Adapting Existing Games for Education using Speech Recognition

by

Carrie Jun Cai

M.A. Education, Stanford University (2008)
B.A. Human Biology, Stanford University (2008)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

Author . . . . . .
Department of Electrical Engineering and Computer Science
May 21, 2013

Certified by . . . . . .
Stephanie Seneff
Senior Research Scientist
Thesis Supervisor

Certified by . . . . . .
Robert Miller
Associate Professor
Thesis Supervisor

Accepted by . . . . . .
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Adapting Existing Games for Education using Speech Recognition

by

Carrie Jun Cai

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Although memory exercises and arcade-style games are alike in their repetitive nature, memorization tasks like vocabulary drills tend to be mundane and tedious while arcade-style games are popular, intense and broadly addictive. The repetitive structure of arcade games suggests an opportunity to modify these well-known games for the purpose of learning. Arcade-style games like Tetris and Pac-man are often difficult to adapt for educational purposes because their fast-paced intensity and keystroke-heavy nature leave little room for simultaneous practice of other skills. Incorporating spoken language technology could make it possible for users to learn as they play, keeping up with game speed through multimodal interaction. Two challenges exist in this research: first, it is unclear which learning strategy would be most effective when incorporated into an already fast-paced, mentally demanding game. Secondly, it remains difficult to augment fast-paced games with speech interaction because the frustrating effect of recognition errors highly compromises entertainment.

In this work, we designed and implemented Tetrilingo, a modified version of Tetris with speech recognition to help students practice and remember word-picture mappings. With our speech recognition prototype, we investigated the extent to which various forms of memory practice impact learning and engagement, and found that free-recall retrieval practice was less enjoyable to slower learners despite producing significant learning benefits over alternative learning strategies. Using utterances collected from learners interacting with Tetrilingo, we also evaluated several techniques to increase speech recognition accuracy in fast-paced games by leveraging game context. Results show that, because false negative recognition errors are self-perpetuating and more prevalent than false positives, relaxing the constraints of the speech recognizer towards greater leniency may enhance overall recognition performance.

Thesis Supervisor: Stephanie Seneff, Senior Research Scientist

Thesis Supervisor: Robert Miller, Associate Professor

Acknowledgments

To my advisor, Stephanie Seneff, thank you for your daily encouragement, advice, and support. To my advisor, Rob Miller, thank you for your thoughtful insights and rigorous guidance. I am a better thinker and researcher because of you both.

To my Mother, Grandfather, and family back home, thank you for your unconditional love and encouragement during this new and exciting phase of my life.

To the Spoken Language Systems group, thank you for giving me a research home. My life has been enriched by our birthday traditions and defense suit-ups, our alternative "reading groups," spontaneous music breaks, and multilingual conversations. Thank you to Ian McGraw for showing me the ropes and helping me through numerous technical challenges, from WAMI to PMM and everything in between. Thank you to Victor Zue, for inspiring me to pursue graduate studies at MIT; Jim Glass, for your valuable suggestions and calm leadership; Scott Cyphers, for working your magic on almost anything and everything breakable; Wade Shen, for taking interest and believing in my research; and Marcia Davidson, for skillfully handling logistical issues and making my life easier as a result.

To the User Interface Design group, thank you for brainstorming, creating, analyzing, evaluating, and iterating with me. My weeks would not be the same without our regular interactions at Tea and paired research. Thank you for inspiring me to cultivate my nerdy side as well as my human-loving side, and helping me realize that it is actually possible to do both.

To Amazon Mechanical Turkers, thank you for taking a risk on my long HITs, for patiently enduring false negative speech recognition errors, for having fun with Tetrilingo, and for learning so many nonce words.

To Sidney Pacific, Smallfeet, and Friday Night Dinner, thank you for giving me a home from the day I stepped foot onto the east coast. My snowy days are made warmer by your company, and my sunny days are made brighter by your willingness to take part in my crazy antics.

Contents

1 Introduction **9**

2 Related Work **13**

2.0.1 Game-based Learning 13

2.0.2 Memory Rehearsal and Retrieval Practice 15

2.0.3 Speech-enabled Games for Learning 16

3 System and User Interface Design **19**

3.1 Interface Improvements 20

3.1.1 Paper Prototyping 20

3.1.2 Digital Prototyping 22

3.2 Final Implementation 26

3.2.1 Speech Recognition Architecture 26

3.2.2 System Architecture 27

3.2.3 Game Modes 28

3.2.4 Order of Word-Picture Presentation 30

4 Data Collection **33**

4.1 Amazon Mechanical Turk 33

4.2 Data Collection Interface 34

4.2.1 Phase 1 34

4.2.2 Phase 2 35

4.2.3 Phase 3 37

4.3 Recruiting Users for Controlled Studies on Amazon Mechanical Turk 39

5 Learning Assessments **43**

5.0.1 Retrieval Practice vs. Study Practice 43

5.0.2 Free-recall vs. Multiple Choice Retrieval Practice 52

5.0.3 Conclusions from Learning Assessments 57

6 Speech Recognition Results **59**

7 Improving Speech Recognition Performance **65**

7.0.4 Dynamic vs. Static Vocabulary 65

7.0.5 Deepening N-best Hypotheses 67

7.0.6 Training on High Confidence User Utterances 69

8 Conclusion and Future Work **71**

List of Figures

1-1 Tetrilingo: a Tetris game modified for learning 10

3-1 Tetrilingo: a Tetris game modified for learning 19

3-2 Paper prototype 21

3-3 Setup for paper prototype user testing 22

3-4 Open source Tetris implementation. 22

3-5 Traditional speech recording interface 23

3-6 Modified speech recording interface 24

3-7 Horizontal chute in Tetrilingo interface 25

3-8 Tetrilingo system architecture 27

3-9 Wami javascript API 28

3-10 Tetrilingo game modes 29

3-11 Hint interface in multiple-choice retrieval mode 30

3-12 Leitner flashcard system 31

4-1 Initial Mechanical Turk HIT preview page 35

4-2 HIT game instructions 36

4-3 Initial HIT microphone test interface 36

4-4 Final HIT microphone test interface 38

4-5 HIT game tutorial 39

4-6 HIT submission page 41

5-1 Vocabulary words learned in user studies. 44

5-2 Study vs. Free-recall Retrieval conditions 45

5-3 Interface of learning evaluation 46
5-4 Study vs. Free-recall Retrieval learning outcomes 47
5-5 Comparing in-game performance to learning outcomes. 51
5-6 Free-recall Retrieval vs. Multiple-choice Retrieval conditions 53
5-7 Free-recall Retrieval vs. Multiple Choice Retrieval learning outcomes 54
5-8 User rankings of Study condition, Free-recall Retrieval condition, and Multiple-choice Retrieval condition. 55

6-1 Effect of poorly ranked words on speech recognition performance . . . 60
6-2 Speech recognition performance of Letter-to-Sound lexicons 61
6-3 Recognition performance by speaker 62
6-4 Speech anomalies by category 63

7-1 Performance of static vs. dynamic recognizer 66
7-2 Recognition accuracy at increasing N-best depths 67
7-3 False negatives vs. false positives at increasing N-best depths 68
7-4 Comparing Pronunciation Mixture Model-trained lexicons to classic Letter-to-Sound lexicons 70

Chapter 1

Introduction

Although memory exercises and arcade-style games have similarly repetitive structures, memorization tasks like vocabulary drills tend to be mundane whereas arcade games are fun, intense and broadly addictive. The repetitive structure of arcade games suggests an opportunity to modify these games for education through embedding memory rehearsal strategies. However, existing arcade games are typically difficult to modify for learning because their fast-paced nature leaves little room for simultaneous practice of other skills. Spoken language technology may offer an opportunity to overcome this challenge by enabling users to keep up with game speed through speech interaction, thereby learning as they play. In this work, I design and implement Tetrilingo, a modified version of Tetris (Figure 1-1) that is augmented with educational features and speech recognition. Using this system, I investigate techniques to adapt existing arcade games for education, and evaluate methods for improving speech recognition in this fast-paced game environment.

There are several challenges in this research. From a design standpoint, it is unknown what user interface changes are appropriate for encouraging learning amidst an already cognitively intensive and fast-paced game. Moreover, it is unclear which learning strategies commonly used in custom-made games or standard classrooms would remain effective when incorporated into existing arcade games. Although previous studies have evaluated the effectiveness of memory retrieval strategies, to the best of our knowledge there have been no studies measuring the impact of these tech-

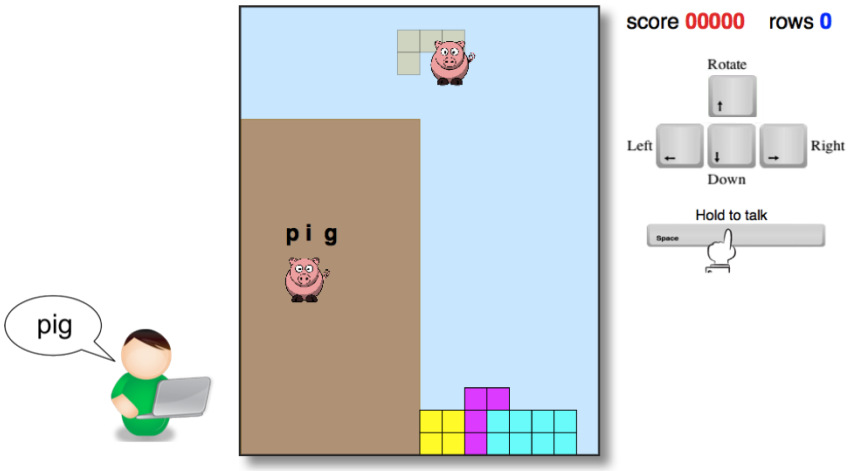


Figure 1-1: Tetrilingo: a Tetris game modified for learning. Saying the correct word unlocks block rotation.

niques within a time sensitive, arcade-style game. Because retrieval strategies exert greater cognitive effort than non-retrieval alternatives, and the interactivity of gaming introduces an additional cognitive load [19], it is unclear whether retrieval practice would remain advantageous in an intense game setting. Cognitive research suggesting somewhat independent working memory channels for visuo-spatial and phonological processing [2] gives some hope to the possibility of uniting vocabulary learning with visually demanding arcade games. Lastly, although incorporating spoken language technology could enable users to keep up with game speed, the frustrating effect of recognition errors could be detrimental to overall enjoyment. The question remains as to which types of errors are most common in such game contexts, and in what ways speech recognition performance can be enhanced without substantial re-engineering on the backend.

To address these questions, I first conduct several pilot user studies using paper and digital prototypes. Through insight from initial user studies, I identify several techniques to make learning more natural in a fast-paced game setting, and incorporate these modifications in the final implementation of Tetrilingo, the speech-augmented Tetris game. I then conduct two online user studies to investigate the extent to which different learning strategies embedded in Tetrilingo impact learning and engagement. Finally, using utterances collected from these user studies as

a speech corpus, I evaluate several techniques for enhancing speech recognition performance, all of which use in-game context to provide additional information to the recognizer.

The remainder of this thesis is organized as follows. I first present an overview of related work in Chapter 2. Chapter 3 presents the game design, system implementation, and iteration on the user interface design through feedback from pilot user testing. Chapter 4 describes the process for data collection, challenges encountered in collecting data, and modifications to data collection methods in order to overcome these challenges. In Chapter 5, I present two user studies to measure educational effectiveness and user enjoyment, and analyze the results from those studies. Lastly, motivated by the speech recognition results described in Chapter 6, I present several strategies for improving recognizer performance and evaluate their effectiveness in Chapter 7.

Chapter 2

Related Work

2.0.1 Game-based Learning

The pervasive spread of computer games has made a significant impact on game-based learning as a serious topic in the field of education. Research by Bisson and Luckner [4] has shown that fun can have a positive impact on the learning process by suspending one's social inhibitions, reducing stress, and creating a state of relaxed alertness. In particular, fun and enjoyment are central to the process of learning because they increase learner motivation. An activity is said to be intrinsically motivating if people do it "for its own sake," driven by an interest or enjoyment in the task itself as opposed to being motivated by some external reward such as money or status [22]. According to psychologists Piaget [28] and Bruner [8], intrinsically motivated play-like activities lead to deeper learning; individuals who are intrinsically engaged not only engage in the task willingly, but also tend to devote more effort to learning and will use it more in the future. In contrast, extrinsic reinforcement may sometimes degrade the quality of learning and performance [21].

Games are a potentially powerful means for learning because they embody core elements of intrinsic motivation such as challenge, fantasy, competition, and recognition [22]. In the 1980s and 1990s, Mihaly Csikszentmihalyi defined *flow* as a mental state in which a person is so absorbed in an activity that it persists purely by virtue of intrinsic motivation [26]. Today, flow is widely accepted to be one of the fundamental

reasons for gameplay. Games are highly engaging because they are simultaneously challenging and achievable to players, keeping the player in a flow state for an extended period of time [25]. The motivational effectiveness of gaming has brought a new genre of *serious games*, referring to games that are designed for some primary purpose other than pure entertainment, often for educational purposes [1].

To some, games not only balance between challenge and competency, but also offer an opportunity for players to perform before they are fully competent by offering just-in-time support to help players overcome challenges [11]. This notion of just-in-time support is in line with the educational principle of instructional scaffolding [7], defined as the support given during the learning process that is tailored to the needs of the student, with the intent of helping the student achieve his or her learning goals. According to educational theorists, support is most effective when it is in the learner's *zone of proximal development* [37], described as the space in which a learner would be able to achieve beyond what he or she could achieve alone, through receiving help by another individual. Although such principles imply that students learn best when they are empowered beyond their individual level of competence, schools ironically tend to require that students gain competence before they can perform in a particular domain. In contrast, games often offload some of the cognitive burden from the learner to the virtual world, allowing the player to begin to act with some degree of effectiveness before being really competent. Players gain competence through trial, error, and feedback, and consolidate mastery through "cycles of expertise," only to be challenged again when faced with new hurdles [11]. Indeed, many effective games are rooted in the thrill of self-challenge and near-failures. If players could only perform after demonstrating full competence, games would become predictable and lifelessly mundane.

To date, research on educational games has focused primarily on the design of custom-made learning games or elaborate extensions of adventure-style frameworks. Most adaptations of existing games emerge from turn-based models like card games [24] or from complex virtual environments [35], perhaps due to less time pressure on learners and greater amenability to structural changes in comparison to fast-paced

arcade games. However, arcade-style games such as Tetris and Pacman are advantageous in that they are logically much simpler to manipulate and have open source code bases.

Popularized in the 1970's and 1980's, arcade games (e.g. Tetris, Pac-man) have been characterized by their short levels, simple control schemes, and rapidly increasing difficulty. Originally, this was due to the arcade environment, where players rented the game until they ran out of tokens or failed at a particular level. Compared to adventure-style games, arcade games do not require much initial learning time, and moreover do not require a specific time commitment from players. Just as flashcards enable students to review vocabulary on the run, arcade games allow players to either indulge in short spurts or stay indefinitely. Today, many of the most popular arcade-style games are freely available online, due to developers creating their own implementations of such games on different platforms and in a variety of programming languages.

2.0.2 Memory Rehearsal and Retrieval Practice

The simple, repetitive nature of arcade-style games makes them natural environments for embedding a form of repetition that aids in the retention of memories. This process, known as memory rehearsal [13], lies at the core of flashcard use in studying. Learners use flashcards to strengthen memories by repeatedly prompting themselves to review or recall mappings, one card at a time. Similarly, arcade games have a highly consistent trial-by-trial structure that repeatedly challenges players to overcome new hurdles. For example, users playing Tetris must repeatedly place blocks without overflowing the screen, and those playing Pac-man must repeatedly find paths to consume dots without being defeated by enemies. This repetitive structure is perhaps made most salient in the popular game of Snake, which incrementally grows the tail of a snake every time the user succeeds at a task, doubling as both a reward for accomplishment and a new challenge. Such games could potentially convert the explicit memorization task implied by repeated review of flash cards into one where the internalization of mappings can be incidental to the game's goals.

That memory can be enhanced via repeated recall is a finding that has emerged through decades of memory research. Tulving's pioneering work in 1967 revealed that tests not only assess learning, but also produce learning in ways that are as effective as studying [34]. This notion of the *testing effect* subsequently sparked a burst of research surrounding the impact of *retrieval practice* on memory. Retrieval practice is the act of repeatedly attempting recall from memory in multiple trials. Karpicke and Roedinger found that retrieval practice not only benefits learning as much as non-recall studying, but also improves long-term retention more than study alternatives [17]. These findings are consistent with other studies showing that testing leads to better long-term retention than repeated study, even though studying often produces a boost shortly after learning [32][38]. Bjork's work [5] further indicated that techniques which make initial learning slower or more effortful often enhance long-term retention. In the case of retrieval practice, the additional effort required to recall an item, as opposed to merely reviewing the item, appears fruitful for long-term retention. Retrieval practice is posited to be powerful because it offers opportunities to strengthen memory encodings through multiple exposures to memory cues. In some cases, retrieval practice has demonstrated an advantage even over more complex active learning strategies, such as elaborative studying with concept mapping [16].

2.0.3 Speech-enabled Games for Learning

In practice, the incorporation of retrieval practice into existing arcade games may be challenging because it requires adding additional components to an already fast-paced and potentially mentally demanding game. It is unclear, for instance, whether it is possible to make room for learning if a game is already optimized for speed, challenge, and flow. Moreover, since players are typically pre-occupied by rapid keyboard interactions, the addition of more manual interactions may be infeasible.

The hands-free nature of speech interaction may offer advantages in the adaptation of games for learning. Because speech is a typically unused input channel during traditional arcade gameplay, speech interaction could enable users to keep up with the original game speed more so than text input. More generally, the speed of voice

interaction also enables users to potentially work or play as fast as they speak instead of as fast as they type or move the mouse. Interface designers have turned to spoken language input and output as a way of alleviating manual manipulations in certain conditions, such as for users suffering from motor disabilities [27] or for those who must concurrently operate vehicles [15].

In educational games, speech production offers a significant benefit for learning because it is a central component of vocabulary acquisition. Second language acquisition (SLA) research has shown that spoken output is as much a channel for acquiring vocabulary as it is the result of learning a language. Speaking out loud strengthens memory by providing learners with phonological input back to the mind, thereby strengthening word knowledge [6]. Over the last decade, automatic speech recognition (ASR) and Voice over IP (VoIP) have made it possible to develop systems for computer assisted language learning (CALL) and computer assisted vocabulary learning (CAVL). For example, the commercially available software package, Rosetta Stone, allows students to choose from a set of pictures associated with spoken descriptions that get progressively longer. Similarly, Duolingo is a free language-learning website and crowdsourced text translation platform that is supplemented by functionality for users to record speech and be scored on their pronunciation. Recent work has also produced more complex dialogue systems and frameworks to practice translating or question-answering in the second language [40][39]. Speech-augmented games have emerged in the form of turn-based games such as Rainbow Rummy [42] and Scrabble [33], or custom-made adventure style mobile games for improving literacy in developing countries [18]. A speech-enabled game for Hispanic children based on Guitar Hero [31], for example, showed promising results in the specific genre of arcade games. However, the broader question remains as to which user interface issues are most pertinent in the adaptation of such games, and how retrieval practice compares to less cognitively intensive alternatives in these arcade-style settings.

Despite the potentially large benefit that speech interaction could bring to educational games, fast-paced games offer an unusual challenge because their motivational effectiveness depends heavily on the rhythm and flow of the game, along with clear

accountability for progress [26]. The thrill of playing a fast-paced game could be seriously dampened by the frustrating effect of speech recognition errors, a reason that perhaps explains the limited adoption of speech technology in this genre. Recent work has explored the use of context to enhance speech understanding. For example, some have explored using personal data such as address book, location and time to customize the recognizer's language model in information-access systems on mobile devices [30]. Others have leveraged dialogue context, such as a hybridization of parse scores and knowledge about dialogue progress, to reduce recognition error rates in dialogue systems for second language learning [41]. However, little research has been devoted to enhancing speech recognition systems in time-sensitive settings for rapid gameplay. Fast-paced arcade style games may offer the advantage of providing even more fine-tuned contextual information, due to simpler game logic and a more granular trial-by-trial structure compared to non-arcade games and conventional search systems.

Two primary issues emerge from this body of previous work: the challenge of modifying existing arcade games for learning, and the obstacles presented by speech recognition errors amidst gameplay. The following chapters will address these concerns through the design, implementation, and evaluation of Tetrilingo, a speech-enabled game based on Tetris.

Chapter 3

System and User Interface Design

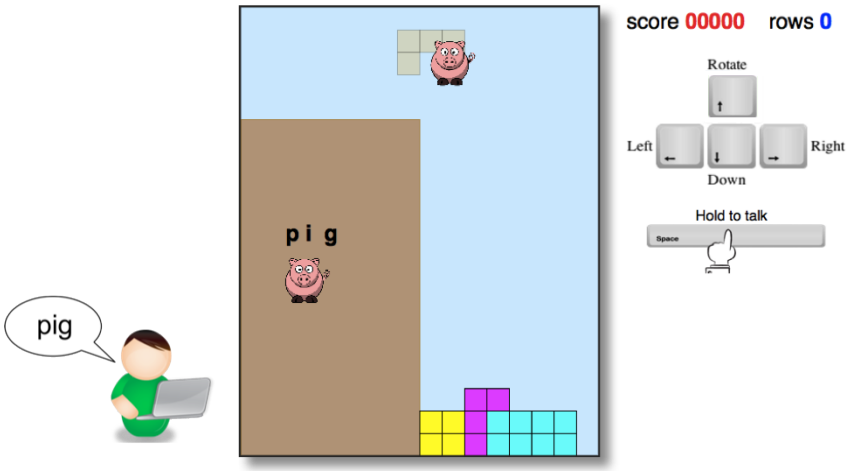


Figure 3-1: Modified Tetris game interface. Saying the correct word unlocks block rotation.

Tetrilingo, our speech-enabled Tetris game, modifies traditional Tetris rules to offer an incentive for learning any set of paired associations, such as capitals and countries or names and faces. Figure 3-1 shows our specific implementation which teaches the meanings of words by mapping words to their picture representations. Each player sees a Tetris block attached to a picture and must correctly speak the word associated with the picture before block rotation can be unlocked for the trial. We selected this particular rule modification because it most closely aligns word-learning incentives with the core means to success in the original game; traditional Tetris rewards players who can skillfully rotate and place blocks as a means to clear

rows. To give players some ability to play the game even if they do not succeed in pronouncing the word, we allow players to still move blocks left, right, and down regardless of how they perform on the learning component.

As in traditional Tetris, a block can only be maneuvered while it is still falling. Once it has dropped, the next block with a new picture immediately appears. Although our specific implementation teaches word-picture associations, the framework can in practice be applied to non-pictorial cues such as foreign-language words or definitions. Furthermore, the framework is not limited to the practice of words and their meanings. Players can use the game to learn or rehearse any set of paired associations, such as historical events and the dates on which they occurred.

We enhanced traditional Tetris with speech interaction for two reasons. First, prior research in speech-based literacy games has indicated that productive speech practice strengthens word knowledge by providing learners with phonological input back to the mind [18]. Secondly, speech is also a typically unused input channel during traditional arcade gameplay. It could therefore enable users to keep up with the pace of the original game speed more so than text input.

3.1 Interface Improvements

3.1.1 Paper Prototyping

We first created a lightweight paper prototype to test the feasibility of a speech-enabled Tetris game. Our goal was to gather initial feedback from users before investing time into software development. Figure 3-2 shows an image of the prototype, which was constructed using colored paper cut-outs as Tetris blocks and printed images of animals as word prompts. Because block rotation is a key component of the original Tetris game, we presented users with a computer keyboard for block manipulation rather than paper alternatives, so that this interaction could be as natural as possible (Figure 3-3). A graduate researcher simulated block animations by manually moving the block in response to the user pressing arrow keys on the keyboard.

Figure 3-2: Paper prototype

Two graduate students were recruited to interact with this tabletop prototype in a quiet setting. From our observation, it was not immediately clear to users that speaking the correct word would unlock block rotation for the entire trial. For example, one user spoke a word multiple times in a row, assuming that each correct utterance would enable one single rotation. Users also indicated that the picture prompt was positioned too far away from the block.

Addressing these concerns, we modified the design to provide more clarity with respect to block rotation. In the new design, the picture prompt initially appears on top of the block as a visual indication that the block is in locked mode. Once the user speaks the correct word, the picture disappears from the block, indicating that the block is now free to be manipulated for the rest of the trial. Removing visual occlusion of the block not only indicates a change of state, but also supports the user's shift in focus within a trial: the picture is in clear view when the user needs to see it for vocabulary recall, after which user attention naturally shifts towards rotating the block in order to succeed in the game. Hence, visual occlusion of the block is appropriately removed only when a clear view of the block's shape becomes necessary for the activity at hand.

Figure 3-3: During user testing, the user controlled blocks using a physical keyboard.

3.1.2 Digital Prototyping

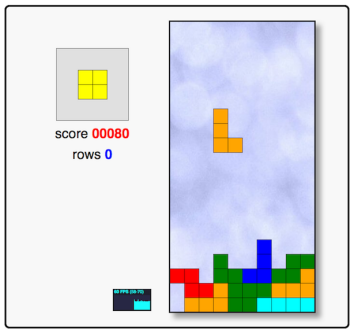


Figure 3-4: Open source Tetris implementation.

Following feedback on the paper prototype, we created a digital prototype by modifying an existing open source web implementation of Tetris (Figure 3-4)[1]. The system was iterated and refined over a period of six weeks. In total, eight MIT students play-tested successive versions of the game.

Building a functional software interface required careful consideration of the affordances for speech input in the context of a time-sensitive game. Conventional web-based speech interfaces typically require users to record speech via a two-step process of clicking a button once to record and again to stop recording (Figure 3-5). Unlike conventional search interfaces that tend to handle multi-word phrases or

[1] http://codeincomplete.com/posts/2011/10/10/javascript_tetris

sentence-long utterances, the speech-enabled Tetris game instead processes shorter, one-word utterances. Thus, clicking twice for each speech recording imposes more burden on users due to a high number of clicks within a short period of time. At the same time, the motor effort involved in holding a button while speaking is also less taxing because utterances are short. To increase the ease and efficiency of user input, we modified the recording interface to use a spring-loaded, hold-to-talk functionality. Rather than clicking twice, users instead record in one motion (press, talk, and release), making it easier to keep up with game speed.

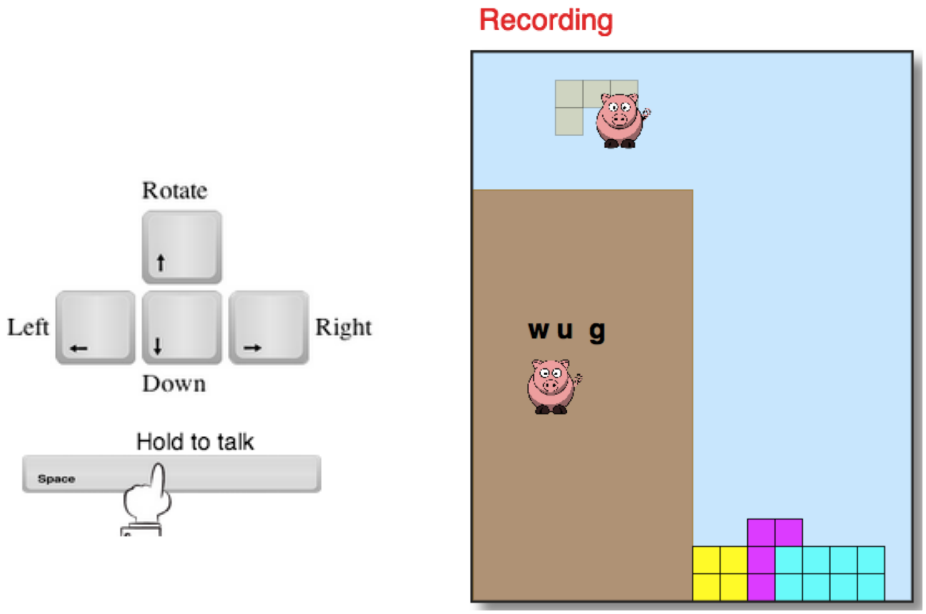
Figure 3-5: Example of a traditional speech recording interface: user presses once to record and once to stop recording. Button is highlighted red (left) to indicate that system is in record state.

Because users rarely use the computer mouse while playing Tetris, and instead anchor their hands on the keyboard to maneuver blocks, we placed the record functionality on the keyboard rather than in an on-screen button. Although we initially placed the hold-to-talk functionality on the R key (R for Record), users took some time to locate the key and also indicated that it felt somewhat unnatural. The inefficiency of locating a small key in the middle of the keyboard is consistent with Fitts' law [10], which asserts that the time required to rapidly move to a target area is a function of the size of the target and distance to the target. We thus relocated the record functionality to the spacebar, with the goal of increasing efficiency due to its larger surface area. Because the spacebar was positioned at the edge of the keyboard, users could also overshoot slightly without missing the target, thus making the effective size of the spacebar even larger.

The mappings of keys to affordances were displayed as one group (Figure 3-6a) on the right side of the screen to correspond spatially to the right-side placement of arrows on typical keyboards. Although we had intended for users to interact with the

spacebar using their left hands and arrow keys with their right hands, we found that some users used their right hands to control both of these affordances, a behavior that was inefficient for gameplay. To encourage bimanual interaction, we added an image of a left hand pushing the spacebar as a hint to users.

Furthermore, our decision to use a keyboard-based affordance for recording required us to devise an on-screen method for delivering feedback regarding the record state. We thus displayed the record state in red text directly above the game canvas (Figure 3-6b). Although highlighting the on-screen spacebar image may have been more effective, taking advantage of direct mapping, users appeared to understand the red text and behaved as expected.



(a) Tetris keys. (b) Visual indication of record state.

Figure 3-6: Modified speech recording interface.

From initial usability tests, we observed users struggling to keep up with the limited time allotted to speaking the target word in addition to block rotation. Players also instinctively started maneuvering blocks as soon as they appeared, even though this multitasking distracted them from word learning. To address these issues, we redesigned the blocks to move horizontally for a few seconds before dropping (Figure 3-7). The horizontal chute provided extra time for the learner to recall and speak words

without undue pressure to maneuver the block.

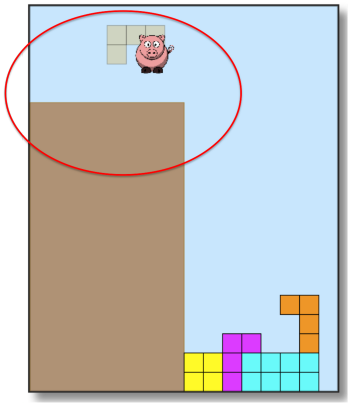


Figure 3-7: Tetris piece moves horizontally through a chute to give the learner more time.

Due to the additional time allocated to speaking, users cleared rows less frequently than in regular Tetris games. Recognizing that the pace of feedback on progress is crucial to the experience of flow [26], we shortened each row from 10 to 8 columns wide so that users could experience progress at a rate more similar to that of regular Tetris. Users indicated that easier row clearing did not negatively impact the game in light of the additional stimulus gained from word learning.

Because the Tetris blocks were the primary animation on the interface, users tended to be visually focused on the block at the time they spoke the word, often at the expense of missing feedback delivered elsewhere on the screen. We found that the text nature of feedback further decreased efficiency, as it required users to read a message before knowing whether they had succeeded, at which point the block would have already dropped further down the screen. To enhance feedback efficiency, we added multisensory feedback to be delivered at the locus of attention. In the revised design, whenever users pronounce the correct word, they not only hear a "success" game sound, but also witness the picture on the block disappear as the block transforms from translucent to opaque. As expected, we found that users no longer read the message once we implemented symbolic feedback. Users noted that they relied on the visual transformation of the blocks and game sounds for feedback on performance, rather than the "Good job! You said [target word]" text displayed

at the top of the screen.

Along the same lines, users tended to be completely focused on block rotation immediately after successful word recall, such that any further efforts to motivate learning during this time were fruitless. For example, to see whether users could learn from additional exposure to the word-picture mappings, we modified the interface to display the picture again as the block was dropping. The majority of users indicated that they did not pay attention to the picture because they were focused on rotating the block. To determine whether users might pay more attention if they were not preoccupied with block rotation, we also displayed the picture briefly once the block had finished falling at the end of each trial. Even though the picture was displayed next to the fallen block, most users did not recall seeing the picture until they were reminded in follow-up interviews. They reasoned that their anticipation of the next block and word-picture pair may have overshadowed any attempt to garner their attention at the end of the previous trial.

3.2 Final Implementation

3.2.1 Speech Recognition Architecture

To recognize speech input, we used the WAMI (Web-Accessible Multimodal Interface) toolkit [14], a client-server framework that allows the majority of the computation to be performed remotely. Figure 3-8 depicts a block diagram of the platform architecture. A large component of the underlying technology is Asynchronous Javascript and XML (AJAX), which allows the browser and Web server to communicate freely and enables development of highly interactive browser-based user interfaces. The WAMI platform provides a standard mechanism for linking the client GUI and audio input/output to the server. When a user opens the Tetris webpage, the core WAMI components first test the browser's compatibility and notify the client if the server is not compatible with the client's browser. Then, the WAMI GUI Controller and Flash component connect to the server. Once both are connected, the web server notifies

the WAMI GUI Controller, which passes the message along to the Tetris GUI so that it can initiate interaction.

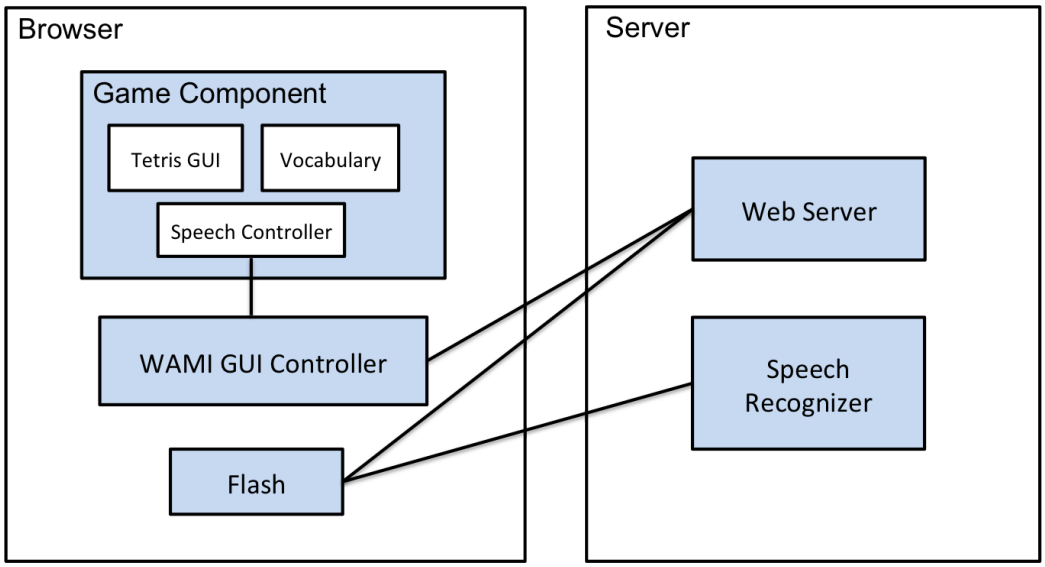


Figure 3-8: Tetrilingo system architecture.

Audio is captured at the web page through Flash and transmitted to the SUMMIT speech recognizer [12] running remotely. When the user finishes speaking, the recognizer's hypothesis is routed back to the WAMI Controller, which then informs the Game Component to formulate a response. Although WAMI's core platform can support more complex natural language processing such as dialogue management and natural language generation, we use a more lightweight version of its development model because our system only requires word-level recognition. On the client side, we provide the language model written using the JSGF (Java Speech Grammar Format) standard[2]. Figure 3-9 shows an example grammar for this lightweight WAMI interface and how it is passed to the WAMI javascript API.

3.2.2 System Architecture

The game module maintains game state through three main components, shown in Figure 3-8. 1) the Tetris GUI component controls game-specific displays and visual

[2]http://www.w3.org/TR/jsgf/

```
my_grammar =
"#JSGF V1.0;\n" +
"grammar vocab;\n" +
"public <animals> = pig | cat | mouse;";

var options = {
       grammar : my_grammar,
       guiID : 'gui-div',
       devKey : 'MY_KEY',
       onRecognition : function(result){...}
}

var wamiApp = new Wami.App(options);
```

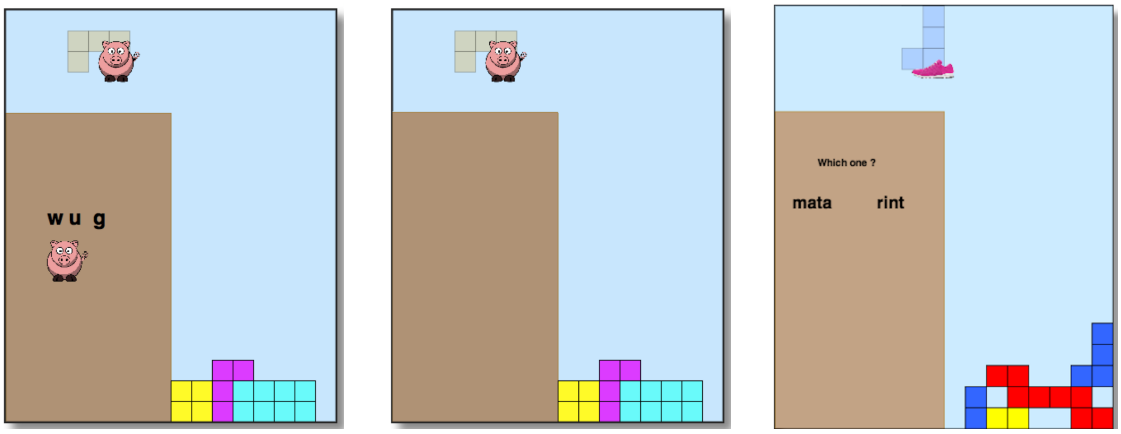
Figure 3-9: WAMI javascript API. The language model is set via the grammar option, and hypothesis results are captured in the onRecognition callback.

animations such as block movements, row completions, visual feedback on user utterances, and auditory game tones. 2) The Vocabulary component handles logic related to the ordering and frequency of word-picture pairs, and sends word-picture pairs to the Tetris GUI component to display to the user. 3) The Speech component is the interface to WAMI and speech recognition components. It initializes the language model for the recognizer and receives speech recognition hypotheses from the WAMI Controller. Lastly, a supplemental component logs all game interaction and user state for use in data analysis.

3.2.3 Game Modes

The final implementation of the game can be set in three different modes: study mode, free-recall retrieval mode, and multiple-choice retrieval mode.

In study mode, the word associated with the picture is presented each time the picture appears (Figure 3-10a). In free-recall retrieval mode, learners see the word-picture pair only the first time it appears, and in subsequent trials only see the picture displayed as a vocabulary cue (Figure 3-10b). The word is revealed inside the brown box as a hint if the learner says nothing after four seconds, or as soon as the learner records a response regardless of correctness. The purpose of the hint is to give learners



(a) The study condition presents the word along with the picture every trial.

(b) In the free-recall retrieval condition, learners must recall the associated word every trial after the first one.

(c) The multiple-choice retrieval condition displays two answer options.

Figure 3-10: Tetrilingo game modes

more support and to keep them engaged throughout the trial even if they have no recollection of the target word. This functionality was based on insight from an initial user test, in which a user expressed that not being able to recall the target word led to a helpless feeling of "having nothing to do" for the remainder of the trial.

Lastly, similar to free-recall retrieval mode, multiple choice mode displays the word associated with the picture only the first time it appears. However, in subsequent trials, learners are aided by the display of two word options to choose between (Figure 3-10c). One is the target word and the other is a distractor word randomly chosen among all other words the user has seen thus far in the game. The two words are placed side by side, and their horizontal positions are randomized so that the target word does not appear consistently in the left or right position.

Unlike free-recall mode, in which the target word is completely withheld from the user during recall, in multiple choice mode the target word is in effect visible to the user during the entire trial. Whereas free-recall mode displays the target word as a hint to the user, multiple choice mode requires a different visual hint since the word is already on-screen. We thus changed the text color of the target word to yellow as a hint (Figure 3-11), and subsequently to green once the correct word had been spoken. During local user testing, users tended to speak the target word when its

color changed to yellow, indicating that they had correctly understood the hint. If the yellow signal had been interpreted as representing the distractor, then we would have expected users to speak the distractor word when the target word turned yellow.

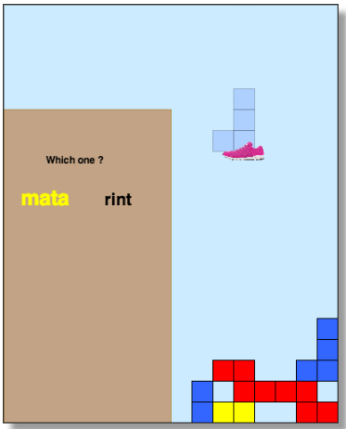


Figure 3-11: In multiple-choice retrieval mode, the target word turns yellow as a hint to the user.

3.2.4 Order of Word-Picture Presentation

During pilot user testing, we implemented a general version of the game that adapted the frequency and ordering of word-picture presentations mid-game based on user performance. The algorithm presented new word-picture pairs incrementally and at increasingly spaced intervals, an educational technique known as spaced repetition [29]. This algorithm is rooted in the notion that items which the learner finds difficult should be reviewed more frequently, and items that the learner succeeds on should be reviewed less frequently. Specifically, our implementation is modeled after the Leitner system [20], a method classically used to order the presentation of flashcards.

The Leitner method places flashcards into different bins based on how well the learner knows each flashcard (Figure 3-12). For example, items that the user has never been exposed to start in bin one. If the learner succeeds at recalling the solution for a particular flashcard, the flashcard is moved to the next bin. If the learner fails to recall the correct answer, the flashcard is placed back in the first bin. Each bin is associated with a certain frequency at which the user is required to revisit the cards

in that bin. Thus, the first bin is visited most frequently, and the last bin is visited least frequently because the learner has demonstrated competence in recalling those items.

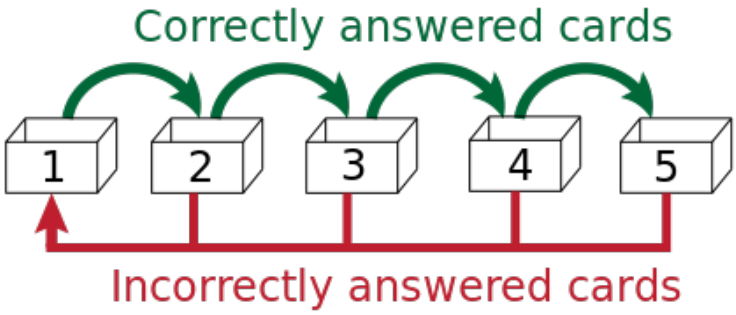


Figure 3-12: Leitner flashcard system. A flashcard advances to the next bin if the learner succeeds. Otherwise, it is sent back to the first bin. Each succeeding bin is visited less frequently than the bin before it. Image from http://en.wikipedia.org/wiki/Leitner_system

In our implementation, each word-picture pair is equivalent to one flashcard in the Leitner method. Unlike a conventional Leitner system which assumes that all flashcards are initially in bin one and thus introduced in bulk, our system requires a more gradual introduction of new word-picture pairs so as not to overwhelm the learner. We thus modified the algorithm to incrementally add new words to bin one every time it became non-full. To prevent users from learning words in a predictably ordered fashion, we also shuffled the word-picture sequences at each round so that the word-picture pairs would appear in a randomized order while still preserving their frequency requirements. Lastly, because repetitions in close proximity are less effective for retrieval practice, we added constraints so that the same word-picture pair would not be displayed twice in a row.

To determine a pace that would feel natural to users, we collected feedback from users regarding whether words were being introduced or reviewed too quickly or slowly, and adjusted the frequency level of each bin accordingly. The higher the frequency associated with each bin, the more frequently familiar word-picture pairs

would continue to be revisited, and the slower the new words would appear. We ultimately increased the frequency of each bin due to feedback from users that words were being introduced too rapidly.

With our modified, speech-enabled Tetris game system and improved user interface, we proceed to data collection and evaluation, described in the following chapters.

Chapter 4

Data Collection

To evaluate the educational effectiveness and speech recognition performance of Tetrilingo, we invited remote participants to play the fully speech-enabled Tetris game on the Amazon Mechanical Turk web service.

4.1 Amazon Mechanical Turk

Amazon's Mechanical Turk is a popular web service that pays humans to perform simple computation tasks. Workers on the system (turkers) are typically paid a few cents for Human Intelligence Tasks (HITs) that can be done within a few minutes. In the past few years, Mechanical Turk has been used by industry and academia for a variety of micro-tasks, ranging from labeling images and categorizing products to more complex tasks such as finding answers online and producing hypothetical search queries for the purpose of natural language processing. More recently, researchers have explored the idea of incentivizing turkers to play "games with a purpose" [36]. Such games typically provide casual online entertainment for two players, with the the covert side effect of performing some useful task, such as labeling an image.

We turn to Amazon Mechanical Turk as our core source of data for several reasons. First, unlike onsite user studies, deploying the system to remote users better captures the intended purpose of computer-aided learning games. Because users cannot be aided or prompted by any instructions provided by the facilitator of the study,

the user interface must itself be intuitive, easy-to-learn, and simple to interact with, just like online games in-the-wild. Second, remote data collection also allows us to observe challenges and anomalies resulting from remote speech interaction, such as degradation resulting from low quality microphones, hardware incompatibility, noisy backgrounds, as well as user distractions and multi-tasking. Lastly, because typical workers on Amazon Mechanical Turk are adults, the demographics of Mechanical Turk also match our intended target population. A major advantageous of adapting existing games for education is to lower the initial learning curve for adults by leveraging prior familiarity with game rules. Capturing data from adults also allows us to better observe the behavior of those who may not be in the daily routine of learning or memorizing vocabulary, a practice that may be more familiar to college students commonly recruited for onsite university lab studies.

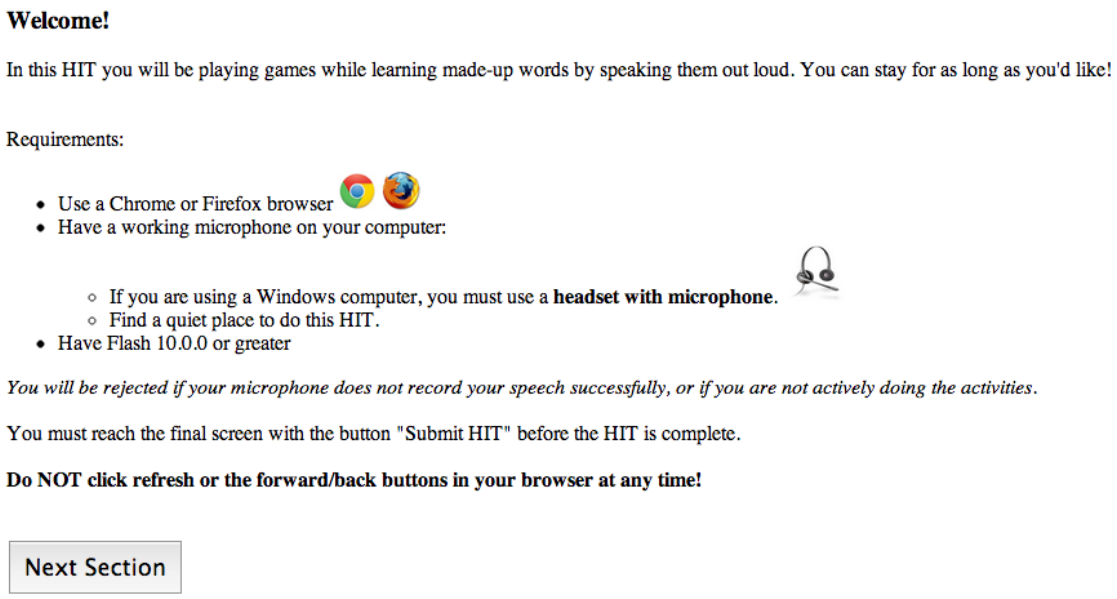
4.2 Data Collection Interface

To evaluate the feasibility of launching a speech-enabled game on Amazon Mechanical Turk, we posted a set of pilot HITs and iteratively improved the instructional interface of the HIT based on the user behavior we observed through data logs. In particular, we were interested in whether users could learn how to play Tetrilingo, whether users could successfully record speech by following our instructions, and whether the quality of the recorded speech was adequate for speech recognition.

4.2.1 Phase 1

The initial HIT interface featured a preview page with instructions (Figure 4-1) suggesting workers to use a headset microphone for higher quality speech capture. Subsequent pages showed workers how to allow Flash to access their microphones, as well as step-by-step game rules explaining how to play the modified Tetris game (Figure 4-2). Turkers were then shown the game interface and played the speech-enabled Tetris game for several minutes while learning words.

We collected data from ten turkers and observed from data logs that the vast



Welcome!

In this HIT you will be playing games while learning made-up words by speaking them out loud. You can stay for as long as you'd like!

Requirements:

- Use a Chrome or Firefox browser
- Have a working microphone on your computer:
  - If you are using a Windows computer, you must use a **headset with microphone**.
  - Find a quiet place to do this HIT.
- Have Flash 10.0.0 or greater

You will be rejected if your microphone does not record your speech successfully, or if you are not actively doing the activities.

You must reach the final screen with the button "Submit HIT" before the HIT is complete.

Do NOT click refresh or the forward/back buttons in your browser at any time!

Next Section

Figure 4-1: Initial Mechanical Turk HIT preview page.

majority (8 out of 10) almost never had a chance to rotate the Tetris blocks because their utterances were repeatedly mis-recognized as incorrect. The high frequency of mis-recognitions suggested that these users had low quality microphones and were not using headsets. We also observed that several workers had pressed the correct key to record speech but never received any feedback from the recognizer, probably because Flash was not enabled to access their microphones. We conducted further user testing in a local setting to discover why this was the case, and observed that users frequently skipped over the Flash-related step of the instructions because they interpreted the Flash box to be an instructional image rather than an interactive interface. Ironically, in our attempt to position the Flash box in a location well-aligned with the rest of the instructions, we had caused users to ignore it altogether.

4.2.2 Phase 2

In the next iteration of the task, we sought to filter out users with poor quality microphones by modifying the HIT to include a microphone test that workers were required to pass before they could proceed in the task. Figure 4-3 shows the inter-

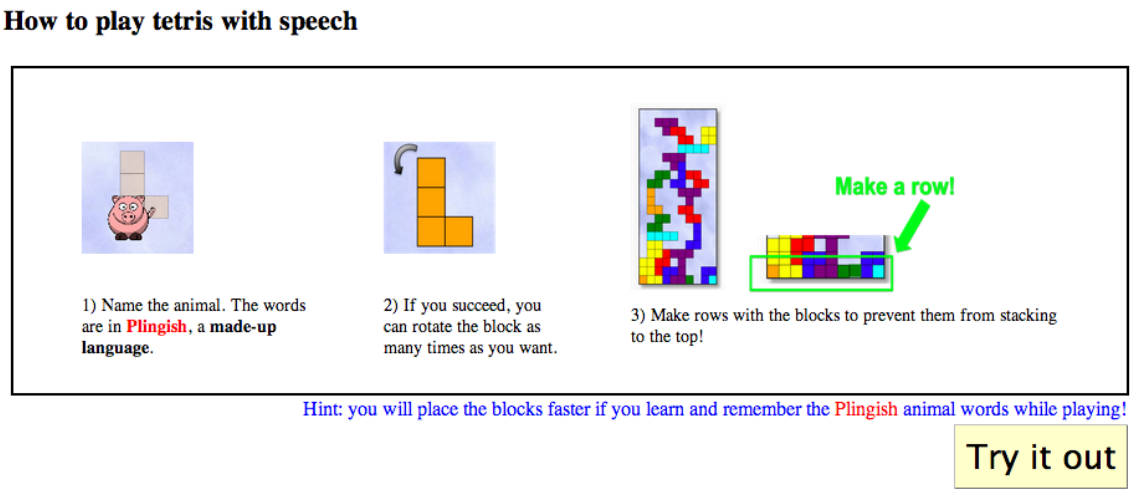


Figure 4-2: Game instructions in the HIT.

face for the microphone test. Users were required to speak the word "pig" and be correctly recognized by the speech recognizer. We purposefully seeded the recognizer with competing vocabulary acoustically similar to the word "pig" to ensure sufficient quality in the speech recordings. We also added functionality for users to click and hear their own voice recordings.

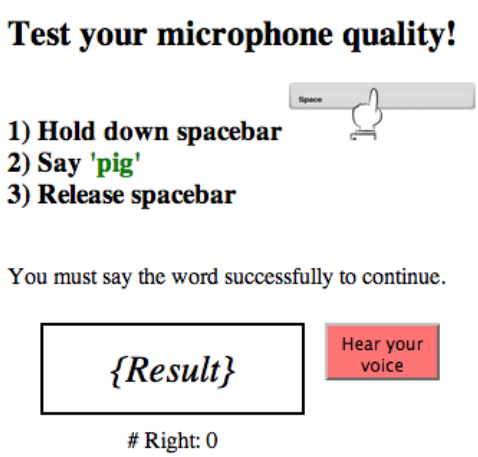


Figure 4-3: In the HIT microphone test: users must be successfully recognized speaking the word "pig" before they can continue.

To deter those with poor microphones from completing the task, the instructions also indicated that HIT submissions with poor quality speech would likely be rejected.

Addressing the Flash box issue, we also repositioned the Flash box to partially obscure non-essential objects on the page so that users would not mistake it for an image.

With this new interface, we re-launched the HIT with ten more Mechanical Turk users. Surprisingly, many users still were not progressing in the Tetris game despite the changes we had made. Although most were now successfully recording their voices (indicating that Flash had been properly enabled), audio recordings revealed a significant amount of static noise in the background, indicating that many users were still operating on poor quality microphones, despite passing the microphone test.

Upon evaluating the speech logs, we found that some users had attempted the microphone test multiple times in an effort to pass the test, and that a few were able to pass by chance after numerous tries. It occurred to us that asking workers to hear their own audio was not an effective way to filter out those with poor quality microphones. Not only are such evaluations largely subjective, but noisy speech that is poorly suited for speech recognition may still seem perfectly comprehensible to a human ear.

4.2.3 Phase 3

In the next iteration of the task, we prevented chance successes by requiring that users be correctly recognized three times in a row before they could progress. Users were allowed to re-try up to a maximum of 12 utterances and were prompted to exit the HIT once they had exceeded the maximum allowable number of tries. To help users better judge their own recording, we provided an example of a high quality utterance with no background noise and required that they compare this utterance with their own recording before continuing. Figure 4-4 illustrates the final microphone test interface.

Because users who failed the microphone test would not be paid for the HIT, we moved this portion of the instructions to the preview page of the HIT. In this way, turkers could view all information up front and could even try the microphone test before accepting the HIT. Placing the microphone test on the preview page also meant that workers would not waste time reading other instructions about the HIT

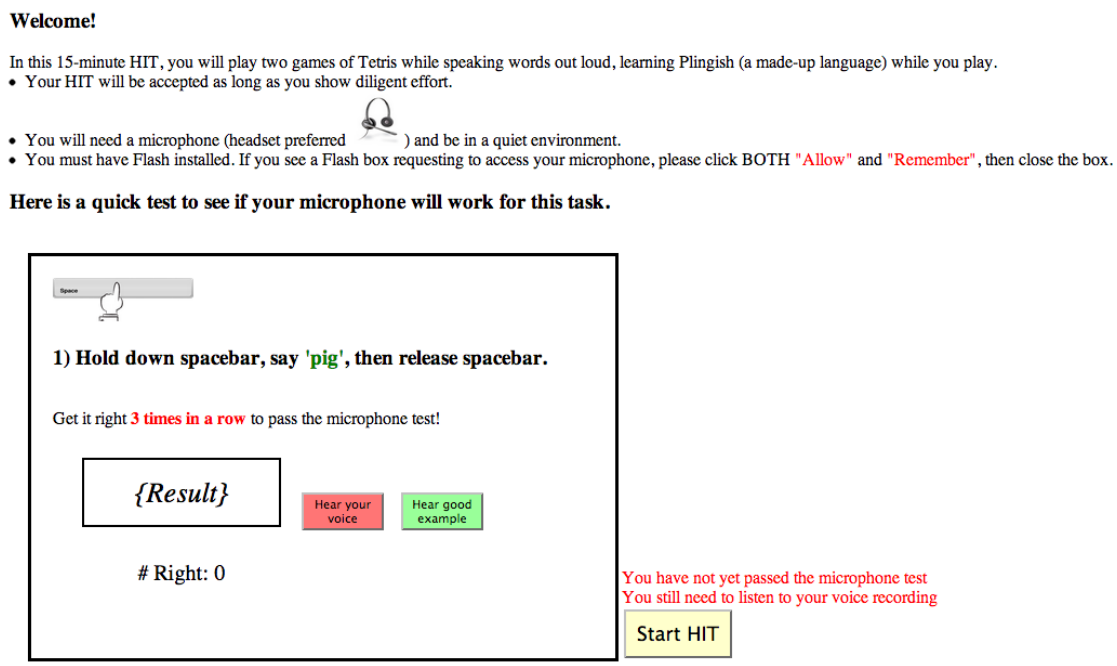


Figure 4-4: Final microphone test interface on the HIT preview page.

before discovering they were unqualified for the task.

To discover whether turkers understood game rules, we added an interactive Tetrilingo tutorial (Figure 4-5) after the microphone test and before the game. The tutorial required users to insert the block into a pre-configured slot that was oriented in such a way that only users who were able to rotate the block could succeed at the task. Because users could not rotate the block unless they had correctly spoken the target word, passing the tutorial was an indication that the user had not only maneuvered the block correctly on the game board, but also produced adequate speech recordings for recognition. To avoid presenting the user with too many instructions, the tutorial first instructed users to speak the word (via the prompt "Name the animal"), and subsequently prompted them to rotate the block once their utterance was recognized as correct. Users were allowed to re-try this activity by clicking the "Start Over" button, which reset the canvas to the original block configuration. After nine unsuccessful trials, users would be re-directed to the submission page. We paid users the full amount even if they failed the game tutorial because the microphone test and

Tetris tutorial were already worth a non-trivial amount of effort.

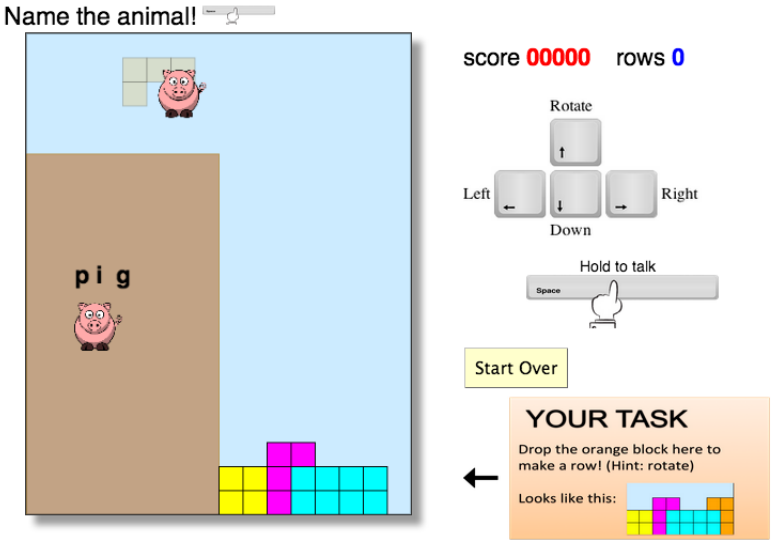


Figure 4-5: The interactive game tutorial guides users to speak the word and rotate the block into a pre-configured slot.

From the data collected, it appeared that approximately one third to one half of all turkers who attempted the microphone task succeeded. The actual percentage of turkers with quality microphones is potentially even lower, because some users seeing the microphone requirement on the preview page may have turned away immediately. In contrast, we found that only 2 users who passed the microphone test failed the Tetris tutorial, indicating that the main roadblock had been microphone quality as opposed to game understanding.

4.3 Recruiting Users for Controlled Studies on Amazon Mechanical Turk

With our revised HIT interface and Tetris tutorial, we implemented two full-scale user studies, collecting data from a total of 16 users (12 male, 4 female) between the ages of 21 and 51, with a mean age of 31.6. All participants were native English speakers located in the United States. Speech recognition performance tends to be poorer

for females, a characteristic that may have contributed to fewer females passing the microphone test and ultimately completing the study. Details on the experimental design and comparison conditions will be described in Chapter 5. In the remainder of this section, we describe our approach to recruiting and retaining users for these user studies.

Unfortunately, implementing controlled studies to measure learning on Amazon Mechanical Turk is not easy, particularly when speech recognition is involved. Because educational studies necessarily require more than a few minutes for learning to take place, and because within-subject studies require each user to experience more than one condition, a single task becomes much longer than the typical Mechanical Turk HIT, often lasting more than 10 or 15 minutes. It is thus difficult to recruit turkers, the majority of whom are more accustomed to completing a series of short micro-tasks. While a between-subject study can decrease the amount of time spent per user, between-subject studies demand a large number of users, a requirement that would be difficult to fulfill given the paucity of turkers with good quality computer microphones or headsets.

In an effort to incentivize turkers to complete our 15-minute-long studies, we paid a high price of $3.00 per HIT. Because the validity of our study depended on the integrity of turkers to be honest on learning evaluations, we further guaranteed that workers would earn full payment so long as they demonstrated adequate effort regardless of learning outcomes, and logged user interaction during gameplay to verify that this was the case. To obtain high quality workers with a greater likelihood of following instructions and producing useful speech results, we also limited the task to workers with a minimum of 98% acceptance rate on prior tasks, and required workers to be located in the United States as a way of filtering out non-native speakers who may not understand the instructions or have accents incompatible with the English recognizer.

Educational studies also typically involve a short distractor task to diminish the effects of short term memory reliance before a learning assessment, both of which further extend the time period necessary for a successful learning study on Mechan-

ical Turk. Because the Mechanical Turk platform does not explicitly support the functionality of requiring users to complete follow-up tasks, it is difficult to assess long-term retention without providing strong incentives for users to return.

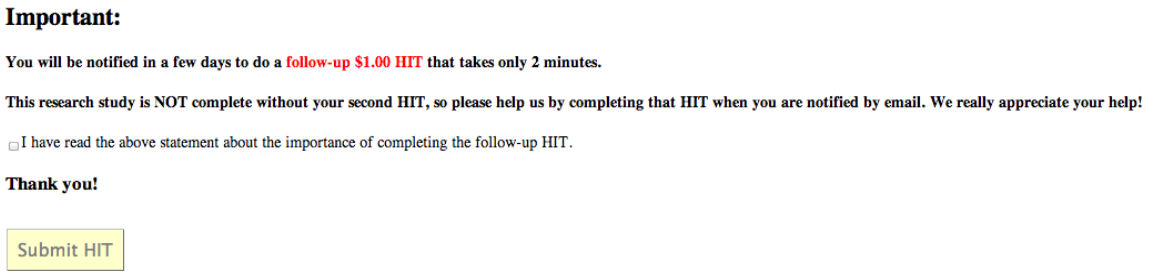


Figure 4-6: The HIT submission page asks users to confirm that they can return for a followup evaluation before the user can submit the task.

To increase the likelihood that users would return for the follow-up evaluation, we required turkers completing the initial HIT to select a checkbox stating that they would agree to complete a short, 2-minute follow-up task (Figure 4-6). Several days later, we emailed users who had completed the study, asking them to complete the 2-minute follow-up assessment. The email message underscored the importance of the follow-up task by explaining that their results would not be complete otherwise. We also set a high HIT payment of $1.00 in spite of the short task duration, as extra incentive to turkers. These strategies appeared to pay off: all but one worker who completed an initial study successfully returned for the corresponding follow-up study.

Chapter 5

Learning Assessments

In order to evaluate learning gains without giving any user an advantage due to prior knowledge, the games in the following two studies taught artificial vocabulary rather than existing words in the English language. These novel words, listed in Figure 5-1, were generated using a probabilistic model on English phonemes[1]. The 28 words were mapped to pictures of familiar animals and household objects.

5.0.1 Retrieval Practice vs. Study Practice

Although retrieval practice has been studied extensively in memory and cognition communities, there is limited work exploring the dual effects of retrieval practice on both learning and entertainment within the context of a time sensitive game. We conducted a study via a 15-minute HIT on Amazon Mechanical Turk to understand this issue.

Each user was given 1-2 minutes to practice playing the modified Tetris game. Participants then played two sessions of the game, once in study mode and once in free-recall retrieval mode. In the retrieval condition, learners saw the full word-picture pair the first time it appeared (Figure 5-2a), and subsequently only saw the picture displayed (Figure 5-2b). The word was revealed if the learner had said nothing after four seconds, or as soon as the learner recorded a response regardless of correctness.

[1] http://ibbly.com/Pseudo-words.html

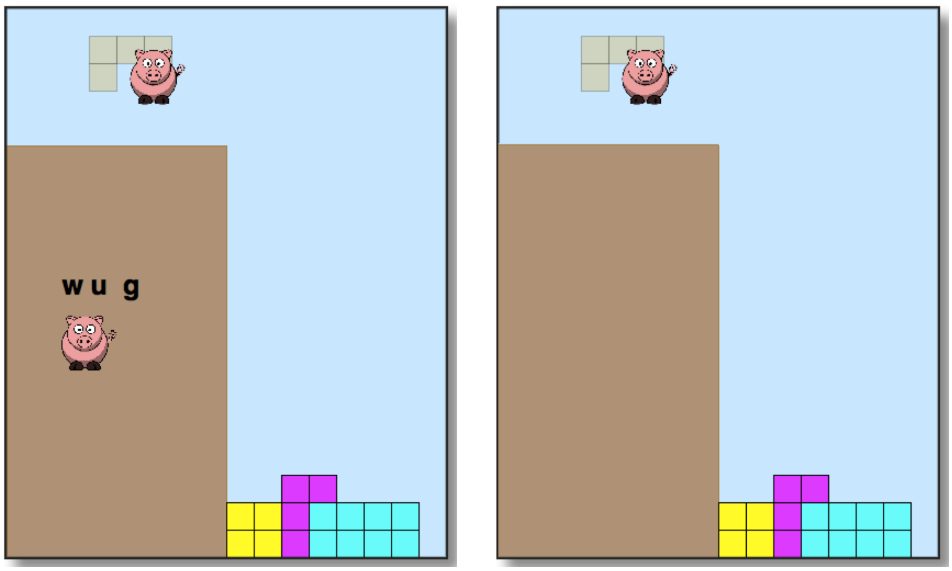
Vocabulary Words	
wug	blicket
speff	dax
pimwit	zigant
nanose	gazzer
tusket	toma
intess	fendle
priole	moffer
unty	illo
rint	del
mata	blas
pos	omma
tranco	atter
musker	corros
henne	barnel

Figure 5-1: The artificial vocabulary that users learned while playing the speech-enabled Tetris game. These words were mapped to common animals and household objects.

In the study condition, learners instead saw the word-picture pair displayed every trial (Figure 5-2a). Order of presentation and word sets were fully counterbalanced, and participants were randomly assigned to conditions.

In each condition, users learned the meaning of seven artificial words during game-play. The order of word-picture presentation was hard-coded for the purpose of controlling word exposure between the two conditions. The words were split into two groups for initial introduction and two rounds of practice, followed by two repetitions of all seven words. Thus, within a condition, each word-picture pair appeared five times, once for introduction and four times for rehearsal, totaling 35 trials for the seven words per game.

After each condition, participants completed a 45-second distractor task consisting of simple arithmetic questions, followed by a two-part quiz. The quiz consisted of a production component, in which participants saw each picture and filled in the associated word (Figure 5-3a), and a multiple-choice component (Figure 5-3b), in which participants selected the answer from the word list given a displayed picture. On both evaluations, pictures were displayed one at a time rather than simultaneously



(a) First appearance of the word-picture pair.

(b) In the retrieval condition, learners must recall the associated word every subsequent time.

Figure 5-2: Study vs. Free-recall Retrieval Conditions

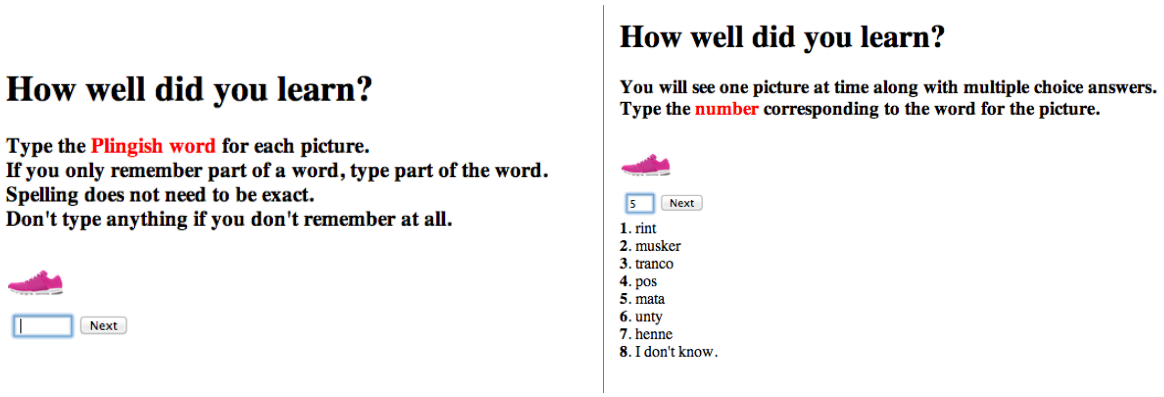
so as to prevent users from simply matching pictures to targets using process of elimination. After the quiz, users completed a Likert scale survey with questions on demographic information, level of enjoyment, self-assessed amount of learning, speech recognition performance, and prior experience playing Tetris and other video games. For the purpose of assessing long-term retention, a delayed post-test was administered to the Amazon Mechanical Turk workers between 3 and 5 days after gameplay.

Informed by previous work on retrieval practice, we have the following hypotheses for this study:

H1: Retrieval practice will be advantageous for long-term retention, an advantage that will be more salient in the production evaluation due to practice in active recall.

H2: Those who struggle with word learning will find the study condition more enjoyable, due to the high cognitive burden of learning while playing a game.

Of the sixteen users who participated in this user study, three participants were removed from the analysis. One participant indicated in the post-study questionnaire that he had stopped midway through the study and started over, so could not be included due to unequal exposure to the vocabulary words in the two conditions.



(a) Example item on the production quiz.

(b) Example item on the multiple choice quiz.

Figure 5-3: Two types of learning evaluations administered 45 seconds after the initial study, as well as 3 to 5 days after gameplay.

Another completed the initial task but did not complete the follow-up study, and the third experienced technical difficulties in one condition but not the other. Thus, we evaluate results on the remaining thirteen users.

On both evaluations, learners received one point for each word correctly produced. On the production evaluation, this included words that were misspelled but acoustically correct. We gave zero points to incorrect answers, blank answers on the production quiz, and multiple-choice selections of the answer choice "I don't know." A summary of results is illustrated in Figure 5-4.

On the multiple choice evaluation 45 seconds after playing Tetris, users on average retained 5.69 out of 7 words in the study condition and 5.62 on the retrieval condition. After 3-5 days, users still retained 4.62 words in the study condition and 5.08 words on the retrieval condition; the relatively steady performance over time on the retrieval condition is particularly impressive. Performance was lower on the production quiz: learners on average scored 4.15 out of 7 in the study condition and a slightly higher 4.85 in the retrieval condition, and this performance dropped further in the follow-up evaluation (2.38 out of 7 in the study condition and 1.23 in the retrieval condition). Such outcomes are reasonable given the extra challenge of recalling a word entirely from scratch on the production evaluation, compared to simply choosing among alternatives on the multiple choice quiz. During local user testing, for example, users taking the production quiz sometimes appeared to have the word on the "tip of the

tongue," as if they strongly recalled learning the word but could not produce it on their own. Many expressed a moment of realization upon seeing the word as an option on the multiple choice evaluation.

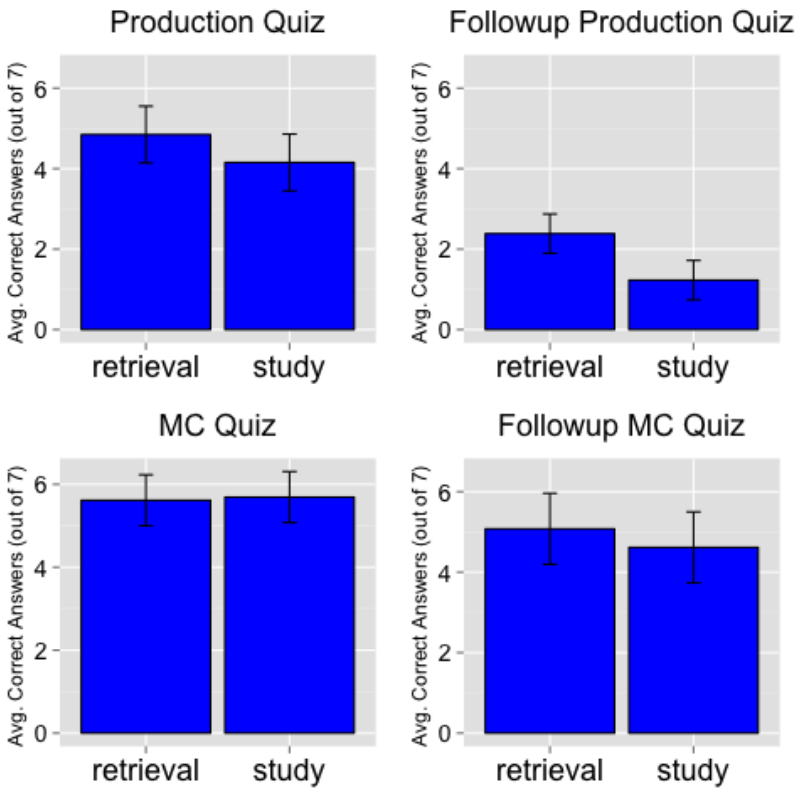


Figure 5-4: Quiz scores with 95% confidence interval, immediately after playing Tetris (two graphs on left) and 3-5 days after exposure (two graphs on right).

Despite poorer performance on the production evaluation, multiple choice results suggest that with even as little as 15 minutes of arcade-style game play, the average learner can recall 9 or 10 of the 14 total word-picture associations as many as 3-5 days after initial exposure, if provided some options to choose among. We get this approximation by combining the number of words recalled from both study and retrieval sessions. This gain is particularly promising given that, unlike many other studies on educational games, the words are introduced entirely during the game rather than in a pre-game training period.

We ran a 2x2x2 (practice type x presentation order x word set) repeated measures ANOVA on the dependent variable quiz score. Although no significant differences

were found in multiple choice evaluations, the retrieval condition demonstrated a significant advantage over the study condition ($F(1,9)=19.47$, $p=0.002$) in the follow-up production quiz taken 3 to 5 days after initial gameplay. Thus, there is critical evidence supporting a significant long-term production benefit of retrieval practice over study practice, even when the user is actively producing speech utterances in both conditions. It is possible that the benefits of free-recall retrieval practice are better captured in the production quiz, which forces users to recall the word entirely from memory, similar to the retrieval session. The multiple choice quiz, which helps support memory retrieval with a display of options, may not have achieved statistical significance as a result of being less challenging. As such, it may not significantly disambiguate learning gains between the two conditions.

Interestingly, the advantage of retrieval practice manifests more strongly in the setting where players experience the retrieval condition after the study condition, suggesting that acclimation to the game may be a prerequisite for benefitting from retrieval practice. Consistent with this, production quiz results revealed a minor interaction effect of practice type and order on quiz score ($F(1,9) = 4.71$, $p = 0.058$) 45 seconds after gameplay. Despite the practice session that users received before starting the study, it appears that more time is needed for users to gain familiarity, especially given the challenging nature of the retrieval condition. Moreover, presenting the study session before the retrieval session introduces users to the game in a more incremental fashion, which could feel more natural to new users. As one participant noted in the post-study questionnaire: "I had more fun on the second session because I had gotten the hang of it."

However, the question remains as to whether the potent benefits of retrieval practice are worth the potential risk of reduced enjoyment, particularly for learners who struggle with word learning. In the post-study questionnaire, we measured enjoyment using two 7-point Likert scale items, one for each condition. Users responded to the item "Overall, the [first/second] tetris session was interesting and fun to do," with 1 meaning "strongly disagree" and 7 meaning "strongly agree." Because users experienced the two conditions in different orders, we mapped answers to the appropriate

condition after the fact. Users on average rated their enjoyment of the study session higher (mean=6.31) than that of the retrieval session (mean=5.85).

To discover whether user enjoyment differed between high and low performers, we divided participants in half, separating them into two groups based on quiz scores. High performers rated their enjoyment equally between the two conditions, with only one person preferring the study condition. One high performer rated the two conditions equally in Likert scale questions, but indicated in a comment that "the [retrieval] session was a little more fun because it was more challenging. During the [retrieval] session, the animal names didn't immediately pop up every time a new game piece appeared, so you had to try harder to remember them." In contrast, only half of the low performers rated them equally. The other half preferred the study condition, noting that it was less stressful as it gave them more time and opportunity to learn the associations. Two low performers enjoyed the study condition more even though they admitted that the retrieval condition helped them learn better. For example, one said that "there was less 'stress' by trying to remember the names on demand," yet also commented that the retrieval session "required more learning to play as quickly as I was able in the first, so it seemed to be more of an incentive to memorize the names." However, other low performers attributed their enjoyment of the study condition directly to greater learning gains: "The [study] session was much more fun to do as I was told what the word was for each piece, and I could spend more time learning the words."

Interestingly, we observed minor differences between perceived and actual learning gains even though participants completed self assessments after they had taken all evaluation quizzes. Although actual performance on evaluations was not revealed to users, learners were implicitly made aware of words they could not recall during the evaluations, and thus had a channel for gauging their own learning. We measured self-perceived learning gains via the survey question "How many words (out of 7) do you think you learned well, in the [first/second] tetris session?" Oddly, three low-performers expressed having learned more words in the study condition, even though they had in fact performed better on words from the retrieval section, on either the

production or multiple choice quiz. The asymmetry suggests that, for some users, an overly challenging experience may be disproportionately perceived to be disadvantageous for learning, even when inconsistent with reality. An irrational bias could perhaps be explained by the strong role of enjoyment on one's intrinsic motivation to learn; an unpleasant experience could convince a learner that the activity was not worth the effort, even if it had in fact been beneficial. Such behavior is consistent with the theory of cognitive dissonance in psychology [9], which states that human beings have a motivational drive to reduce dissonance between conflicting beliefs or emotions in order to create a consistent belief system. Applied to this situation, several low performers may have ignored or downweighed benefits of the retrieval condition in order to remain consistent with their belief that unpleasant experiences are ineffective for learning. This negative effect may have been augmented by the fact that users learned artificial words in the study rather than words of a real language. If learners are more intrinsically motivated to learn real words, then it is possible that they would have been less sensitive to or more tolerant of the retrieval condition. However, this question remains inconclusive since we used only artificial words in our study.

In contrast, it was never the case that a user who reported learning more words on the retrieval condition performed better on the study condition. When asked to compare and contrast how well the two sessions helped them learn words, many high performers directly pinpointed retrieval practice as the beneficial factor: "The first was easier to learn because it did not list the animal names on the side...so you had to remember quicker and I got better at it faster." Another participant also connected retrieval practice to game motivations: "Since the words didn't immediately pop up every time, you had to remember the word to be able to move your game piece." Even those who did not recognize the benefits of retrieval practice accepted that they performed better on the retrieval condition. For example, one person wrote that "I was having a harder time during the [retrieval] session but I seemed to retain those words better. I don't know why." It is possible that those who are more prepared to reap the benefits of retrieval practice were also more able to accurately discern

performance gains, especially after a corroborating experience on the production quiz.

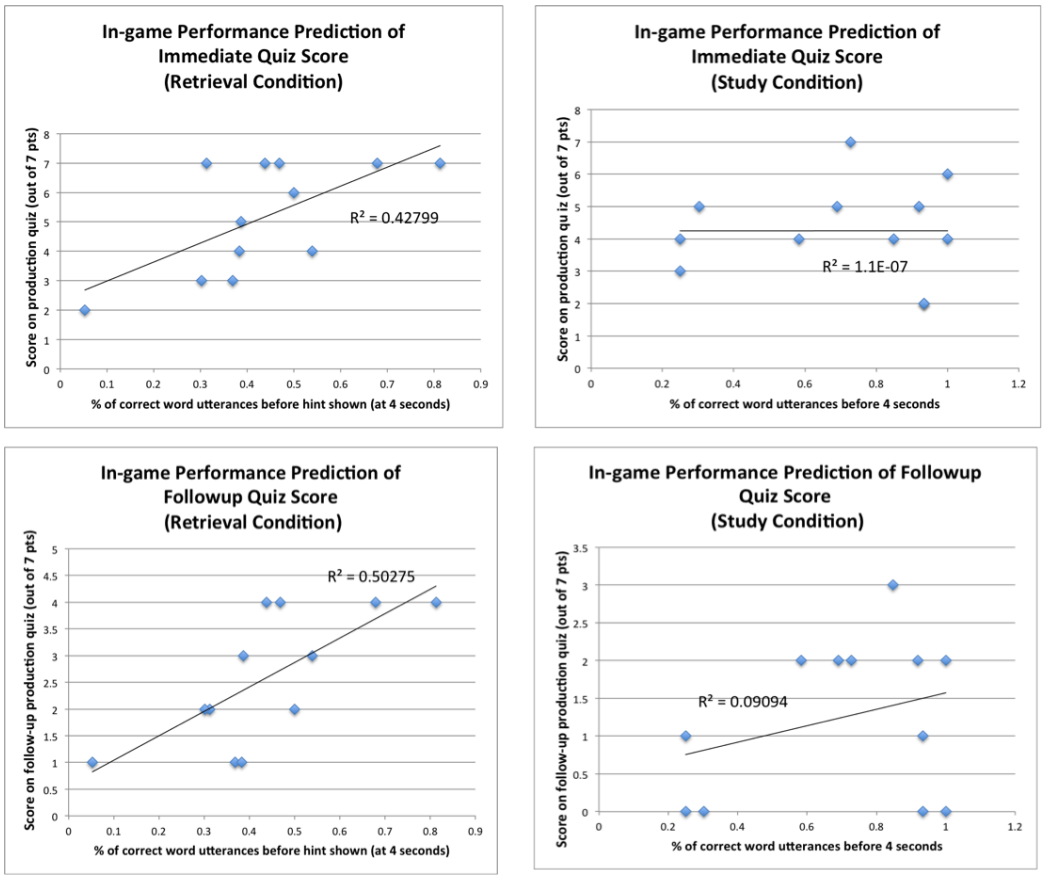


Figure 5-5: Comparing in-game performance to learning outcomes. Performance during free-recall retrieval game sessions (left) exhibit mild correlation with actual performance, while performance during study practice (right) showed no significant correlation.

Finally, we assessed the extent to which learning outcomes correlate with in-game performance. We measured each user's in-game performance by calculating the percentage of all utterances produced by the user that were correctly mapped to the picture prompt. In the retrieval condition, because users may have spoken the correct word only due to the hint provided at the four second mark of each trial, we considered only utterances produced before four seconds as being eligible in either condition. Similarly, we considered only utterances produced before four seconds in the study condition. User utterances were approximated by speech recognition hypotheses delivered mid-game rather than post-hoc ground-truth labels because we were unable to recover the appropriate timed information corresponding to individual

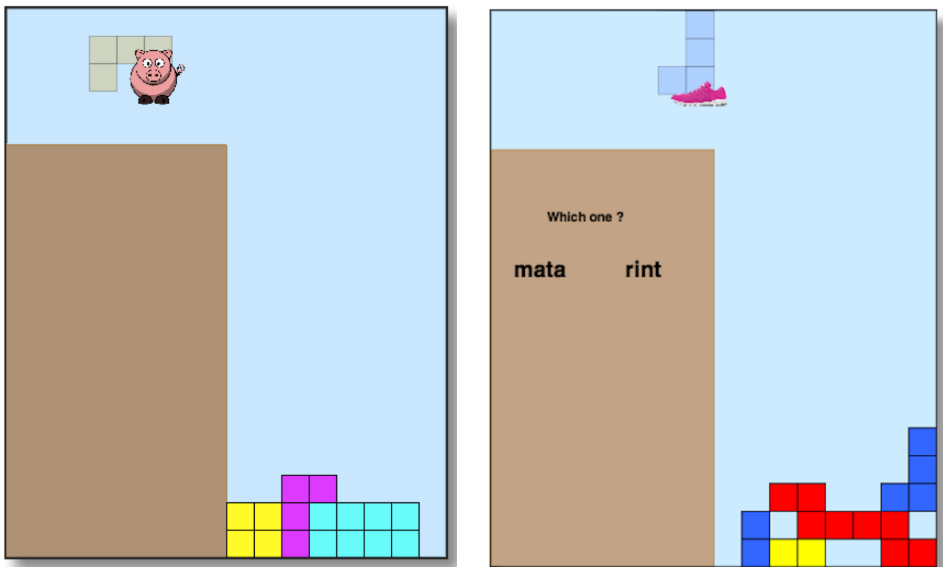
utterances after the study. Figure 5-5 shows that, while in-game performance is not correlated to learning outcomes in the study condition, in-game performance during retrieval practice is strongly predictive of both the immediate production quiz (R^2 = 0.428) and the follow-up production quiz (R^2 = 0.503). In both conditions, the multiple choice quiz exhibits no correlation with in-game performance, perhaps due to its lower sensitivity as a measurement tool. Our results may be complicated by the fact that data based on speech recognition results tend to be noisy. Nevertheless, the suggested correlation between in-game performance and learning in the retrieval condition makes retrieval practice a potentially powerful means for on-the-fly assessment of user progress.

Overall, our findings from this user study show that retrieval practice may be advantageous for both memory retention and in-game assessment of learning, but that these advantages may come at the cost of decreased user engagement, particularly when the act of retrieval imposes a cognitive burden on slower learners. Because the experience of flow hinges more on perceived than on actual skills and challenges [26], an in-game educational feature that gives players more opportunities to succeed mid-game may sustain engagement for a longer time despite slower learning.

5.0.2 Free-recall vs. Multiple Choice Retrieval Practice

To explore the effects of balancing between learning and engagement, we evaluate a different version of retrieval practice that prompts multiple choice selection rather than free recall of the target word. Similar to the free-recall retrieval mode (Figure 5-6a) described in the previous study, multiple choice retrieval mode (Figure 5-6b) displays the word associated with the picture the first time it appears. However, in subsequent trials, learners are aided by the display of two word options to choose between. One is the target word and the other is a distractor word randomly chosen among all the other words the user has seen thus far in the session. The two words are placed side by side, and their horizontal positions are randomized each trial.

We use multiple choice mode to evaluate how a less cognitively straining version of retrieval practice compares to study practice and free-recall retrieval practice, both in



(a) In the free-recall retrieval condition, the learner must recall the word associated with the picture.

(b) In the multiple-choice retrieval condition, the learner is shown two options (target and distractor) to choose between when prompted to recall the word associated with the picture.

Figure 5-6: Free-recall Retrieval vs. Multiple-choice Retrieval conditions

terms of learning and user enjoyment. To address this question, we conducted another within-subjects study on Amazon Mechanical Turk, using 14 artificial words mapped to common household items. Due to fatigue effects associated with a within-subjects study, as well as inherent difficulty in recruiting Mechanical Turk participants willing to complete long HITs, we limited our within-subjects study to only two of the three conditions. In this study, we compare free-recall retrieval practice to multiple choice retrieval practice. To evaluate user enjoyment across all three conditions, the users we recruited were limited to only those who had completed the previous study, and an additional question was added to the questionnaire asking users to rank and compare all three conditions. We hypothesize the following:

H1: Free-recall retrieval practice will still be advantageous on the production evaluation.

H2: Users will find the multiple choice condition more enjoyable than both the study and free-recall retrieval conditions, perhaps due to greater perceived learning gains than study practice and lower cognitive load than free-recall.

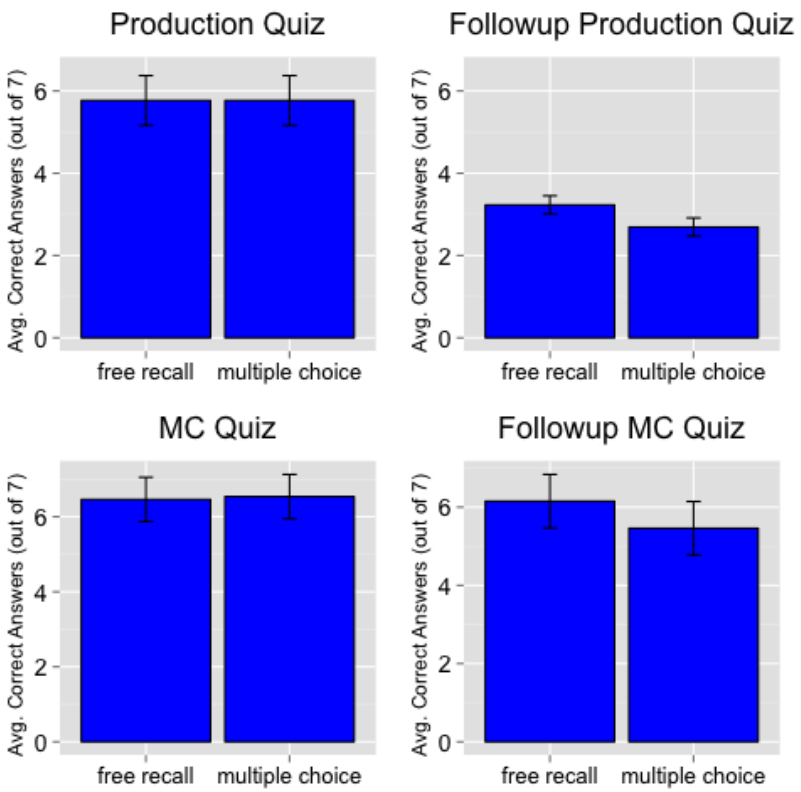


Figure 5-7: Quiz scores with 95% confidence interval, immediately after playing Tetris (two graphs on left) and 3-5 days after exposure (two graphs on right.

A summary of results is illustrated in Figure 5-7. Similar to the first user study, learners generally performed better on the multiple choice evaluation compared to the production evaluation. Overall performance exceeded that of the previous study, with an impressive average of 12.1 out of 14 words recalled on the multiple choice follow-up quiz when we combine scores from the study and retrieval conditions.

We again ran a 2x2x2 (practice type x presentation order x word set) repeated measures ANOVA on the dependent variable quiz score. In the follow-up production quiz, the free-recall retrieval condition demonstrated a significant advantage over multiple choice retrieval ($F(1,9)=13.57$, $p=0.005$), confirming our first hypothesis. Despite no significant differences in any multiple choice evaluations, the free-recall condition exhibited a higher average performance (6.15) than the multiple-choice condition (5.46) on the follow-up multiple choice evaluation. If similarity to quiz format were the main determining factor of performance, we would expect the multiple choice retrieval con-

dition to perform better on multiple choice evaluations. Our findings to the contrary lend some support to the notion that free-recall retrieval practice may offer advantages above and beyond mere similarity to the production quiz. The non-trivial advantage supports prior research on the benefits of the retrieval effect on memory retention.

In the production quiz offered 45 seconds after gameplay, we observe no difference between free-recall retrieval and multiple choice retrieval (mean=5.78 for both conditions). In contrast, for the same evaluation in the previous study, free-recall retrieval exhibited a higher quiz average compared to study practice, and a minor interaction effect of condition and order also favored the free-recall retrieval condition. It is possible that this minor advantage disappeared in the current study because users had become sufficiently familiar with the game, eliminating any ordering effects. Alternatively, free-recall retrieval may in fact be more advantageous over study practice than over multiple choice retrieval. If these claims are true, then the results would lend support to the notion that multiple choice retrieval could be advantageous over mere study practice. However, our results on this issue did not reach statistical significance and are thus inconclusive.

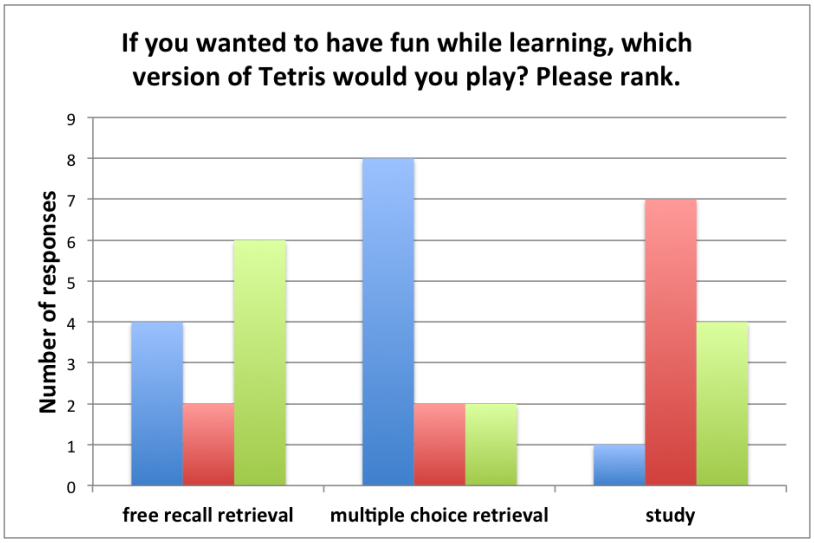


Figure 5-8: User rankings of study condition, free-recall retrieval condition, and multiple choice retrieval condition.

To investigate how overall user enjoyment on the multiple choice condition compares to its alternatives, we asked users to rank the three versions in the post-study

questionnaire item: "Overall, if you wanted to have fun while learning, which version of Tetris would you play? Please rank the three versions by typing 1, 2, or 3," where 1 means "like the most," and 3 means "like the least." In addition to defining the ranking scheme in the instructions, we also verified ranking choices against user comments to ensure that users had correctly interpreted the ranking system. Figure 5-8 displays the rankings for each condition by first, second, and third choice.

Of the thirteen total participants, eight ranked multiple choice retrieval as their first choice. By far the most common ranking pattern (made by six participants) listed multiple choice retrieval as first choice, study practice as second choice, and free-recall retrieval as third choice. Users who ranked in this fashion commented that multiple choice retrieval was a good compromise between learning and entertainment. As one user noted, "I like the middle ground of there being the word every time but you still have to make a choice." Another said that it "would help you learn better while still being fun," adding that the study condition "may be a little repetitive just showing the word every time" and the free-recall condition "would help you learn very well ... but puts on too much pressure to be as much fun." Notably, this ranking pattern also places study practice second, above free-recall retrieval. The stress imposed by an overly challenging game experience may outweigh the potential benefit of greater learning, particularly given that the study condition still produces some amount of learning.

The second most common sequence listed free-recall retrieval first, followed by multiple choice retrieval, and lastly study practice. These users appeared to enjoy the challenge of free-recall retrieval practice and preferred not to be aided by prompts. For example, one person expressed that "I like having a picture there without the word," and another explained that not having a prompt "would allow for better memorization." It therefore made sense that almost everyone who ranked the free-recall condition highest also ranked the study condition lowest. As one user put it, "[in the study condition,] my hand was held through the exercise so to speak."

Some participants also commented on their perceived pace of learning as a differentiating factor. One user commented that multiple choice helped him "learn a lot

faster because it had two words to choose from, so they were all readily in my head" while free-recall "had no helpers on the side and was harder to remember them until well into the game." Even though both versions were perceived to be useful learning tools, the user preferred multiple choice mode because it allowed him to make progress from the very beginning. A sense of progress early in the game may be an important factor for user engagement, supporting the performance-before-competence model of effective game design described in Chapter 1. Transitioning from study practice to retrieval practice in an incremental manner could also empower learners to make progress from the start without significantly sacrificing learning. One participant pinpointed this very notion of hybridizing different models for a more natural game experience: "I think it would be very effective to have it go from the word there every time to the word with a couple choices and then just the picture."

5.0.3 Conclusions from Learning Assessments

Overall, results from these studies confirm that free-recall retrieval practice is advantageous for learning when compared to study practice and, perhaps to a lesser extent, multiple choice retrieval practice. Multi-trial retrieval practice also offers a powerful means for assessing learning on-the-fly by essentially evaluating the learner's progress mid-game. Such in-game assessments may be useful for tailoring games to individual ability over time.

Despite the potent benefits of free-recall retrieval practice, the challenge inherent in this practice may be detrimental to user enjoyment when it is not well matched to the user's ability to learn or perceived level of learning, particularly in a game context. Our results show that users prefer conditions in which they perceive greater learning, so long as the activity is not excessively stressful. Some users prefer multiple choice retrieval despite admitting to the educational benefits of free-recall retrieval. Others maintain that easier alternatives are better for learning, even in spite of conflicting evidence.

Despite the less stressful nature of multiple choice retrieval practice, this alternative to free-recall retrieval is often unsupported in memory research that focuses

largely on learning benefits and less on learner enjoyment. Memory studies tend to use free-recall retrieval as the model for retrieval practice when comparing against alternative learning strategies, even though variants such as multiple choice retrieval may still yield learning advantages over study practice. In particular, multiple choice retrieval practice may help users gain skills in situations where recognition is more critical than production, such as reading street signs or understanding train announcements in a foreign country. Production processes such as speaking and writing would more likely be enhanced by free-recall retrieval practice.

Chapter 6

Speech Recognition Results

The challenge of augmenting arcade-style games with speech interaction is largely rooted in the frustrating effect of recognition errors, which highly compromises entertainment. Because fast-paced games tend to rely heavily on the user's experience of flow, the motivational effectiveness of such games can be seriously dampened by the frustrating effect of speech recognition errors. We thus evaluate recognizer performance on the in-game speech corpus we collected to better understand these challenges.

In the two studies described in Chapter 5, each user's utterances were logged to a database for the purpose of post-hoc speech recognition analysis. In total, we collected 2584 utterances from the 16 participants. Data for two sessions were not evaluated due to technical difficulties expressed in the user comments in a follow-up questionnaire. However, unlike our evaluation of learning gains, data was included for speech analysis regardless of whether that user successfully returned for follow-up tasks. We thus perform evaluation on a total of 2351 utterances.

The recognizer's performance depends critically on its letter-to-sound (L2S) model [3] used to generate lexical pronunciations for each out-of-vocabulary word. In our user studies, we used an artificial vocabulary not only to prevent any user from having an unfair advantage, but also to better model real-life scenarios in which a learner may wish to customize the game with words that are missing from the recognizer's existing vocabulary. This is a common occurrence particularly because items to be learned are

often academic terminology or proper nouns, such as chemical structures or famous historical figures. To evaluate the robustness of our L2S model, we utilized different pronunciation models ranging from one to twenty-best pronunciation hypotheses per word. The lexicon for speech recognition used an English letter-to-sound model.

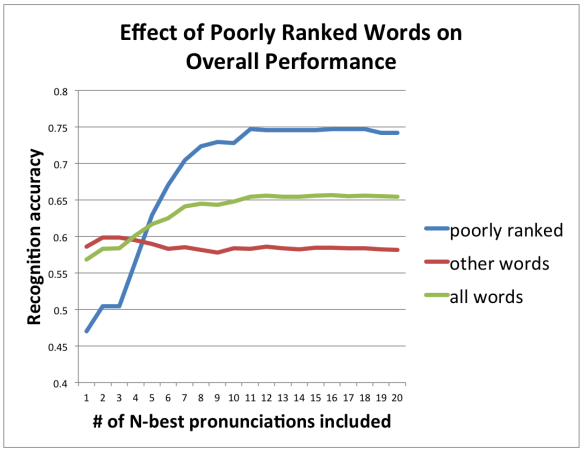


Figure 6-1: Poorly ranked words (9 of 28) account for increased recognition accuracy when more pronunciations are included in the L2S model.

The N-best pronunciations were produced from the SUMMIT L2S model applied to the 28 artificial words. We configured a static recognizer with the full 28-word vocabulary and evaluated it on all utterances in which the speaker had produced any one of the 28 vocabulary words. When only one pronunciation per word was included in the L2S, recognizer performance was surprisingly low at 55%, but accuracy increased to 63% when 20 pronunciations were included per word. Although performance for the majority of words peaked at a small number of included pronunciations, for 9 of the 28 words the most common pronunciation was ranked very low, causing overall performance on the 28 words to suffer in lexicons using only a limited number of L2S pronunciations (Figure 6-1). Hence, the total corpus benefited from an expansion of the lexicon to include more N-best pronunciations. The high risk of missing a key pronunciation commonly produced by users thus appears to outweigh the diluting effect of including greater pronunciation variety.

We also examined the extent to which performance could be enhanced by including L2S confidence scores for each pronunciation (Figure 6-2). Confidence scores [7]

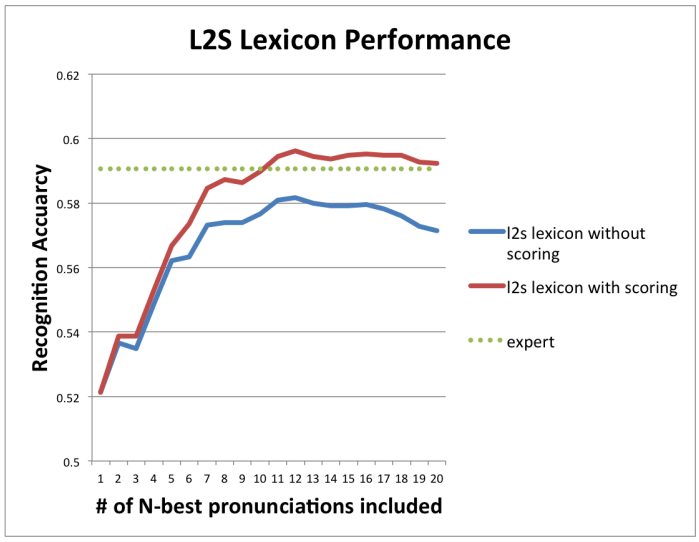


Figure 6-2: Comparison of L2S performance with and without confidence scoring to an expert L2S.

are used to weigh pronunciations based on their likelihood of being correct. For a benchmark comparison, we also evaluated the same corpus on a lexicon built using 1-best pronunciations manually created by an expert. Regardless of the number of pronunciations included, the expert lexicon performed better than an L2S lexicon with no confidence scoring, illustrating the disadvantage of poor pronunciations in the lexicon. However, the inclusion of L2S confidence scores produced a recognizer whose performance surpassed expert lexicon performance when the L2S model included at least ten-best pronunciations, illustrating some tangible benefit to including pronunciation variety on untrained words, particularly if confidence scores are available to down-weigh less likely pronunciation occurrences. In line with this notion, letter-to-sound confidence scores kept performance relatively steady even at the inclusion of a high number of potentially irrelevant pronunciations, a point at which lexicon performance without confidence scores had begun to drop.

Although average recognition performance on a static 28-word recognizer was surprisingly low, recognition accuracy for the highest performing speaker was 94%, and it was above 85% for the top four speakers (Figure 6-3). As our user study was strictly a remote task, the remarkably wide spread among different speakers is partly due to substantial differences in microphone and hardware quality on different

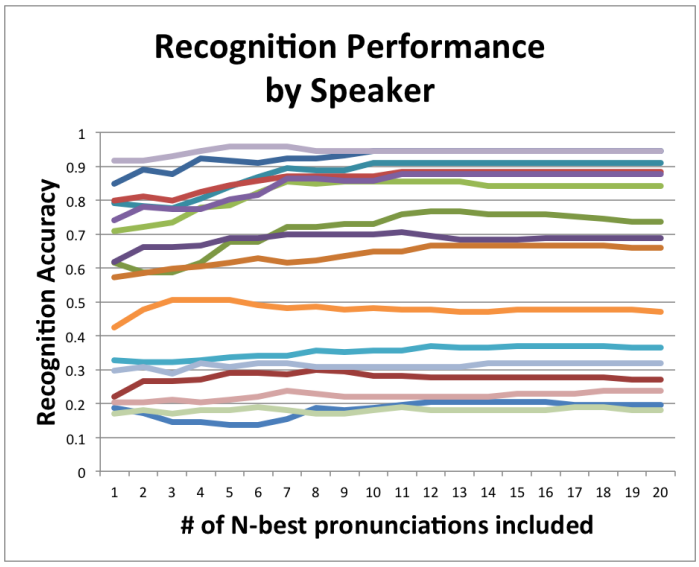


Figure 6-3: Comparing recognizer performance across the 16 different speakers.

computers. To better understand the low average performance and high variance among speakers, we further categorized misrecognitions by false negative and false positive recognition errors. We found that the vast majority of errors were due to false negatives (85%), and only a small number were false positives (2%). The remaining errors (neither false positive nor false negative) were situations in which the learner produced the wrong utterance, but the recognizer hypothesized a third word that was neither the learner's utterance nor the target word.

Interestingly, the alarmingly high false negative rate was partially a function of in-game user behavior. Many users tended to repeat the same utterance multiple times upon experiencing a false negative error, in an attempt to resolve the recognizers mistake. These repeated false negatives widened the performance gap between speakers because a single false negative error would almost always be exacerbated by an ensuing sequence of more false negative errors. This behavior may manifest particularly strongly in fast-paced game settings with short target utterances; the urgency associated with game incentives (i.e. Tetris blocks dropping) is complemented by the fact that one-word utterances are easy to repeat incessantly and thus worth the attempt. To discover the impact of repeated false negatives, we re-evaluated the corpus without false negatives that had been purely due to repetition, and found a

14% increase in overall recognition performance.

False negative speech recognition errors also appeared to have an asymmetric impact on user enjoyment. In a post-study questionnaire on Mechanical Turk, some users reported that false negative errors inhibited their enjoyment of the game. For example, one user wrote that false negatives "made me less engaged, because I felt like [the game] was counting off for something I knew." On the other hand, false positive errors seemed to have a less detrimental effect on user enjoyment. Observations from local pilot testing revealed that false positive errors were more rare because users tended to speak only when they had some confidence or inkling of the correct answer. Moreover, because the target answer was revealed whenever the user succeeded, users often appeared amused rather than misdirected by the small number of false positives that they experienced.

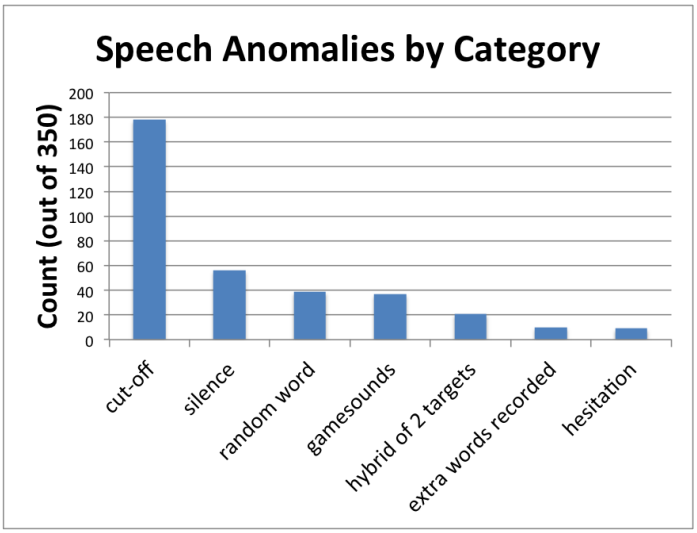


Figure 6-4: Count of anomalous utterances by category.

The combination of time pressure and playful exploration inherent in gameplay may also have contributed to more anomalous utterances, which further increased the number of recognition errors. Anomalous utterances (Figure 6-4) accounted for 15% of the speech corpus and 10% of all recognition errors. For example, because we had changed the input method to be spring-loaded to optimize efficiency, some recordings were partially cut-off due to the player releasing the record button prematurely. At other times, recordings were silent because the user hesitated to speak or accidentally

pressed the record button. On occasion, game sounds such as row-completion ringing tones were also captured in the recording, even though they were designed to not overlap temporally with recorded speech. Furthermore, some users uttered nonsense phrases or English labels for the pictures, perhaps in a playful attempt to test the recognizer or in order to trigger the display of a hint, which is designed to appear once the user has attempted any utterance in a trial. More rarely, users conflated two vocabulary words and spoke a hybrid of two words.

Overall, the most common anomalous cases were cut-off words and silent recordings (51% and 16% of anomalies, respectively). Cut-off recordings could be addressed by having the system constantly listen for speech and pad recorded utterances with extra time on both ends before sending them to the recognizer. Silent recordings could be better handled by incorporating silence into the recognizer's language model such that silence is a competing hypothesis in addition to the existing vocabulary words. In cases where the recognizer hypothesizes silence, the game interface can give feedback to the user to try again or speak louder. We leave these improvements for future work and instead focus on improving overall performance regardless of anomalies.

Chapter 7

Improving Speech Recognition Performance

The disheartening effect of false negative recognition errors on user enjoyment suggests that relaxing the constraints of speech recognition to be more lenient could benefit engagement. The difficulties inherent in optimizing a letter-to-sound model for out-of-vocabulary words might also be alleviated by training lexicons on user-produced pronunciations mid-game that are detected to be likely correct. To this end, game-based constraints could be leveraged to provide strong contextual clues for maintaining high recognition accuracy in the face of greater leniency. To explore the viability of this approach, we identify several potential techniques for modifying the speech recognizer and re-evaluate the collected speech corpus on alternative recognizer configurations.

7.0.4 Dynamic vs. Static Vocabulary

Effective educational approaches tend to focus the learner's attention on only a few words or concepts at a time until their meanings have been internalized by the learner through repeated practice. In an intense and time-sensitive game setting, the gradual introduction of small sets of words is also critical for reducing the learner's cognitive load imposed by existing simultaneous interactions. Unlike typical speech interac-

tions in which the set of possible user utterances may be large and uncertain, speech interactions amidst a learning game have implicit constraints that can be leveraged for enhancing speech recognition. Specifically, the game environment enables us to both constrain the recognizer vocabulary size and dynamically add additional words to the vocabulary as they are introduced to the learner. Constraining the vocabulary size can hopefully decrease the likelihood of false negative errors by preventing the recognizer from hypothesizing a word that the learner is unlikely to produce.

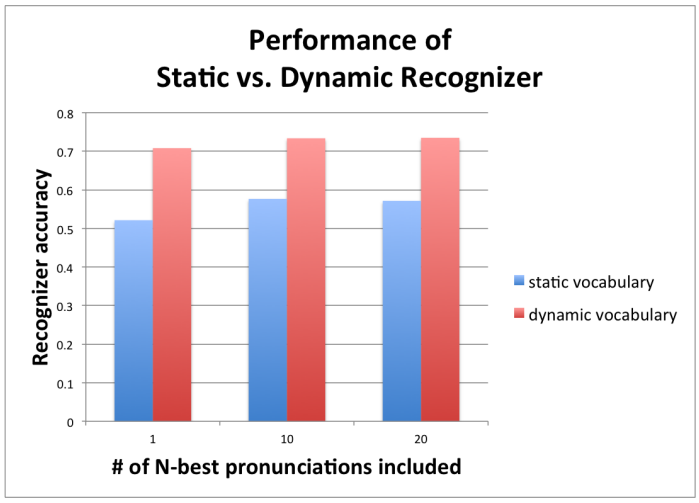


Figure 7-1: Performance of static vs. dynamic recognizer at 1, 10, and 20 included pronunciations in the L2S model. The advantage of the dynamic recognizer remains fairly consistent across different numbers of N-best pronunciations.

To determine the potential impact of this approach, we compare recognition accuracy between a static vocabulary of 28 words and a dynamic vocabulary (Figure 7-1), at varying numbers of pronunciations included in the lexicon. In the dynamic condition, we add a new word to the vocabulary only once it has appeared in the game, and constrain the maximum vocabulary size to only the words that the learner has seen within any particular game session (seven words maximum). The dynamic vocabulary demonstrated a 27% increase in accuracy over the static vocabulary when 10-best L2S pronunciations were included, and this benefit appeared fairly consistent across different numbers of N-best pronunciations included. The benefits were largely due to the substantial reduction in false negative errors, which were the source of most recognition errors.

7.0.5 Deepening N-best Hypotheses

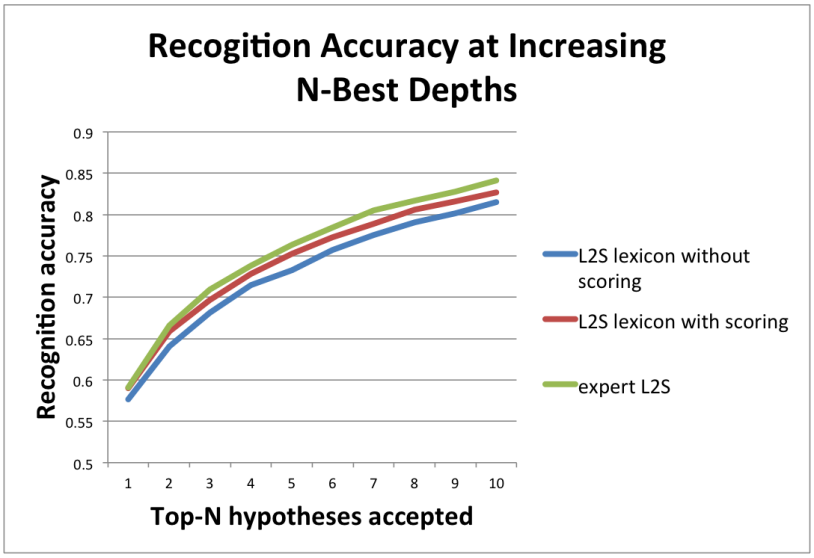


Figure 7-2: Recognition performance, varying the number of N-best hypotheses accepted. Utterance is deemed correct when any top-N hypothesis matches the target word. Uses 10-best L2S pronunciations.

Game-based settings also provide strong contextual information about the target item on a trial-by-trial basis. Because the game keeps state of which target item is being presented to the user at every turn, a more lenient system could deem the learner correct if the target word appears in any of the top-N recognition hypotheses. This approach assumes that the recognizer has some room for error and that, because the learner is likely to have spoken the target word, it is safer to check the top few hypotheses for the correct response before deeming the utterance incorrect. Figure 7-2 illustrates a substantial increase in overall word accuracy simply by expanding the N-best depth from one (59%) to four (73%), all with a static vocabulary of 28 words. In practice, even though recognition accuracy could be further boosted with more hypotheses accepted, it would be preferable to set a limit on this number so that the user does not assume that the recognizer will accept any response.

A primary concern surrounding N-best depth expansion is the increased risk of false positive recognition errors. In the case of false positive errors, learners may mistakenly believe they have correctly recalled the word for a particular picture,

with the consequence of strengthening an incorrect mapping. Hence, a trade-off may exist between decreasing frustration due to false negatives and increasing incorrectly learned mappings due to excessive leniency.

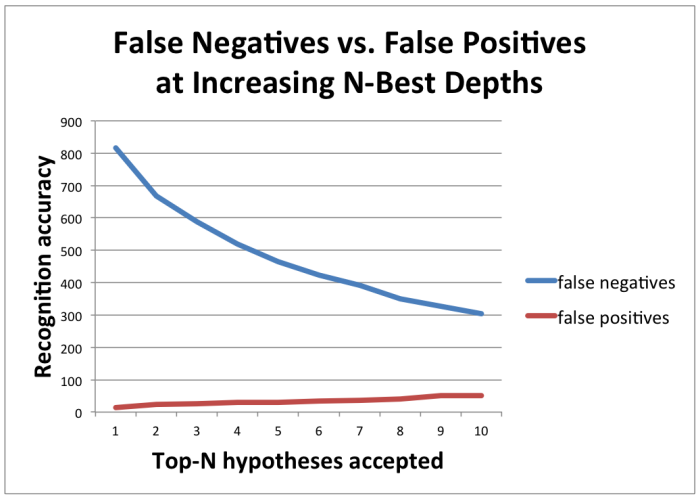


Figure 7-3: Comparing the number of false negative and false positive utterances at an increasing number of N-best hypotheses accepted.

To examine this potential trade-off, we measure the number of misrecognized utterances due to false negative and false positive errors at increasing N-best depths. Figure 7-3 shows that, as the number of accepted hypotheses increases, the number of false negative errors decreases dramatically, with only a minor increase in false positives. The significant decrease in false negatives is magnified by the elimination of repeated false negative errors due to learners re-attempting the same utterance after experiencing a false negative. Nevertheless, we find a very similar trend even after removing such repetitions from the dataset.

We further analyze false negatives and false positives among anomalous utterances, and find that anomalous recordings account for a substantial 80% of all false positive errors, compared to only 24% of all false negative errors. Because the majority of false positives are anomalies, and because a sizeable number of those are due to users producing random utterances, learners may find false positives more transparent and potentially less impenetrable than false negatives. In general, false positives are also less frustrating because they do not unfairly hinder the player's in-game progress. After a false positive, the player immediately focuses his or her attention

on block rotation rather than being forced to re-attempt the utterance, making those experiences potentially more forgettable. These patterns lend support to the notion of adapting in-game speech recognition systems to be more lenient.

7.0.6 Training on High Confidence User Utterances

Lastly, out-of-vocabulary terminology can be detrimental to recognition accuracy and game enjoyment. Unlike acoustic and language models that learn the values of their parameters from training data, word pronunciations in a recognizer's lexicon are typically specified manually, often by an expert. Hence, a user wishing to review out-of-vocabulary words might encounter frequent recognition errors due to a letter-to-sound model that has been trained using only existing lexicons.

Recent work on pronunciation mixture models (PMM) has made it possible for experts to specify a set of pronunciations, but leave the weighting of these pronunciations to the PMM using speech data collected on the fly [23]. Yet, in a game-based learning context, it is unclear how unlabeled utterances can be used for training a PMM live, due to a chicken or egg problem of learners being unreliable agents for speaking the correct target item.

Nonetheless, we make a key insight that players are typically first introduced to the word-picture pair before the word is withheld for memorization practice. Because the learner sees both the word and cue on the first trial by way of introduction, the first utterance the player produces for any word has a high likelihood of being correct. In the Tetris game we have designed, the learner also hears the word pronounced out loud when it is first introduced, making it even more likely that the learner will speak the target word correctly, particularly in the case of second language learning. On the other hand, first utterances may also be riskier for training since they could contain more anomalies such as hesitation and silence due to the user's unfamiliarity with the new item.

We thus evaluate speech recognition using pronunciations obtained by training a pronunciation mixture model solely on the user's first utterance of each word as a replacement lexicon (Figure 7-4). As a benchmark, we compare these results against

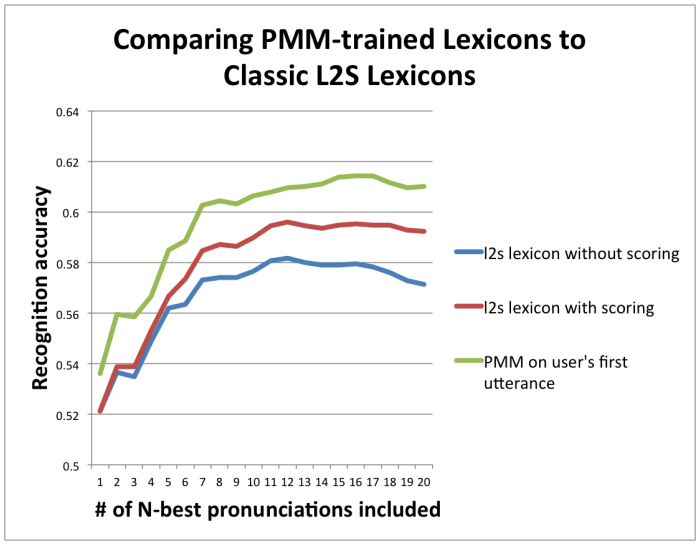


Figure 7-4: Comparing performance of a PMM trained on the first utterance of each word to that of normal L2S lexicons.

lexicons produced using the letter-to-sound model. Because the test set for the PMM condition does not include any of a user's first utterances, we similarly remove all first utterances when evaluating recognition on the normal letter-to-sound lexicons. Remarkably, the PMM trained purely on the user's first utterances demonstrated a 3% improvement over the L2S lexicon (averaged over results from one to twenty pronunciations included), despite having no ground-truth labeling of any first-trial utterances. A PMM trained on other learners' first utterances produced no significant advantage over the L2S lexicons, suggesting that speaker-dependent characteristics may be critical to effective recognition.

The promising speech recognition enhancement obtained by training only a small number of high confidence user utterances suggests further exploration of opportunities to perform user-specific PMM training using high confidence in-game scenarios. For example, starting a game in study mode before transitioning to retrieval mode could not only give the learner more time to develop familiarity with new items, but also offer an advantage for speech recognition enhancement. One could imagine collecting utterances during the study phase to produce a true mixture of multiple utterances produced by the same user for each word.

Chapter 8

Conclusion and Future Work

In this work, we have proposed several design techniques for leveraging existing arcade-style games in a learning context, such as adjusting for lost time and optimizing efficiency of input and feedback. In our controlled study, we have also demonstrated that, even when embedded in a fast-paced game, retrieval practice offers a potent production vocabulary gain over studying, but that this benefit may come at the cost of reduced engagement, particularly for slower learners.

While existing arcade-style games have experienced limited adoption in the realm of education compared to custom-made arcade games or virtual environment games, our promising results suggest that this domain deserves more attention. In particular, greater emphasis should be placed on understanding potential trade-offs between learning benefits and user enjoyment in the educational adaptation of fast-paced games. The results from this study also open a gateway for exploring other forms of retrieval practice that may be less cognitively demanding and include more instructional scaffolding, such as displaying a few word choices. Gradual adaptation of learning challenges, perhaps by offering a combination of study and retrieval trials, should also be explored in future research.

Our research has further shown that a speech recognizer designed for traditional purposes may be unnecessarily strict when placed in a fast-paced game context, particularly because false negative recognition errors are both self-perpetuating and detrimental to learner enjoyment. We have proposed several techniques for improving

performance, such as using a small and dynamic recognizer vocabulary, expanding the set of N-best accepted hypotheses, and using high confidence in-game utterances to retrain out-of-vocabulary words. Although a more lenient recognizer may run the risk of accepting learner errors, we found these occurrences to be surprisingly rare, and well worth the trade-off of decreasing the significant frustration associated with false negatives. It would be worthwhile to evaluate whether first utterances remain advantageous for PMM training in a second language learning context, despite learner inexperience in the target language.

While speech recognition has experienced limited adoption in fast-paced educational games compared to alternatives such as adventure style games, our results suggest that tailoring the recognizer to the unique needs of time-sensitive game environments could be key to increasing adoption. Future work should explore methods for handling speech anomalies specific to learning amidst rapid gameplay, such as using voice activity detection or time padding to prevent cut-off speech, and a silence model to handle accidental or hesitant recordings. Finally, automatic detection of words that are likely to be poorly ranked by the recognizers letter-to-sound model, perhaps by comparing PMM scores to default L2S rankings of out-of-vocabulary items, would be a worthwhile venture for future research.

Bibliography

[1] Clark C. Abt. *Serious games*. University Press of Amer, 1987.

[2] Alan D. Baddeley, Graham J. Hitch, et al. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[3] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008.

[4] Christian Bisson and John Luckner. Fun in learning: The pedagogical role of fun in adventure education. perspectives. *Journal of Experiential Education*, 19(2):108–12, 1996.

[5] Robert A. Bjork. Assessing our own competence: Heuristics and illusions. 1999.

[6] Kees Bot. The psycholinguistics of the output hypothesis. *Language learning*, 46(3):529–555, 1996.

[7] Jerome Bruner. Child's talk: Learning to use language. *Child Language Teaching and Therapy*, 1(1):111–114, 1985.

[8] Jerome Seymour Bruner. *On knowing: Essays for the left hand*. Harvard University Press, 1979.

[9] Leon Festinger, Henry W Riecken, and Stanley Schachter. When prophecy fails. 1956.

[10] Paul M Fitts. The information capacity of the human motor system in controlling the amplitude of movement. 1954.

[11] James Paul Gee. What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, 1(1):20–20, 2003.

[12] James R. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech & Language*, 17(2):137–152, 2003.

[13] E. Bruce Goldstein. *Cognitive psychology: Connecting mind, research, and everyday experience*. Wadsworth Publishing Company, 2008.

[14] Alexander Gruenstein, Ian McGraw, and Ibrahim Badr. The wami toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 141–148. ACM, 2008.

[15] Alexander Gruenstein, Jarrod Orszulak, Sean Liu, Shannon Roberts, Jeff Zabel, Bryan Reimer, Bruce Mehler, Stephanie Seneff, James Glass, and Joseph Coughlin. City browser: Developing a conversational automotive hmi. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4291–4296. ACM, 2009.

[16] Jeffrey D Karpicke and Janell R Blunt. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018):772–775, 2011.

[17] Jeffrey D Karpicke and Henry L Roediger. The critical importance of retrieval for learning. *science*, 319(5865):966–968, 2008.

[18] Anuj Kumar, Pooja Reddy, Anuj Tewari, Rajat Agrawal, and Matthew Kam. Improving literacy in developing countries using speech recognition-supported games on mobile devices. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1149–1158. ACM, 2012.

[19] Katsuko Kuwada. The effect of interactivity with a music video game on second language vocabulary recall. *About Language Learning & Technology*, 74, 2010.

[20] Sebastian Leitner. *So learnt man Lernen: Der Weg sum Erfolg (Learning to learn: The road to success).*

[21] Mark R Lepper, David Greene, and Richard E Nisbett. Undermining childrens intrinsic interest with extrinsic reward: A test of the. *Overjustification Hypothesis. Journal of Personality and Social Psychology*, 28:129–37, 1973.

[22] Thomas W. Malone. What makes things fun to learn? a study of intrinsically motivating computer games. *Pipeline*, 6(2):50–51, 1981.

[23] Ian McGraw, Ibrahim Badr, and J Glass. Learning lexicons from speech using a pronunciation mixture model. 2013.

[24] Ian McGraw, Brandon Yoshimoto, and Stephanie Seneff. Speech-enabled card games for incidental vocabulary acquisition in a foreign language. *Speech Communication*, 51(10):1006–1023, 2009.

[25] Curtiss Murphy. Why games work and the science of learning. In *Interservice, Interagency Training, Simulations, and Education Conference*, 2011.

[26] Jeanne Nakamura and Mihaly Csikszentmihalyi. Flow theory and research. *Oxford handbook of positive psychology*, pages 195–206, 2009.

[27] Jan Noyes and Clive Frankish. Speech recognition technology for individuals with disabilities. *Augmentative and Alternative Communication*, 8(4):297–303, 1992.

[28] Jean Piaget. *Play, dreams and imitation*, volume 24. New York: Norton, 1962.

[29] Paul Pimsleur. A memory schedule. *The Modern Language Journal*, 51(2):73–75, 1967.

[30] Stephanie Seneff, Mark Adler, James R. Glass, Brennan Sherry, Timothy J. Hazen, Chao Wang, and Tao Wu. Exploiting context information in spoken dialogue interaction with mobile devices. In *Proceedings of the International*

Workshop on Improved Mobile User Experience (IMux). Helsinki Institute for Information Technology. Retrieved November, volume 16, page 2007, 2007.

[31] Anuj Tewari, Nitesh Goyal, Matthew K. Chan, Tina Yau, John Canny, and Ulrik Schroeder. Spring: speech and pronunciation improvement through games, for hispanic children. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 47. ACM, 2010.

[32] Charles P. Thompson, Steven K. Wenger, and Carl A. Bartling. How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, 4(3):210, 1978.

[33] Zuzana Trnovcova. Chinese scrabble: a web-based speech-enabled game for chinese vocabulary building. Master's thesis, Massachusetts Institute of Technology, 2010.

[34] Endel Tulving. The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 6(2):175–184, 1967.

[35] Erik D. Van der Spek, Pieter Wouters, and Herre van Oostendorp. Code red: Triage. or, cognition-based design rules enhancing decisionmaking training in a game environment. In *Games and Virtual Worlds for Serious Applications, 2009. VS-GAMES'09. Conference in*, pages 166–169. IEEE, 2009.

[36] Luis von Ahn. Games with a purpose. *IEEE Computer Magazine*, 39(6):96–98, 2006.

[37] Vygotskiĭ. *Mind in society: The development of higher psychological processes.*

[38] Mark Wheeler, Michael Ewers, and Joseph Buonanno. Different rates of forgetting following study versus test trials. *Memory*, 11(6):571–580, 2003.

[39] Yushi Xu and Stephanie Seneff. Speech-based interactive games for language learning: Reading, translation, and question-answering. *Computational Linguistics and Chinese Language Processing*, 14(2):133–160, 2009.

[40] Yushi Xu and Stephanie Seneff. A generic framework for building dialogue games for language learning: Application in the flight domain. In *Speech and Language Technology in Education*, 2011.

[41] Yushi Xu and Stephanie Seneff. Improving nonnative speech understanding using context and n-best meaning fusion. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4977–4980. IEEE, 2012.

[42] Brandon Yoshimoto, Ian McGraw, and Stephanie Seneff. Rainbow rummy: A web-based game for vocabulary acquisition using computer-directed speech. *Proc. SIGSLaTe, Warwickshire, England*, 2009.