memory*, vol. 22, no. 5, pp. 509–524, Jul. 2014.

[62] G. E. Bodner and A. Taikh, "Reassessing the basis of the production effect in memory," *J. Experim. Psychol., Learn., Memory, Cognition*, vol. 38, no. 6, pp. 1711–1719, 2012.

[63] J. M. J. Murre and J. Dros, "Replication and analysis of Ebbinghaus' forgetting curve," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0120644.

[64] P. Supitayakul, "Online resource: Source code for displaying visual stimuli and recording audio," Okayama Univ., Okayama, Japan, Tech. Rep., 2021, doi: 10.6084/m9.figshare.19429745.v1.

PARISA SUPITAYAKUL received the B.E. degree in software and knowledge engineering from Kasetsart University, Thailand, in 2018, and the M.S. degree in information science from Okayama University, Japan, in 2021, where she is currently pursuing the Ph.D. degree with the Division of Industrial Innovation Sciences, Graduate School of Natural Science and Technology. Her research interests include human behavior analysis, in particular learning, memory, and attention.

ZEYNEP YÜCEL (Member, IEEE) received the B.S. degree in electrical engineering from Bogazici University, Turkey, and the M.S. and Ph.D. degrees in electrical engineering from Bilkent University, Turkey, in 2005 and 2010, respectively. She was a Postdoctoral Researcher with ATR Labs, Kyoto, Japan, for five years, before being awarded a JSPS Fellowship, in 2016. She is currently an Associate Professor with Okayama University, Japan. Her research interests include robotics, signal processing, computer vision, and pattern recognition.

AKITO MONDEN (Member, IEEE) received the B.E. degree in electrical engineering from Nagoya University, in 1994, and the M.E. and D.E. degrees in information science from the Nara Institute of Science and Technology (NAIST), in 1996 and 1998, respectively. He is currently a Professor with the Graduate School of Natural Science and Technology, Okayama University, Japan. His research interests include software measurement and analytics, and software security and protection. He is a member of ACM, IEICE, IPSJ, and JSSST.