The Use of “Talking Head” Video to Support Second Language Comprehension

Penny Thompson
Michigan State University

Abstract

The effects of “talking head” video (video showing the face of the speaker) on L2 comprehension were investigated. Fifty-seven university students in a Spanish language class were randomly assigned to one of three video treatment groups: still picture with audio, low quality video with audio, and high quality video with audio. Within each treatment, students were presented with a three-minute video of a native Spanish speaker talking about herself, using the vocabulary of the current class lesson. The students’ comprehension of the spoken text was measured using a 10-question fill-in-the-blank quiz. The results indicate that the presence of talking head video did not significantly affect listener comprehension.

Introduction

Constructivist learning theorists emphasize the importance of giving learners a meaningful context for the new material they are learning. Rather than memorizing facts in isolation, learners need to understand the real-world relevance of what they are learning (Jonassen, Peck & Wilson, 1999; Duffy & Cunningham, 1996). Theorists in the field of L2 acquisition have addressed this need for authentic context by shifting the focus away from the study of vocabulary and grammar in isolation, and instead stressing the importance of authentic communication in the target language. For example, Krashen’s (1981) Monitor Theory addresses the importance of authentic language inputs in combination with more structured language study. Krashen makes a distinction between language learning and language acquisition. Language learning occurs when the learner memorizes vocabulary and grammar rules, focusing on the proper form of communication. Language acquisition is a deeper, internalized ability to use the language for communication with other people. Acquisition occurs when the learner focuses on content rather than form.

According to Krashen (1981), language acquisition requires more than mere exposure to a language. The L2 input must be at the right level and must engage the learner. The ideal input is at a level just above the learner’s comprehension level, but with enough scaffolding that the learner can still understand it. He referred to this as the “i + 1” (input plus one) concept. Scaffolding in this context comes from extra-linguistic information present in the immediate environment. This idea of scaffolding to bring the learner up to the next level is similar in some ways to Vygotsky’s (1978) Zones of Proximal Development. However, the i + 1 concept focuses on the individual’s processing of language input rather than on the sociocultural aspects of learning (Kinginger, 2001).

Achieving authentic language input is more difficult for students studying a foreign language in their home country, compared to students living where the target language is spoken (Krashen, 1981). Technology has helped to bring authentic L2 input to these students. Early uses of technology for language learning included radio, television, and the language lab where students listened to cassette tapes and occasionally watched authentic video (Salaberry, 2001). These tools allowed learners to hear natural dialog in the varied accents of native speakers. Modern equivalents of these tools include authentic internet radio broadcasts and digital video streamed over the internet from a course web site. Newer technologies allow even more realistic language use. The Horizon Wimba Voice Tool application, for example, allows asynchronous conversation (via mp3 files) between learners and native speakers anywhere in the world.

These technologies all bring different aspects of natural human conversation into the language learning process, and thus all contribute to a realistic context for L2 learning. However, because they emphasize audio-only conditions (or, in the case of authentic video, dialog between characters on the screen), they are missing one element that is present in a natural face-to-face conversation: a sustained view of the speaker’s face.

Although the usefulness of seeing the speaker’s face has not been extensively studied in L2 listening comprehension, research on speech comprehension in the first language suggests that visual processing is an integral part of listening to speech (McGurk & MacDonald, 1976; Kellerman, 1990). If this visual information is also
important in L2 comprehension, then the use of audio-only listening exercises might make the listening task unnecessarily difficult. Kellerman (1990) compares these audio-only situations to listening over a telephone line, one of the most intimidating experiences language learners encounter. If seeing the speaker’s lips, eyes, and facial expressions is an important part of natural conversation, then it would make sense to include this information in second language listening exercises and proficiency tests. One way to do this would be to add what is known in the television industry as “talking head” video, a head and shoulders view that allows listeners to focus sustained attention on the speaker’s face.

If this type of video does provide a benefit to L2 learners, instructors will want to provide this material for their students and encourage students to use it frequently to practice their listening skills. Since course web sites are an increasingly important means for instructors to distribute course material, the issues related to distributing video over the internet must be considered. Digital video files are very large, and must be compressed in order to perform well for someone viewing them over an internet connection. The more a file is compressed, the smaller the file size. Since compression reduces the quality of the image, however, it is necessary to balance the competing demands of small file size versus acceptable video quality. Instructors and students would benefit from a set of guidelines for achieving the smallest file size possible without losing any benefit from the talking head video.

The purpose of the present study was to explore whether the use of talking head video facilitates L2 listening comprehension, and whether different qualities of video provide different levels of benefit. In this paper I review the literature relevant to the use of video in both L1 and L2 comprehension, describe the methods used for the research, present the results, and examine the implications of the study and some ideas for future research.

Review of Literature

Visual information in a multimedia educational program must be used carefully. Multimedia learning aids have the potential to facilitate deep cognitive processing, but also carry a risk of distracting learners or overloading their cognitive processing ability (Mayer, 2001; Mayer & Moreno, 2002). When learners receive both verbal and visual information from a multimedia presentation, they must integrate these two sources of information with each other, and also integrate them with prior knowledge. This requires a more active, generative process than receiving only verbal or only visual information, and also makes greater cognitive demands on the learner. Mayer (2001) provides this concise summary of this cognitive theory of multimedia learning: “people have separate visual and auditory channels; …the channels are limited in capacity; and …meaningful learning involves actively selecting, organizing, and integrating incoming visual and auditory information.” (p. 189) Therefore, in order to facilitate second language comprehension the video will need to help the learner organize and integrate information, and should minimize cognitive demands that increase the burden on limited working memory. To achieve this delicate balance, Mayer’s (2001) cognitive theory of multimedia learning emphasizes the integration of audio and visual information and the avoidance of any input that is not directly relevant to the learning task. It is helpful to review the research on how visual information affects language comprehension against the backdrop of Mayer’s (2001) theory.

Although the use of “talking head” video for second language comprehension has not been extensively studied, the research on speech perception in the first language provides some clues about how viewing a speaker’s face affects language comprehension. Kellerman (1990) provides a summary of the research that has been done on how infants use visual information when acquiring their first language. Visually handicapped infants take longer to produce certain sounds, such as the “b” in “boy” or the “f” in “fine,” where the articulation is easy to see. For consonants where the articulation is less visible, such as the “k” in “kick,” there is no difference between the visually impaired and the sighted infants. Sighted infants appear to observe and imitate the mouth movements of the adults around them, and to use this information to learn sounds that are considered “easy to see and difficult to hear.” While the process of learning a second language is quite different from learning a first language, this research with infants suggests that language comprehension is a bimodal process where both visual and auditory information are used.

To demonstrate the way adults use visual information in understanding spoken language, Kellerman (1990) refers to the foundational work of McGurk and MacDonald (1976). They presented adult volunteers with video segments where the speaker’s lip movements did not match the sound on the audio track. For example, a video of someone articulating “ga” was accompanied by an audio track of a speaker saying “ba.” Ninety-eight percent of the listeners reported hearing “da,” a fusion of the two conflicting pieces of information. This phenomenon is known as the McGurk effect, and demonstrates that visual information does contribute to language comprehension, at least for distinguishing phonemes.
Rudman, McCarley and Kramer (2003) demonstrated how adults use visual information to understand a speaker in a noisy room. They presented the voices of six actresses reading different texts at the same time, and asked participants to focus on one designated target speaker. The participants were better able to focus on the target speaker when they were shown video of that speaker. An analysis of eye movement revealed that it was not necessary for listeners to focus on the speaker’s lips; they only needed to see the face. This suggests that even when distinguishing specific phonemes is not possible because of extraneous noise, a view of the speaker’s face helped to focus the listener’s attention, and thus aided comprehension.

Much of the research on the use of video in second language comprehension has centered on “authentic video,” such as movies and serial television programs. Weyers (1999) studied a group of university students studying Spanish as a second language, who watched a Mexican telenovela (soap opera) as part of their regular course work. These students showed significantly greater increases in their post-test scores compared to students in the control group, who attended class for the same number of hours but were taught the standard curriculum. The author attributed these higher gains to the motivating aspect of the video and to the presence of visual cues (e.g., action, gestures, etc.) supplementing the aural information. These visual cues helped learners select and organize the information needed to understand the dialog, which resulted in meaningful learning.

Another use of video that has been studied extensively in second language comprehension is the use of a series of still photos or graphics to depict either content (the subject matter being discussed), or context (information about the speaker). The results of this research have been mixed. Ginther (2002) looked at the use of graphics on the listening comprehension portion of the Test of English as a Foreign Language (TOEFL). In this test the audio track was supplemented by a series of still photos showing either the speakers or something related to the content being discussed. The results showed that the presence of content visuals slightly improved students' comprehension. The context visuals, on the other hand, only improved comprehension when they helped to distinguish between speakers, such as when both speakers were the same sex and had similarly-pitched voices. When the speakers could be clearly distinguished by their voices, the use of a series of context visuals seemed to slightly suppress comprehension. Changing from one picture to another appears to have been distracting in situations where it provided no useful information. This is consistent with Mayer’s (2001) principle that all extraneous visuals should be avoided.

Context graphics do not appear to distract the listener when a single picture is used instead of a series. Mueller (1980) studied the use of simple line drawings to accompany a listening exercise for students learning German. A control group and two experimental groups listened to the same taped interview. The control group did not view any visual aids, but the experimental groups were shown a simple line drawing of the speakers either just before or just after hearing the audio tape. The post-test showed that novice learners were able to write a more complete summary of the interview if they had seen the drawing, while there was no significant difference in the scores of more proficient learners. The author suggests that the visual aids provided contextual information that would not otherwise have been understood by the novice listeners.

Other research has addressed the use of video with captions in the target language (Shea, 2000; Jakobsdottir & Hooper, 1995) and the use of authentic newscasts that include titles, graphics, and short segments of video (Gruba, 2004). All of these additional elements appeared to facilitate listening comprehension, although in many cases they provided more benefit to novices than to advanced learners. It is interesting to note that adding captions in the target language conflicts with one of Mayer’s (2001) principles. When information is presented in the listener’s native language, text on the screen that merely duplicates the audio track is a distraction that can suppress learning (Mayer, 2001). In the second language the result appears to be the opposite. It may be that for L2 listeners the scaffolding provided by captions is beneficial enough to compensate for the cognitive burden of the redundant text.

Sueyoshi and Hardison (2005) addressed the use of “talking head” video directly when they studied the effects of facial cues and gestures on a group of English language learners. They studied the effect of video on listening comprehension for university students enrolled in English as a second language classes. The participants were identified as either high or low proficiency English speakers, and participants from each level performed one of three listening tasks. One group watched a video that included the audio track and a close view of the speaker’s face. The second group watched a video that used the same audio track, but showed both the face and the hand gestures of the speaker. The control group listened to the audio track with no video. The listening exercise was followed by a multiple choice comprehension test. The results showed that the low proficiency group had the highest scores on the post test when they could see both the face and the gestures of the speaker. For the high proficiency group, the highest scores occurred in the group that saw only the speaker’s face. The audio-only condition showed the lowest test scores for both proficiency levels. The authors suggested that the higher proficiency learners’ greater level of experience with English enables them to use visual information from the
speaker’s face more efficiently. The fact that learners of both proficiency levels showed the lowest performance under the audio-only condition, however, supports the hypothesis that seeing the speaker’s face facilitates comprehension.

Sueyoshi and Hardison (2005) focused on English language learners who were living in the United States and thus immersed in their target language. The learning conditions are different for native English speakers learning a second language in the U. S., where their exposure to the target language is limited. It would be useful to know if there is a clear benefit to providing talking head video to L2 learners in this situation. Because of the challenges of delivering high quality video over the internet, it would also be useful to develop guidelines for minimum standards for video quality required to realize those benefits. Therefore, this study addressed the following questions:

1) does the use of talking head video facilitate listening comprehension for second language learners?
2) is the facilitating effect of talking head video, if there is one, different at different levels of video quality?

Method

This chapter presents the method used in this study, including a description of the participants, materials, experimental design, dependent measures, and data analysis.

Participants

Ninety students from an accelerated beginning Spanish class at a large midwestern university were invited to participate in this study. Sixty-six students volunteered and signed consent forms, and fifty-seven of those students completed both the pre-test and post-test components of the study. The sample included 37 females and 20 males. The average age of participants was 19.15 years (SD = 1.46), with two participants declining to report their age. The course fulfills the language requirement for all undergraduates in the College of Liberal Arts. Students were placed in this course based on a standard placement test created by the Spanish department and administered by the university’s Language Testing Center.

The participants were enrolled in four different sections of the class and taught by two different teaching assistants, who were supervised by the same senior lecturer. All sections of the course used the same text book (Conexiones, 3rd Edition), covered the same material, and used a standardized WebCT Vista course web site. The video exercises that made up this experiment were part of the regular curriculum for this course, so all students completed the exercises, but the instructors provided the researcher with the scores for only those students who signed consent forms to participate in the research.

Materials

The materials used in this study consisted of two three-minute videos featuring natural speech from native Spanish speakers, and two ten-point comprehension quizzes corresponding to the two videos.

Videos: The videos were each three minutes long and featured native Spanish speakers talking about a topic related to the current week’s lesson. The speakers deliberately used vocabulary and grammatical constructions (e.g. verb conjugations) that had been covered in the lesson. The first video featured a male speaker from Puerto Rico talking about his favorite sports. This video, which came from existing course material, was prepared in QuickTime format and was provided online to the students on the course website.

The second video featured a female speaker from northern Spain, who described her personality. This video was prepared explicitly for this study. It was recorded in a quiet room using a Panasonic model AG DVC7 video camera with an external PZM microphone, edited using iMovie. The video footage was exported from iMovie and compressed for playback using the Flash 8 Video Encoder with the Sorenson Spark codec. The dimensions were 540 by 360 pixels. Three versions of the video were prepared. One version featured only a still picture of the speaker along with the audio track. The audio track was extracted from the source video footage, and a still shot of the speakers face (exported from iMovie in .jpg format) replaced the motion video. The second version included full-motion video processed with a frame rate of 12 frames per second and the “low” quality setting of Sorenson Spark. The low quality setting and low frame rate resulted in video that was clear when the speaker sat still, but became slightly blocky when she moved. The third version also included full-motion video, but this version was processed using the high quality setting and a frame rate of 30 frames per second. The quality of this video appeared comparable to a commercial DVD.
Comprehension Quizzes: The pretest quiz, which corresponded to the first video, consisted of two multiple choice and three short answer questions that asked participants to write down what the speaker had said. Each question was worth two points, so the highest possible score was ten points. Participants took this quiz using the quiz feature in the course web site.

The post-test quiz, corresponding to the second video, consisted of ten fill-in-the-blank questions, each worth one point, where participants were required to write in Spanish the exact word they had heard the speaker say. This second quiz was taken in class using paper and pencil.

The instructor and one teaching assistant reviewed the test questions to assess validity. The original draft of the post-test quiz included a combination of short answer and multiple choice questions, but there was some concern that participants might guess the correct answer even if they didn’t understand the speaker. Therefore, the quiz was revised and all of the multiple choice questions were changed to short answer format. The K-R 21 reliability for the post-test quiz was .75.

Procedure

The experiment took place in a computer lab on campus, with all of the students using Macintosh G5 computers. Participants were randomly assigned to one of the three treatment groups, and upon entering the computer lab, were assigned a specific computer based on their treatment group. The arrangement of the lab allowed participants to be seated in clusters with their group so they could not see the computer screens of the other groups. They watched the video on their computer screens and listened using headphones. They were instructed to watch and listen to the video once before seeing the quiz questions. After they had listened once, the quiz was passed out and they were given ten minutes to complete the questions, replaying the video as needed during the time allowed. The instructor and the researcher observed the class to make sure that each participant worked independently.

Design and Analysis

A one-way analysis of variance (ANOVA) was used to identify any significant differences in post-test scores between the three groups. Alpha was set at .05. The independent variable was video type: still picture, low quality motion video, or high quality motion video. Participants were randomly assigned to only one video type.

A second factor, Spanish proficiency level, was initially considered in the design of this study. However, the pre-test data did not provide any evidence that the participants had significantly different proficiency levels. In addition, the sample size was small, with only 63 participants taking the pre-test, so dividing them by both proficiency level and video type would have resulted in groups that were too small for valid statistical analysis. The dependent variable was learner comprehension as measured by the comprehension quiz.

Results

In this section I report the comprehension scores for the three video treatment types. The dependent variable was the number correct answers on the comprehension quiz. The quiz scores were normally distributed for the Still Picture group, Shapiro-Wilk(19) = .919, p = .110, and the Low Quality Video group, Shapiro-Wilk(18) = .930, p = .197. Scores for the High Quality Video group were somewhat negatively skewed, Shapiro-Wilk(20) = .823, p = .002, partly because one participant earned a score of zero. After confirming with the instructor that the zero was an actual earned score, rather than a missing score, I elected not to remove it from the data set. Table 1 shows the means and standard deviations for each of the treatment groups.
Table 1
Means and Standard Deviations by Treatment

<table>
<thead>
<tr>
<th>Video Type</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Still Picture</td>
<td>19</td>
<td>6.95</td>
<td>2.51</td>
</tr>
<tr>
<td>Low Quality Video</td>
<td>18</td>
<td>6.78</td>
<td>2.24</td>
</tr>
<tr>
<td>High Quality Video</td>
<td>20</td>
<td>7.55</td>
<td>2.78</td>
</tr>
</tbody>
</table>

ANOVA indicated no significant effect of video type on listener comprehension, $F(2,54) = .50$, $p>.05$. Table 2 shows the results of the ANOVA.

Table 2
ANOVA Performed on Comprehension Quiz Scores

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>MS</th>
<th>MSe</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Quality</td>
<td>2</td>
<td>3.18</td>
<td>6.39</td>
<td>.50</td>
<td>.611</td>
</tr>
</tbody>
</table>

Because the ANOVA did not show any significant differences, no follow-up tests were done.

Discussion

This chapter presents a discussion of the results of the study. The purpose of the study was to see if the use of talking head video, showing the face of the speaker, would improve listening comprehension for L2 learners, and to explore whether the quality of the video had any effect on listening comprehension. Participants completed an exercise under one of three different conditions – still picture with audio, low quality video with audio, and high quality video with audio – and their scores on a listening comprehension quiz were compared.

The results showed no significant differences in the listening comprehension scores between the three treatment groups. Therefore, the hypothesis that talking head video facilitates L2 listening comprehension was not supported. Furthermore, the hypothesis that higher quality video has a greater effect than low quality video was not supported.

There are several possible explanations for the non-significant result found in this study. One is the low level of Spanish language proficiency of the participants. Sueyoshi and Hardison (2005) found that listeners with higher language proficiency were better able to take advantage of the visual cues provided by watching the speaker’s face, possibly because they had had more L2 interaction experience and had learned to recognize visible speech cues. Participants in the present study came from an accelerated first-year, first semester class. Although they had taken at least one year of high school Spanish, they had been identified by the Spanish department’s placement test as being low-proficiency learners. It is possible that the results would be different if the experiment were repeated with participants whose Spanish language skills were at the intermediate level or higher.

A second possible explanation is that the three-minute video may not have been long enough to measure the effect of the video. According to Mayer (2001), one of the benefits of providing relevant visual information is
that it helps the learner focus attention on the content of the lesson. This benefit might be more important as fatigue begins to interfere with listeners’ ability to focus attention on detailed information such as specific vocabulary words from the text (van der Linden & Eling, 2006). A possible follow-up study would be to repeat the same experiment using a longer video text.

A third possible explanation for the lack of significant finding is the relatively small sample size. There was a high absentee rate the day of the experiment, perhaps because it occurred during a week that was shortened by a holiday, and because the use of the computer lab required students to attend class in a different building from their regular classroom. A sample size of 30 is often cited as the ideal for statistical analysis, but the groups in this study ranged from 18 to 20.

A fourth possible reason is the limitation of the testing instrument used to measure comprehension. Although an original draft of the comprehension quiz was modified by the instructor to replace the easy-to-guess multiple choice questions with fill-in-the-blank questions, the data still show a ceiling effect. Twenty-one percent of the participants earned the maximum score of ten correct answers. This suggests that the test was not sensitive to variability among the highest-performing students, and therefore had a limited ability to reveal differences between treatment groups.

It is possible that the ceiling effect in these data could have been prevented by making the comprehension quiz even more difficult. However, the listening exercise used in this experiment was an in-class activity with a stated goal of giving students an opportunity to practice their developing Spanish language skills, rather than testing their achievement. The video gave them a taste of how the Spanish language is used outside the classroom, which is consistent with Krashen’s (1981) recommendation to provide a guided but informal environment for language acquisition. The introduction of a stringent assessment component into a class activity that is normally used as a rehearsal of skills might have been discouraging for the students. A future study with a more difficult test would be valuable, but the present study reflects an authentic classroom activity, and therefore may have more relevance to how video would actually be used in L2 teaching.

Although there might be significant effects of video type that were not found in the present study, it is also possible, of course, that there really are no significant effects of talking head video on L2 comprehension. Although disappointing in the context of the present study, language instructors may be encouraged by the fact that this study did not reveal any negative effects from using simple pictures or low quality video. Due to the large file size of video and the bandwidth required for viewing video over the internet, instructors are often limited to posting highly-compressed, low quality video on their course web sites. Despite this limitation, however, students like using video materials in their foreign language classes (White, Easton & Anderson, 2000). The attractiveness of video might make the listening exercise appear easier and put students at ease (Norman, 2004), which could increase students’ motivation to learn. Because the present study did not find a significant difference between still picture, low quality video, and high quality video, it does not suggest any requirement that the video used in foreign language classes be of high visual quality.

Although the results of this study were statistically insignificant and therefore inconclusive, this project may serve as a pilot study on a topic that is worthy of further exploration, especially with more advanced learners and using longer, more complex video materials and testing instruments.

References


