ISCA Archive
http://www.isca-speech.org/archive

Second Language Studies - Acquisition, Learning, Education and Technology
Tokyo, Japan
September 22-24, 2010

Multimodal learning of words: A study on the use of speech synthesis to reinforce written text in L2 language learning

Kevin Dela Rosa, Gabriel Parent, Maxine Eskenazi

Language Technologies Institute,
Carnegie Mellon University, United States of America
kdelaros@cs.cmu.edu, gparent@cs.cmu.edu, max@cs.cmu.edu

Abstract

Past research has shown that the use of multimedia, such as pictures, audio narration, and video, can be beneficial in computer aided instruction. We propose that spoken words generated by speech synthesis can be used to reinforce written text during L2 language instruction, and can lead to a more robust learning experience than providing written language input alone. Two in-vivo studies were conducted with ESL (English as a second language) students to investigate the effect of providing spoken language produced by speech synthesis during different instructional events in REAP, a computer based vocabulary tutor. Our results show that students benefit from spoken language input, particularly when they are strongly encouraged to listen to words. Furthermore, our studies seem to suggest that on demand English text-to-speech synthesis may be good enough to provide added value during computer based L2 language instruction. **Index Terms**: speech synthesis, language tutors, computer assisted language learning, English as a second language, L2 language learning

1. Introduction

Recent efforts in language learning have focused on incorporating computer technology into classroom instruction. One concern in the domain of language learning technology has been how to best incorporate different media types, such as written words, spoken words, sounds, graphics, videos and animations, in various instructional events. While the role of spoken language input is important for L1 (first language) learning to read, its role in L2 (second language) learning is less well known. We propose that two modes of input, namely written and spoken language generated by speech synthesis, during instructional events reinforce each other and result in a more robust learning experience than written language input alone.

In this paper we first discuss past work on multimodal learning, and describe the cognitive theory that motivates our claim that two modes of input can reinforce each other in instructional events. Next, we describe two comparative studies that tested our hypothesis and their associated results. Both studies were conducted using REAP, a vocabulary learning software tool that makes use of documents harvested from the internet. Finally we offer a discussion of the results of the studies and suggest future research directions.

2. Background

Past research work has shown that the appropriate use of multimedia can be beneficial in computer aided instruction [1]. In multimedia instruction, the information in a lesson is presented to students in different modes, or formats, such as text, images, or audio. An area of interest in computer aided instruction has been how the combination of different modes affects learning. In the domain of scientific explanation, Moreno and Mayer showed that students comprehended explanations best when words were presented auditorily and visually as opposed to auditorily only, given that no other visual material was provided concurrently [2]. Their results can be explained by the dual-processing theory for working memory, which states that since the auditory and visual processing channels are independent, "students can hold both representations in working memory at the same time and build referential connections between them" [2].

With respect to L2 vocabulary learning, a few different multimedia formats and environments have been explored. A study conducted by Snyder and Colon [3] showed that providing students with more audio-visual aids, in the form of audio tapes, slides, fill-in pictures, and overhead transparencies, than those in the standard curriculum lead to significantly better vocabulary retention. In another study, Neuman and Koskinen [4] compared learning vocabulary words through reading documents, reading and listening to documents, and watching television. Their results found that the participants learned and retained the vocabulary words best through watching television. Additionally, Jones and Pass [5] conducted a study with English speaking college students enrolled in a French course where students listened to a French passage using a computer program and were provided with either pictorial annotations, written text annotations, or pictorial and written text annotations. Their results showed that students who had both pictorial and written text annotations remembered the translations and the passage more effectively than students who were provided with no annotations or just one type of annotation. Lastly, Al-Seghayer found that using video clips provided in electronic glosses coupled with printed text can be an effective way of teaching unknown vocabulary words [6].

Oftentimes when spoken word input is provided to students, it is produced manually by humans, which can be a laborious and expensive process. Text-to-speech (TTS) synthesis has the potential to replace human recordings in certain applications, and has the benefit of generating speech on demand. While TTS synthesis systems have not been widely accepted in computer assisted language learning (CALL) environments, an extensive evaluation by Handley [7] suggests that the quality of current TTS systems is sufficient to add value in French language learning environments. Further research should be done to determine whether Handley's results generalize to other languages.

One may conclude that providing more modes of input will consistently lead to a better learning experience, but the cognitive theory of multimedia states that this is not always the case with redundant information [1]. Caution must be taken when presenting redundant information to a student, because providing multimedia information in the same mode could overload a student's visual or auditory channel. For example, a student can be given a diagram explaining convection and an audio narration explaining the convective process. Naively, we may assume that providing the redundant

text for the narration will help the student, but it may in fact be disadvantageous. Providing this additional information can overload the student's visual channel, effectively splitting their visual attention, since both the diagram and text have to be seen and must be simultaneously processed with the limited cognitive resources of the visual channel while the narration enters the ears and is processed by the auditory channel, whereas if we exclude the redundant text, the cognitive load is more balanced and there is minimal chances of overloading either of the student's channels.

3. Study Setup

In order to test our hypothesis, which claims that two modes of input can reinforce each other in CALL instructional events, we conducted two in-vivo studies at the University of Pittsburgh's English Language Institute, using the REAP system, where the modes of input were varied in two different instructional events. The first study focused on the effect of varying the mode of input during post-reading cloze questions. The second study focused on assessing the effect of varying the available modes of input during in-class readings. Both studies were conducted using REAP, a language tutor described in the next section.

3.1. Overview of REAP

REAP, which stands for REAder-specific Practice, [8] is a web based language tutor developed at Carnegie Mellon University that makes use of documents harvested from the internet for vocabulary learning and reading comprehensions. REAP has the ability to provide reader-specific passages by making use of user profiles that model a reader's reading level, topic interests, and vocabulary goals.

REAP's interface has a number of features that help to enhance a student's learning experience. A dictionary word lookup system is embedded in the interface which allows students to look up the definition of any of the words they encounter during readings. Another key feature in REAP is that it provides users with the ability to listen to the spoken version of any word that appears in a reading. REAP makes use of Cepstral Text-to-Speech [9] to synthesize words on demand when they are clicked during reading activities, or when a button is clicked during dictionary lookups. Finally, REAP automatically highlights the focus words, which are the words targeted for vocabulary acquisition in a particular reading.

3.2. Study 1: Comparison of Written and Spoken Input for Cloze Questions

In Study 1, we looked at the effect of providing the spoken version of a word, generated by speech synthesis, in *cloze question instructional events*. For this study we had a population of 50 ESL college students, whose native languages included Arabic, Chinese, Japanese, Korean, and Spanish. Individualized readings were given as homework, centered on 50 focus words from the Academic Word List whose written and spoken forms were available to students during the readings, followed by practice cloze questions in two conditions: answer choices in written form, and answer choices in spoken form.

A pre-test was administered at the beginning of the study, which asked students to self-assess their knowledge of words in their written and spoken form. A study by Heilman [10] has shown that self-assessment tests that ask students to make binary decisions about whether they know a word or not, such as the one used in the pre-test of this study, can effectively be used in language tutors to assess vocabulary knowledge. The post-test consisted of cloze questions with the answer choices provided in the two conditions.

3.3. Study 2: Comparison of Multimodal Input in Readings

In Study 2, we looked at the effect of providing spoken versions of a word, generated by speech synthesis, *during reading activities*. Study 1 explored the impact on vocabulary assessment tests when the mode of input during question instructional events were varied, while both modes of input, written and spoken words, were consistently available to students during the practice readings. By contrast, this study varies the modes of input available to students during the practice readings, while keeping the available mode of input during questions constant. Since students in Study 2 had the option to listen to certain words, but were not explicitly required to listen to each word, the presentation of spoken word input was more passive than in Study 1, which explicitly required students to listen to each answer choice during audio cloze questions. Therefore Study 2 was conducted as a follow up to Study 1 to see if a more passive approach to multimodal input can lead to comparable gains in vocabulary improvement.

For this study we had a population of 34 ESL college students, whose native languages included Arabic, Chinese, Japanese, Korean, and Spanish. Group readings were given as class activities, centered on 30 focus words from the Academic Word List, followed by practice cloze questions. During the readings we presented the focus words in two different ways: written form and written + spoken form provided. Students were randomly assigned to one of the conditions for each word. We hypothesized that providing two modes of input will provide a more robust learning experience.

A pre-test was administered at the beginning of the study, which asked students to self-assess their knowledge of words in their written and spoken form. Six pseudo-words were also presented with the focus words, to compensate for guesswork and overestimation of the student's vocabulary. We used a formula for the latter that penalizes a student's raw score if they claimed to know a pseudo-word [11]. The post-test consisted of cloze questions with the answer choices in their written form.

4. Results

In both studies, the general use of the REAP system significantly helped students improve their performance, as made evident by the average overall gains between the pre-test and post-test ($p < 0.01$). Normalized gain is the measure used to measure improvement in both studies, which is given by the following:

If the *post-test score* is greater than the *pre-test score*, then

$$\textit{Normalized gain} = (\textit{post-test score} - \textit{pre-test score}) / (\textit{maximum-possible-score} - \textit{pre-test score})$$

Otherwise,

$$\textit{Normalized gain} = (\textit{post-test score} - \textit{pre-test score}) / (\textit{pre-test score})$$

For Study 1, the overall average normalized gain between the pre-test and post-test was 0.310 (± 0.099). The improvement in spoken word form performance was significantly higher than improvement in written form, as

shown in Figure 1, with average normalized gains of 0.231 (± 0.097) and 0.397 (± 0.124) for the written and spoken answer choice conditions respectively ($p < 0.01$). Over all the readings, on average a student listened to 39.28 (± 13.38) unique words, with a total of 52.14 (± 17.03) synthesized words played. The average time a student spent per question was 142.9 (± 12.61) and 161.0 (± 9.677) seconds for the written and spoken word form answer choices respectively, with the average difference between the spoken and written form questions per student being 18.18 seconds ($p < 0.01$). The overall average time spent per question was 137.6 and 155.5 seconds for the written and spoken word form answer choices respectively.

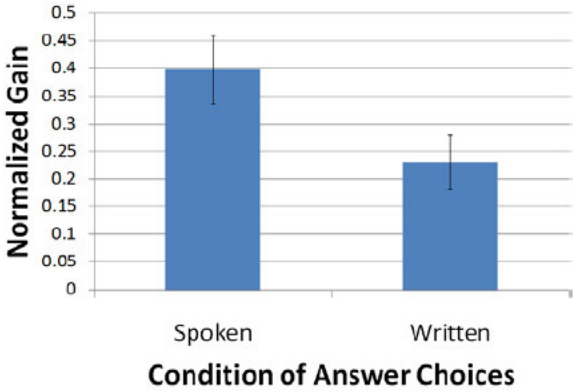

Figure 1: *Study 1, Improvement between pre-test and post-test. Error bars are standard error.*

For Study 2, the overall average normalized gain between the pre-test and post-test was 0.269 (± 0.121). The improvement in spoken + written word form condition performance between the pre-test and post-test was generally higher than improvement in written word form condition, as shown in Figure 2, with average normalized gains of 0.299 (± 0.143) and 0.368 (± 0.149) for the written only and spoken + written conditions respectively. Over all the readings, on average a student listened to 11.53 (± 2.05) unique words, with a total of 27.26 (± 6.48) synthesized words played. Table 1 compares the number of times a real word and pseudo-word was listened to during the pre-test self-assessment. Additionally, the pre-test was given to a small group of native English speakers as well, to see if the overall trends in the number of times words were listened to were the same as with the native speakers; the results are shown in Table 2.

5. Discussion

The results of our studies suggest that using two modes of input, namely written text and spoken words produced through speech synthesis, in instructional events can significantly help students improve their vocabulary, as made evident by their average normalized gains between the pre-test and post-test. In Study 1, the improvement gained from providing answer choices in spoken form, was significantly higher than providing answer choices in written form. Additionally, since both forms of the word were provided to the student in both the pre-lesson instructions and during the readings, and since in both of the answer choice condition students had, on average, statistically significant gains, providing two modes of input seems to be beneficial.

One important thing to consider when providing students with the answer choices in spoken form is that on average students take slightly longer to answer questions with spoken form answer choices as opposed to written form answer choices, as made evident by the fact that in Study 1 the average time spent on spoken form cloze questions per student was 18.18 seconds longer than the average time spent on written form cloze questions per student. In general, whether the additional time spent on questions is critical depends on the particular learning objectives of a given tutor. Therefore, there seems to be a tradeoff between the total time spent on task and improvement in auditory vocabulary performance when considering the usage of spoken word input in cloze questions.

In Study 2, the improvement under the two mode condition (spoken + written) was generally higher than the improvement on the written form condition, though the results were not as statistically significant as in Study 1. This weaker result may be the result of having less dramatic conditions in Study 2 than in Study 1, since people were not required to listen to the spoken version of words, and the stronger results in Study 1 compared to Study 2 may suggest that strongly encouraging students to listen to words (or, in the case of Study 1, requiring students to listen) can lead to better performance.

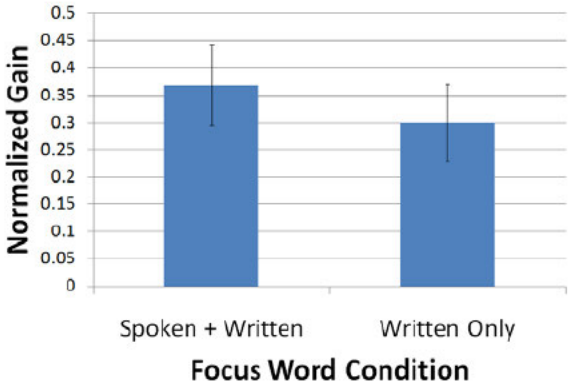

Figure 2: *Study 2, Improvement between pre-test and post-test. Error bars are standard error.*

Table 1: *Study 2, Average number of times a word was listened to during the pre-test self-assessment.*

	Known Words	Unknown Words	All Words
Actual word	1 247	1 594	1.319
Pseudo-word	1.625	1.830	1.824
All words	1 253	1.693	1.403

Table 2: *Average number of times a word was listened to during the self-assessment for native English speakers.*

	Known Words	Unknown Words	All Words
Actual word	1.073	---	1.073
Pseudo-word	2.750	1.397	1.500
All words	1.098	1.397	1.144

As expected, during the pre-test students elected to listen to words they ultimately indicated as "unknown" more often than the words they claimed to know. Additionally, students tended to listen to pseudo-words more often than actual words. Both of these trends were also observed when the pre-test was administered to a group of native English speakers, and in fact all native speakers were able to correctly identify the actual words. These results suggest that the spoken version of words generated by speech synthesis was good enough to allow

people to discriminate words they knew and words that they did not know, and that the speech synthesis for the pseudo-words is most likely *not* the reason people selected them as known. This result is encouraging for the field of CALL, since it suggests that on demand text-to-speech synthesizer systems may be good enough to give additional value to computer systems for L2 language learning under certain conditions.

6. Conclusions

We proposed that the use of two modes of input, namely written and spoken language generated by speech synthesis, can reinforce each other and result in a more robust learning experience than written language input alone for L2 vocabulary learning. Our assertion that written text and redundant spoken words would reinforce each other was based on the cognitive theory of multimedia and the redundancy principle [1]. Two studies were conducted with ESL college students to test our hypothesis.

The results of our studies have shown that students tend to benefit from spoken language input in vocabulary instruction and assessments. We recommend its usage in vocabulary tutors used in language learning laboratories, and in particular we feel that students should be strongly encouraged to listen to words. Furthermore, our studies suggest that using speech synthesis to produce the spoken versions of words can be beneficial to non-native speakers during vocabulary learning lessons, and that current text-to-speech synthesizers may be good enough for use in language learning software. Additionally, there are many aspects involved in knowing a word, such as knowing its meaning, word forms, usage, and lexical relations [12]. Our results contribute more evidence to the aspect of knowing the aural form of a word, since exposure to the spoken words during class readings and practice questions led to gains in auditory vocabulary performance. One possible caveat to using speech synthesis is that some rare words may be synthesized incorrectly, particularly in unrestricted texts; therefore special considerations may need to be taken in their use in L2 vocabulary learning settings.

One related future research problem is to see whether there is a relationship between a student's native language and their performance under visual and auditory learning conditions. We observed that when you break down the improvements in both studies by native language of the students, the trend that spoken performance is higher than written performance does not always hold. Another possible future research direction would be to systematically evaluate the difference between manual human recordings of words by native speakers and current text-to-speech synthesizers, with respect to learning the aural form of vocabulary words. Lastly, one other possible research problem is to investigate whether words with complex grapheme-to-phoneme relations are harder to learn than words with simpler grapheme-to-phoneme relations.

7. Acknowledgements

This project is supported through the Pittsburgh Science of Learning Center which is funded by the US National Science Foundation under grant number SBE-0836012. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

The authors would like to thank Betsy Davis and all the teachers at the University of Pittsburgh's English Language Institute for their participation and input in the studies. Additionally, we would like to thank Adam Skory and Luis Marujo for their help in setting up the studies.

8. Appendix: Words used in Studies

For Study 1, the following words were used:

abandon, accumulate, assume, bond, cease, cite, civil, collapse, commence, comprise, conceive, conflict, consent, controversy, convert, demonstrate, device, dimension, estimate, grant, guarantee, identical, incidence, incorporate, index, induce, legal, liberal, license, minimal, minimum, neutral, outcome, panel, participate, precise, prime, refine, restore, route, subsequent, technology, theme, theory, trace, transport, undergo, visible, welfare, widespread

For Study 2, the following words were used, with the pseudo-words underlined:

accompany, adequate, arple, boldrenite, brief, bulk, circumstance, commit, community, confine, core, debate, enormous, error, eventual, exogle, feature, final, framework, horfe, imply, namlop, network, nevertheless, option, partner, phinoscope, principle, prior, schedule, site, survive, task, ultimate, undertake, via

9. References

[1] Clark, R. C. and Mayer, R. E , *e-Learning and the Science of Instruction Proven Guidelines for Consumers and Designers of Multimedia Learning*, Jossey-Bass/Pfeiffer, 2003.

[2] Moreno, R. and Mayer, R. E., "Verbal redundancy in multimedia learning: When reading helps listening", Journal of Educational Psychology., 94(1):156-163, 2002.

[3] Snyder, H. and Colon, I , "Foreign language acquisition and audio-visual aids", Foreign Language Annals , 21(4):343-384, 1988.

[4] Neuman B. and Koskinen, P., "Captioned television as comprehensible input: Effect of incidental word learning from context for language minority students", Reading Research Quarterly., 27(1), 95-106, 1992.

[5] Jones, L. C. and Pass, J. L , "Supporting Listening Comprehensions and Vocabulary Acquisition in French with Multimedia Annotations", Modern Language Journal , 86(4):546-561, 2002.

[6] Al-Seghayer, K., "The effect of multimedia annotation modes on L2 vocabulary acquisition: A Comparative Study", Language Learning & Technology., 5(1):202-232, 2001.

[7] Handley, Z., "Is text-to-speech ready for use in computer-assisted language learning?", Speech Communication., 51:906-919, 2009.

[8] Heilman, M , Collins-Thompson, K., Callan, J. and Eskenazi, M , "Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension", Proceedings of the Ninth International Conference on Spoken Language Processing., 2006.

[9] Cepstral Text-to-Speech, Online: http://www.cepstral.com/.

[10] Heilman, M. and Eskenazi, M., "Self-Assessment in Vocabulary Tutoring", Ninth International Conference on Intelligent Tutoring Systems., 2008.

[11] Milton, J. and Hopkins, N., "Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners", The Canadian Modern Language Review., 63(1):127-147, 2006.

[12] Nation, P. and Newton, J., "Teaching Vocabulary", in J. Coady and T. Huckin [Eds], Second Language Vocabulary Acquisition, 238-254, Cambridge University Press, 1997.