Jesse, A., Vrignaud, N. & Massaro, D. W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. Interpreting, 5, 95-115.

Running head: MULTIPLE SOURCES IN SI

The processing of information from multiple sources in simultaneous interpreting

Alexandra Jesse, Nick Vrignaud, Michael M. Cohen, and Dominic W. Massaro
Perceptual Science Laboratory
Department of Psychology
University of California, Santa Cruz

Abstract

Language processing is influenced by multiple sources of information. We examined whether the performance in simultaneous interpreting would be improved when providing two sources of information, the auditory speech as well as corresponding lip-movements, in comparison to presenting the auditory speech alone. Although there was an improvement in sentence recognition when presented with visible speech, there was no difference in performance between these two presentation conditions when bilinguals simultaneously interpreted from English to German or from English to Spanish. The reason why visual speech did not contribute to performance could be the presentation of the auditory signal without noise (Massaro, 1998). This hypothesis should be tested in the future. Furthermore, it should be investigated if an effect of visible speech can be found for other contexts, when visual information could provide cues for emotions, prosody, or syntax.

In press, *Interpreting*

Biosketch

Alexandra Jesse is a doctoral candidate in the cognitive psychology program at the University of California, Santa Cruz. She received her MS (2000) in psychology at the University of California, Santa Cruz and completed her undergraduate studies (1998) in psychology at the Philipps-University in Marburg, Germany. She works as a member in the Perceptual Science Laboratory on bimodal speech perception and written word recognition. She is also interested in investigating and modeling the time-course of language perception processes.

Nick Vrignaud received his BA in psychology at the University of California, Santa Cruz in 2001. The third experiment reported here was part of his senior thesis.

Michael M. Cohen is a research associate in the Perceptual Science Laboratory at the University of California - Santa Cruz. His research interests include speech perception and production, speechreading, information integration, learning, and computer facial animation.
Dominic W. Massaro is Professor of Psychology and director of the Perceptual Science Laboratory at the University of California, Santa Cruz. He is currently the book review editor of the American Journal of Psychology and co-editor of the journal Interpreting. His research uses a formal experimental and theoretical approach to the study of speech perception, reading, psycholinguistics, memory, cognition, learning, and decision making. One focus of his current research is on the development and theoretical and applied use of a completely synthetic and animated head for speech synthesis, language tutoring, and edutainment.

We humans can perform the most amazing psycholinguistic feats, spanning the range of skills from identifying phonemes to simultaneous interpretation (SI). This paper is concerned with how we use multiple sources of information not only in simple language perception and understanding but also in SI. There is now good evidence that language processing is influenced by multiple sources of information (Massaro, 1998; Massaro & Shlesinger, 1997). Understanding spoken language is constrained by a variety of auditory, visual, and gestural cues, as well as lexical, semantic, syntactic, and pragmatic constraints. Research questions for psycholinguists and speech scientists include the nature of the sources of information; how each source is evaluated and represented; how the multiple sources are treated; whether or not the sources are integrated; the nature of the integration process; how decisions are made; and the time course of processing. Research in a variety of domains and tasks supports the conclusions (for summary see Massaro, 1998) that a) perceivers have continuous rather than categorical information from each of these sources; b) each source is evaluated with respect to the degree of support for each alternative; c) each source is treated independently of other sources; d) the sources are integrated to give an overall degree of support for each alternative; e) decisions are made with respect to the relative goodness of match among the viable alternatives; f) evaluation; integration; and decision are necessarily successive but overlapping stages of processing; and g) cross-talk among the sources of information is minimal. The fuzzy logical model of perception (Massaro, 1998; FLMP), which embodies these properties, gives the best extant accounts of language processing (see Figure 1).

Insert Figure 1 about here.

Consistent with the FLMP, there is a large body of evidence that bimodal speech perception is more accurate than unimodal perception (Massaro, 1998). This result is found for syllables in isolation, as well as for words and sentences. To illustrate the value of the face in perception and understanding, we present first some results from a new set of studies in which the bimodal vs. unimodal processing of sentences was investigated.

Experiment 1

Method

Participants. Seventy-one students recruited from psychology courses at UCSC participated in the experiment, which lasted about 50 minutes.

Stimuli and Material. The test items consisted of 65 meaningful sentences from the Central Institute for the Deaf (CID) set (Davis & Silverman, 1978), e.g. "We will eat lunch out". The sentences were 3, 4, and 5 syllables in length and consisted of 43 statements, 17 questions, and 5 imperatives.

Procedure. On each trial, one of the two talkers and one of the 65 sentences were randomly selected for presentation. The participants were asked to watch and listen to each sentence and to type in as many words as they could for each sentence. There were two presentation conditions: auditory-alone mixed with speech noise (Grason Stadler noise generator) or bimodally, with these same sentences with a video of a talker. There were also two talker conditions: In natural talker condition, the original natural auditory speech was used with a natural head Gary, a radio announcer (Bernstein & Eberhardt, 1986). In the synthetic talker condition, synthetic auditory speech using the AT&T Flextalk TtS or natural auditory speech was aligned with Baldi, a computer-animated head (Cohen & Massaro, 1993; Cohen, Walker & Massaro, 1996; Massaro, 1998). The audio level was adjusted separately for the two talkers in the mixer to achieve approximately the overall same level on the audio-only trials on the basis of an earlier pilot experiment. In sum, there were 2 talkers times 2 modality conditions times 65 sentences, for a total of 260 trials, occurring in two 20-minute sessions of 130 trials each with a short break in between.

Apparatus. The stimuli were presented either as auditory-alone embedded in speech noise or these same sentences presented with visual information. Two talkers were used: (a), a professional speaker recorded on video disk (Bernstein & Eberhardt, 1986), and (b), our synthetic talker called Baldi (Cohen & Massaro, 1993; Cohen et al., 1996; Massaro, 1998). The video was played on a SONY LDP-1500 laserdisk player while the synthetic talker was generated in real-time at 30 frames/sec on a SGI Crimson-Reality Engine. The audio/video signal from the computer or the laser disk was selected by a PANASONIC MX-50 mixer under computer control, and presented on JVC TM-131SU 13" monitors. Particpants typed their response via TVI-950 terminals connected to the computer. All experimental events and data collection were controlled by the SGI computer.

Results

Given that only Baldi will be used in the following experiments on SI, we limit our analyses to those unimodal and bimodal trials relevant to the advantage provided by having Baldi present

during the auditory speech. For each participant, the results were scored in terms of the proportion of words in the stimulus sentences that occurred in the responses. An analysis of variance oon the percentage of words recognized correctly shows that word recognition in sentences benefits from the additional source of information, the visible speech, $F(1,70) = 451$, $p<.001$, one-tailed. If the auditory speech was aligned with Baldi, the participants recognized an average of 66% of the words. Without this additional information, recognition was only 45%.

Figure 2 shows the performance accuracy for these two conditions for each of 71 participants. The large variability in performance is partly due to individual differences and partly because the participants were tested under a variety of different experimental conditions (such as whether the auditory speech was natural or synthetic, and the amount of noise that was added).

As can be seen in Figure 2, the proportion of correct words was higher for the bimodal than the unimodal condition for all 71 participants. Every single participant benefited from the presence of Baldi during the auditory speech. The advantage provided by the visible speech varied greatly across individuals. Some persons benefited more than others, but much of the size of the benefit can be explained in terms of their performance on the unimodal auditory speech. The visible speech gave a larger benefit to the extent that the correct recognition of the auditory speech was poor.

Insert Figure 2 about here.

Discussion

These results illustrate the value of visible speech in the perception and repetition of short sentences. Adding visible speech gives a larger benefit to the extent that the correct recognition of the auditory speech is poor. The question pursued in the present paper is to what extent visible speech also benefits SI.

Within the framework of the FLMP, we expect that a bimodal presentation would lead to better SI than a unimodal one. A central assumption of the FLMP is that language processing is best described within an information-processing approach in which multiple simple processes interact with one another to produce fairly complex behavior. Following this logic, a fundamental hypothesis would be that the mechanisms supporting language perception and comprehension operate fairly independently of the mechanisms of language production and communication. This

hypothesis is particularly relevant to understanding conversational dialogs but most critically central to simultaneous interpretation.

Given that the FLMP has not been developed extensively to account for both the perception and production of speech, we looked for an existing cognitive architecture that addressed these two components while maintaining the spirit of the FLMP. The executive-process interactive control (EPIC) architecture, developed by Meyer and Kieras (1999) and their colleagues (Schumacher et al., 2001), provides a testable framework for SI. Firstly, their research and theory development should be applauded because it confronts the difficulty of accounting for multiple interacting processes (which is a hallmark of SI). A distinguishing characteristic of the EPIC model is independence in the sense that multiple streams of processing activate relevant responses independently of other streams of processing. Thus, this model stands in sharp contrast to the common assumption of limited capacity—a limit of resources that constrains performance in all situations. One common but justifiable criticism of limited capacity is that it is not a productive explanation either in terms of understanding or for a viable research agenda. The only constraints in EPIC are physical ones at the input or output level. For example, if you are trying to understand a person speaking Spanish, a neighboring conversation in English can have a significant impact because the simultaneous auditory message will necessarily degrade the auditory input from the Spanish speaker. Similarly, if your mouth is busy talking, it is not possible to be chewing gum at the same time.

The independence assumption of the FLMP and EPIC stipulates that complete and optimal processing of the source language would necessarily support better interpretation than would reduced processing of this input. On the other hand, a nonindependence assumption such as a limited-capacity attentional view might predict that a more shallow processing of the source language would free up capacity for interpretation (language production). A theoretical and practical test of this hypothesis is whether watching the source language as well as listening to it leads to higher-quality interpretation than simply listening to the source language. Analogously, we can ask whether bimodal speech processing produces better simultaneous interpretation than unimodal processing. If the mechanisms of perception operate independently from those of production in terms of their capacity, then two input sources should lead to better perception and therefore to better production. Processing of a second source of information, namely the face, should not take away capacity from interpretation.

A limited-capacity interpretation might predict, however, that interpretation should be poorer in the bimodal relative to the unimodal condition. An example of such a limited-capacity

model in SI is the Effort model for SI by Gile (1997). The model assumes that there are three nonautomatic "efforts" involved in SI that add up to the total effort. One of the efforts involves all comprehension processes, such as used for the initial signal analysis, the comparison with words stored in memory, and the final recognition process. This effort would encompass all processes described by the traditional FLMP. In addition, the production effort includes all processes that map this mentally recognized speech signal into a speech output. Finally, the memory effort refers to capacity used to store the signal in the short-term memory while processing. If the interpretation task requires less capacity for each effort available, the interpretation proceeds smoothly. For the present experiment, we would assume that the memory and production effort is the same in both conditions. However, the comprehension effort varies depending on presentation with the speaker in the unimodal (auditory only) or bimodal (auditory and lips) condition. Gile's description of interpreting (1997, 1999) is underspecified in terms of which processes of the comprehension effort are automatic and which are not. The essential question is whether or not the additional information from the lips can be processed without additional effort, as predicted by the FLMP and as shown by psychological studies (for an overview, see Massaro, 1998). If no additional effort is needed when processing the lip movements, then there should be no difference in performance between the two experimental conditions. If additional effort is needed, and if the comprehension effort for the bimodal condition is greater than the comprehension capacity, performance in the bimodal condition should be worse than in the unimodal condition.

There have been only a few studies evaluating the influence of a view of the speaker in simultaneous interpreting. Balzani (1990) found that interpreters produced significantly fewer errors when simultaneously interpreting from a speaker displayed on video then when just receiving auditory information (French to Italian). This effect was only found when the presenting speaker was recorded while given a speech but not when reading a prepared text. However, the video of the speaker showed not only the face but also a complete view of the speaker. It is unclear from this study if it was specifically the information obtained from the lips that facilitated performance or whether information from gestures and facial expressions was responsible.

Anderson (1994) tested simultaneous interpreting when a bimodal video of the speaker was presented or just their auditory speech (French to English). Although the paper does not described what exactly was shown on the videos, the author's motivation for the study was to examine the effect of the "visual context", and he discussed the role of gestures and facial expressions. Thus, it can be assumed more than the face was shown. Professional interpreters' performance was assessed in terms of the intelligibility and informativeness of the translation. No difference in simultaneous

interpreting performance was found. However, participants in this study reported that they did not always watch the video screen as instructed.

To see if information from the lips improves simultaneous interpreting, we carried out the following experiments. Participants listened to short paragraphs in English while they had to translate these as quickly as possible (simultaneously) into their first language, which was either German or Spanish. These short texts are presented only auditorily or aligned with the lip-movements of our computer-animated talking head, Baldi. Baldi displayed no facial expressions. The independent variable is the presence or absence of the face and the dependent variables are various accuracy measures of the translation.

We believe that SI should benefit from multiple sources of information. Therefore, interpreting should be improved by the addition of a talking head accompanying the auditory speech. On the other hand, if multiple sources of information hurt performance because of limited processing capacity, then performance should be worse in the bimodal than unimodal condition.

Experiment 2

Method

Participants. Twelve native German speakers participated in the experiment. Although only one of them was considered fully bilingual, all of the participants were fluent in English and had lived in the U.S. at least for the past year for their studies and/or work. Their time in the U.S. ranged from 1 to 7 years, with an average of 2.8 years. All participants were affiliated with the University of California, Santa Cruz (visiting exchange students, graduate students, and staff). Their average age was 28.1 years. Five of the participants were female and seven were male. Participants were paid $8 for their help.

Stimuli Material. Eight short passages from a book called "The House on Mango Street" by Sandra Cisneros (1991) were chosen as stimuli. The texts were slightly edited wherever translations into German seemed to be impossible (e.g. there is no German word for "marshmallow") or the vocabulary seemed to be extremely difficult. The average length of the passages was 16 clauses, or 98 words.

Design and Procedure. The order of presentation for the eight texts was random and each text was also randomly assigned to either a unimodal bimodal presentation condition, with the

constraint that for each of the participants half the texts were presented bimodally and the other half unimodally. After an initial block of 4 unimodal and 4 bimodal passages, all 8 passages were repeated in the same order but with the opposite condition of presentation modality. For example, if the participant first listened to a given text in an auditory-only condition, then the same text participant would be presented later in a bimodal condition. By completing these two blocks, each participant was exposed to all cells in the experimental design.

Participants were also yoked in pairs: Every second participant had the same order of presentation as the previous one, but texts were shown in opposite modality conditions. Thus, the second participant of each pair had the same order of texts as the first participant but with the opposite modality: every second participant saw first block 2 and then block 1 of the previous participant. Across all twelve participants, each text was presented six times in each presentation condition (unimodal or bimodal) in each of the two blocks.

Each participant was first interviewed in German by a native speaker (AJ) about their English and German skills. The participants were asked about how well they could write, read and understand both languages and also on the extent of exposure to the languages. Another goal of the interview was to give participants practice in speaking in German, since lacking a German community in Santa Cruz, the participants rarely spoke German.

The participants then read the instructions in English. They were told that the focus of the study was simultaneous interpretation and that their task is to listen to several short passages in English and simultaneously translate them orally into German. It was emphasized that it was important that they watch the computer screen throughout the task. They were also warned that the passages might be repeated. Participants were informed that on some trials they might only hear a voice whereas on other trials a face would accompany the voice. Furthermore, they were instructed to simply restart translating at any point, in case they fell behind or had difficulties. The participants were encouraged to give their best in this difficult task. Participants were free to take short breaks between passages (while the program was waiting for them to continue with a key press). Then the experimenter emphasized again, that it is important to look at the screen during the whole experiment. It was emphasized that although some people might think they should look on the floor or close their eyes to concentrate, the participants should instead look at the screen.

Each participant was seated directly in front of the computer and instructed on how to use the mouse to work through the experiment. Voice recording occurred during the complete session.

The experiment was started by the participant by clicking with the mouse on a “Begin experiment” button after the experimenter had left the testing cubicle. The first text was then presented. After presentation of each text, a button labeled “continue” appeared on the screen. Participants clicked on the “continue” button in order to proceed to the next text. Every participant was exposed to a total of 16 passages, and attempted to provide simultaneous oral interpretations for each of them. When each participant completed the experiment, the experimenter entered the testing cubicle, stopped recording, and thanked and debriefed the participant.

Apparatus. Baldi, our animated conversational 3D agent, was used within the Center for Spoken Language Understanding (CSLU) Speech toolkit. The CSLU toolkit is a comprehensive set of tools for researching spoken language (see e.g. Sutton et al., 1998, for description). It includes a rapid application developer (RAD), text-to-speech synthesis, and Baldi. Baldi’s visual speech is presented through facial animation synchronized with synthetic auditory speech (Massaro, 1998). Baldi’s facial animation can be driven by and synchronized with speech synthesized from text, as done for the present experiment, or speech recorded by a human speaker. In this experiment, Baldi’s auditory synthesized speech was driven by the FESTIVAL 1.3.1 text-to-speech synthesis (TTS) system (Black & Taylor, 1997). The texts were presented at about 137 words per minute.

The participant’s translations were recorded online via a Plantronics headset microphone connected to a Sharp minidisk portable recorder (MD-MS701H(S)2). The same headset was used to present Baldi’s auditory speech, which was presented at a constant comfortable listening intensity.

Baldi was displayed in a 7.4” by 9.4” window (Zoom 36). The parameters for the window were 200 for Near Clip and 800 for Far Clip, adjusting Baldi’s distance. Baldi’s auditory speech had a basic pitch of 110Hz with a range of 19Hz. Baldi was displayed in a frontal view with eyes, lips, skin, teeth and tongue, but with neither ears nor the back of his head. He also blinked and showed saccades during speaking. His emotional display was set to 100% neutral. The target frame rate was 30.

Scoring. All responses were later transcribed by a native German speaker (AJ), who was not cognizant of the actual experimental condition of each text translation. The original texts were divided in clauses, and for each clause the goodness of the translation was rated. The translations were described as either more or less correctly translated, ranging from perfect translations to translations with minor grammatical or content errors, or as incorrectly translated. Out of all translations that were perfect or had minor grammatical or content errors percentage correct was

calculated. Incorrect complete translations were used to calculate the percentage of incorrectly translated clauses. We also looked at how many clauses were attempted overall, whether the participant completed the translation or not. We also calculated the percentage of clauses that have an incomplete translation. A translation was scored as incomplete when the participant stopped interpreting after a few words of a clause. These were not just clauses that missed words, but were actually too short to rate on goodness of translation. For this reason, incomplete translations were not included in the ratings for goodness of translation. Percentages of all these measures were calculated based on the number of clauses in the original text. The total number of spoken words was also recorded for each participant. We did not analyze some other possible dependent variables, such as number of false starts, inclusion of English words in the translation, or number of repetitions, since their overall occurrence was very low.

Results

One-way ANOVAs with modality of presentation (2 levels) as a within-participant factor was conducted. The dependent variables were percentage of correct translations, percentage of incorrect translations, percentage of clauses attempted, percentage of incomplete translations, and number of spoken words. We analyzed only block 1, since we were not primarily interested in the effect of learning or the influence of text repetition, but simply the influence of facial information on interpreting performance.

Table 1 shows the individual participant means for percentage of correctly translated clauses over all texts presented under each of the modality conditions. For seven out of the twelve participants, the bimodal presentation helped performance, although modality had no statistically significant influence overall.

Insert Table 1 about here.

Table 1 also indicates that our participants were able to handle this complex and difficult task successfully. Although no direct comparisons to professional interpreters are available, we are confident that a grand mean of 64% on percentage of combined goodness of translation is an impressive performance. This means, that on average, our novice participants, who were not even

true bilinguals, were able to translate 64% of all clauses in the original text more or less correctly. This performance differed from an average of 63 % for unimodally presented paragraphs to 65.2% correct for bimodally presented paragraphs. Although this difference did not reach significance, the trend is consistent with the prediction of the FLMP.

Even though the means of all dependent variables for the two modality conditions showed small trends in direction predicted by the FLMP, there was no significant modality effect for any of the dependent measures. There were slightly more correctly translated clauses (*M*=65.17%) in the bimodal condition than in the unimodal condition (*M*=62.50%). Similarly, there were slightly more correctly spoken words (*M*= 68.33) in the bimodal condition than in the unimodal condition (*M*=67.67). The percentage of incorrectly translated clauses (*M*=3.88%) and the percentage of incomplete translations (*M*=6.56%) were slightly higher in the unimodal condition than in the bimodal condition (*M*=3.75%; *M*=4.77%).

Discussion

The results of the first study indicated that facial information did not seem to significantly improve simultaneous interpretation. However, it also does not seem to hinder the simultaneous interpreting process. One concern with our choice of participants in this experiment could be that they were, except one, not true bilinguals. The participants had learned English as their second language in high school for 7.5 years on average. Given this classroom learning, their exposure to English was probably mainly auditory, so it is possible that these participants never learned to use the cues provided by lip-movements for perception of English. In our second experiment, we therefore tested self-claimed bilinguals, namely native Spanish-American English speakers. Design and text stimuli were identical to the first experiment. Again, we hypothesized that, according to the FLMP, perception and therefore interpretation performance should be better when texts were presented bimodally rather than unimodally.

Experiment 3

Method

Participants. Eight undergraduate students (4 male, 4 female) participated in the experiment. The participants were undergraduate psychology students at the University of

California at Santa Cruz, and received course credit for participation. Participants all claimed to be bilingual and proficient in both Spanish and American English.

Stimuli Material. The same eight passages as in the first experiment were chosen. However, the original texts were presented, not the slightly modified version from the first experiment. The average length of the texts was 16 clauses, ranged from 87 to 114 words in length.

Design and Procedure. The design of experiment 2 differed from experiment 1 in the sense that a sequence of altering unimodal and bimodal presentation conditions was followed, again with the constraint that half of the texts were presented in each of the two presentation conditions (in the first experiment, the order of presentation was completely random). As in the first experiment, texts were again randomly assigned to their conditions. Participants saw two blocks of the eight texts with alternating modalities but with the same order for the second block. Basically, for every second participant texts were shown in the same order and modality condition within a block as the previous participant, but the order of blocks was reversed. Each text was presented four times in each presentation condition (unimodal or bimodal) in each of the two blocks over the eight participants. The procedure of the experiment was identical to the first experiment, except that no prior questionnaire was given.

Scoring. All participants' responses were transcribed and rated for their goodness by a bilingual Spanish speaker, who was unaware of the actual experimental condition of each text translation and was paid for her work. The original texts were divided in clauses similar to the first experiment and for each clause the goodness of the translation was rated. For the analysis of this experiment, we chose percentage of clauses correctly translated, percentage of clauses incorrectly translated, and percentage of clauses attempted. All percentage scores were based on the overall number of possible clauses in the original text.

Results

Average scores for each participant for each modality condition were determined by averaging performance across the different texts. A one-way Analysis of Variance with modality of presentation (2 levels) as a within-participants factor was conducted with the three dependent variables described above for each block.

Table 2 shows the individual participant means of percentage of correctly translated clauses over all texts presented under each of the modality conditions separately for both blocks.

Four of the participants show better or equally good performance for texts presented bimodally in comparison to texts presented only auditorily at time 1. At time 2, five participants show a trend in accordance to the FLMP's prediction that participants benefit in their performance from multiple sources of information, although overall there is no significant effect of modality.

Insert Table 2 about here.

For the first block, we did not find any significant modality of presentation effect for any of the three dependent variables. The trend in the means was small, but in the predicted direction There were more clauses attempted in the bimodal condition (*M*=64.13%) than in the unimodal condition (*M*=63.37). There were also slightly more incorrectly translated clauses for the unimodal presentations (*M*=6.86) than in the bimodal presentations (*M*=8.00%). However, there were slightly more clauses translated correctly in the unimodal (*M*=47.25%) than in the bimodal condition (*M*=45.13%), which is not in line with the prediction of the FLMP.

For the second block of presentation, there were significantly more incorrectly translated clauses in the unimodal condition (*M*=10.48%) than in the bimodal condition (*M*=6.08%), (*F*(1, 7)=5.83, *p*<.05). There were slightly more clauses correctly translated when the face was given in addition to the auditory signal (*M*=54.07%) than when just the auditory signal alone was presented (*M*=51.09%). However this difference was not significant (*F*(1, 11)=.42, *p*=.54). There was also no significant difference in percentage of attempted clauses between the two modality conditions (*F*(1, 11)=.03, *p*=.86).

A second analysis was conducted with modality and block (2x2) as within-participants factors. We separately tested these two factors and their interaction with a repeated measure ANOVA for each of the three dependent variables. The analysis yielded only in a significant block effect for percentage of clauses attempted. There were significantly more clauses attempted in the second block (*M*=71%) than in the first one (*M*=64%), (*F*(1, 7)=15.13, *p*<.01). For percentage of correctly translated clauses the difference between blocks was only marginally significant (*F*(1, 7)=4.48, *p*=.07). There were slightly more clauses correctly translated in the second block (*M*=53%) than in the first one (*M*=46%). There was no significant difference between blocks for percentage of

incorrect translated clauses. Modality had no significant influence on any of the three dependent variables. There was also no significant interaction between modality and blocks.

Discussion

The results of the second study replicate our finding in the first experiment that facial information does not seem to improve simultaneous interpretation performance. Our bilingual participants were able to handle this difficult simultaneous interpretation task, but they were on average able to translate only 46% of all clauses correctly at the first time of presentation. This overall performance is lower than the performance of the German participants in the first experiment. This difference could be due to differences in difficulty of interpreting from English to Spanish in comparison to interpreting difficulty from English to German. Of course, other explanations for these differences are also possible.

It is possible that the failure to find an advantage of visible speech in simultaneous interpreting was due to the noise-free auditory speech. Noise degrades perception of the message because in our view there is less information in the speech signal. This loss of information would impact simultaneous interpreting. Gerver (1974), for example, found that adding noise to an auditory signal in a unimodal presentation impacts the performance of professional interpreters. Previous research has found an advantage of visible speech in sentence processing only when noise is added to the auditory speech (Benoit, Mohamadi, & Kandel, 1994; Sumby & Pollack, 1954). In a series of experiments reported in Massaro (1998, Chapter 14), participants were asked to perceive spoken nonsense sentences. The participant was then presented a minimal pair of words and had to make a forced-choice indicating which one was in the presented sentence. The position of the test word in the sentence was randomly varied. The sentences were presented with auditory synthetic speech or auditory synthetic speech plus Baldi, at three different rates of speech. Although performance declined as the rate of the speech increased, adding the talking face did not improve performance at any rate. A fourth experiment, however, presented the sentences at a moderately fast speech rate in +10dB noise. Here, a small but significant advantage of bimodal over unimodal speech was found.

Thus, the absence of noise may be responsible for our failure to find an advantage in simultaneous interpreting a bimodal relative to a unimodal passage. In order to investigate further whether people use visual information in simultaneous interpreting, we plan to present sentences in noise. Our goal is to first replicate the modality effect found in sentence identification (Figure 2) in the first experiment, and then see if the same modality effect for sentences will occur in

simultaneous interpreting. Although isolated sentences are not usually found in simultaneous interpreting, we would have a finer-grained analysis of the test material and responses.

Insert Figure 3 about here.

Another reason that the bimodal speech might not have been beneficial is that Baldi was not programmed to provide any facial expression and head movements. There is now some evidence that head movements can improve intelligibility of spoken language (Risberg & Lubker, 1978; Nicholson, Baum, Cuddy, & Munhall, 2002). Furthermore, Baldi was presented as a disembodied talking head eliminating any potential contribution of gesture. These limitations probably had a significant influence on the effectiveness of the visible speech, as can be seen from the results of the natural face condition in Experiment 1. The same sentences were presented with a radio announcer, Gary, in the auditory and auditory plus Gary's face aligned with his auditory speech. Figure 3 gives the results of these two conditions. As can be seen in the figure, there was a significant advantage of the bimodal condition (M=77%) relative to the unimodal condition (M=50%). More importantly, comparing these results to those in Figure 2, the natural face gave a larger advantage than the synthetic face (27 versus 21%), $F(1,70)=21$, $p<.001$. Thus, it remains distinctly possible that interpreting performance would have benefited more from a natural face than the synthetic face. In addition, we expect that improving Baldi's facial accuracy, adding head movements, and including gestures would make it more likely to obtain an advantage of bimodal speech, especially in contexts where the identification of emotional content, prosody or syntax (e.g. question vs. statement identification) would be beneficial in simultaneous interpreting.

General Discussion

A large body of evidence supports the view that the bimodal presentation improves speech perception (Massaro, 1998). In the present study, we investigated the question whether bimodally presented paragraphs would lead to better perception and therefore also be easier to simultaneously interpret than unimodal presented paragraphs. Our results clearly show that there seems to be no statistically significant difference between the two conditions. This result was found with both native German speakers who had learned English as a second language as well as with Spanish-

American English bilinguals. In terms of our hypotheses, we can conclude that either perception is not improved by the second source of information (facial information) or that enhanced perception does not lead to better interpretation. According to a limited-capacity model, it could be that bimodal perception takes away capacity from the production processes and therefore should then lead to worse performance. However, the present study showed no decrement in the bimodal condition.

There are other reasons to still be encouraged to show that visible speech not simply does not lower performance, but even improves it. Given the vast body of evidence that we cannot help but use visible speech, it seems very unlikely that this would be different for simultaneous interpreting. As can be seen in Figure 2, visible speech improves the recognition of words in sentences. In the simultaneous interpreting experiments, however, the auditory signal was not presented in noise nor ambiguous in any obvious sense. The high quality of the auditory speech might explain why the visual source of information did not provide any additional help in interpreting. Furthermore, our participants may not have mastered speechreading in English to the extent that native English speakers have. Accordingly, comparing the results in perception and comprehension shown in Figure 2 to the results on SI, there is both more auditory information and less visual information in these SI studies. According to the FLMP (Massaro, 1998), a source of information is informative only to the extent that another source of information is ambiguous. It follows that visual speech would make less of a contribution in the present SI experiments.

Notwithstanding the current negative findings, we believe it would be helpful to arrange simultaneous interpreting situations in a way that the interpreter is facing the talker. This could be easily accomplished with the help of a video system.

A somewhat less obvious implication of our theoretical framework is that having both the written message and the spoken message would facilitate simultaneous interpretation. Thus, a simultaneous machine transcription of the spoken input into its written form could contribute positively to simultaneous interpretation. This might be the case even if the transcription into a written form is not perfectly accurate. This prediction is based on the premise that the written form adds to the message. For example, the written form can distinguish among homophones such as *sea* and *see*. Simultaneous transcription could be carried out in real time using any of several continuous speech recognition systems that are commercially available. One potential limitation of most of these systems for this application is that they require training with the talker to implement speaker-dependent recognition. Most individuals requiring translation (such as politicians), however, are

involved in repeated situations in which simultaneous interpreting is required, and it would not be unreasonable to invest in some training time for these individuals.

Thinking laterally, one distant goal of speech science and facial animation technology is to automatically recognize and understand language as it is being spoken translate it into another language, and to produce it by a computer-animated talking head. Our research team and many other researchers have developed several component technologies that bring us closer to this goal. One is texture mapping of a person's face onto the computer-animated talking head and another is the ability to drive our talking head directly from the symbolic input. The source language would be translated and this translation would drive a computer-animated talking head. A technology is also being developed that, given a short sample of a person's speech, synthetic speech can be produced that sounds like that person. Given the appropriate texture mapping and synthetic speech, therefore, the speaker of the source language can be seen producing the target language (Waibel, 1996). The scenario would be that your Japanese colleague in Japan is speaking to you in Japanese. On your computer screen, you see him speaking to you in English. And he sees you speaking to him in Japanese even though you do not know a word of Japanese. Given our high state of globalization in all kind of domains, this application would be a very useful tool for video conferencing, where most often a simultaneous interpreter is not available.

References

Anderson, L. (1994). Simultaneous Interpretation: Contextual and Translation Aspects. In Lambert, S., & Moser-Mercer, B. (Eds.*). Bridging the Gap: Empirical Research in Simultaneous Interpretation (pp.101-120).* Amsterdam and Philadelphia: John Benjamins.

Balzani, M. (1990). Le contact visuel en interprétation simultanée: resultats d'une expérience (Français-Italien). In Gran, L., & Taylor, C. (Eds.*). Aspects of Applied and Experimental Research on Conference Interpretation (pp.93-100).* Udine: Campanotto Editore.

Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research, 37,* 1195-1203.

Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme- to-phoneme stimulus-response alignment. *Journal of the Acoustical Society of America, 95,* 3617-3622.

Bernstein, L. E., & Eberhardt, S. P. (1986). *Johns Hopkins lip-reading corpus videodisk set*. Baltimore, MD: The Johns Hopkins University.

Black, A., & Taylor, P. (1997). *Festival Speech Synthesis System: System documentation (1.1.1),* Human Communication Research Centre Technical Report HCRC/TR-83, Edinburgh.

Cisneros, S. (1991). *The House on Mango Street*. Vintage Books: New York, NY.

Cohen, M.M., & Massaro, D.W. (1993) Modeling coarticulation in synthetic visual speech, In Models and Techniques. In D. Thalmann and N. Magnenat-Thalmann (Eds.), *Computer Animation (pp. 141-155).* Tokyo: Springer.

Cohen, M.M., Walker, R.L., & Massaro, D.W. (1996). Perception of synthetic visual speech. In D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by humans and machines* (*pp. 153-168).* New York: Springer.

Davis, H., & Silverman, S. R. (1978). *Hearing and Deafness* (4th Ed.). New York : Holt, Rinehart and Winston.

Gerver, D. (1974). The effects of noise on the performance of simultaneous interpreters: Accuracy of performance. *Acta Psychologica, 38,* 159-167.

Gile, D. (1997). Conference Interpreting as a Cognitive Management Problem. In Danks, J.H., Shreve, G.M., Fountain, S.B., & McBeath, M.K. (Eds.*). Cognitive Processes in Translation and Interpreting (pp.196-214).* Thousand Oaks: Sage.

Gile, D. (1999). Testing the Effort Models' tightrope hypothesis in simultaneous interpreting – A contribution. *Journal of Linguistics, 23,* 153-172.

Massaro, D.W. (1998*). Perceiving talking faces: From speech perception to a behavioral principle.* MIT Press: Cambridge, MA.

Massaro, D.W., & Shlesinger, M. (1997). Information processing and a computational approach to the study of simultaneous interpretation, *Interpreting, 2*, 13-53.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Meyer, D. E., & Kieras, D. E. (1999) Precis to a Practical Unified Theory of Cognition and Action: Some Lessons from EPIC Computational Models of Human Multiple-Task Performance. *Attention and Performance XVII. Cognitive Regulation of Performance: Interaction of Theory and Application* (pp. 17-88). Cambridge, MA: MIT Press.

Nicholson, K.G., Baum, S., Cuddy, L. & Munhall, K.G. (2002). A Case Of Impaired Auditory and Visual Speech Prosody Perception After Right Hemisphere Damage. *Neurocase, 8*, 314-322.

Risberg, A., & Lubker, J. (1978). Prosody and speechreading. Speech Transmission Laboratory *Quarterly Progress Report Status Report, 4*, 1-16.

Schumacher, E.H., Seymour, T.L., Glass, J.M., Fencsik, D.E., Lauber, E.J., Kieras, D.E., Meyer, D.E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the Central Cognitive Bottleneck. *Psychological Science, 12(2)*, 101-108.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.

Sutton, S., Cole, R. A., deVilliers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, D.W., Cohen, M. (1998). Universal Speech Tools: the CSLU Toolkit. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 3221-3224.

Waibel, A. (1996). Interactive translation of conversational speech. *Computer, 29*, 41-48.

Author Note

Alexandra Jesse, Department of Psychology, Dominic W. Massaro, Department of Psychology, Nick Vrignaud, Department of Psychology.

The research and writing of the paper were supported by grants from National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Technology Award, and the University of California, Santa Cruz.

The authors would like to thank Michael M. Cohen, Karl Young, and Slim Ouni for offering technical assistance and providing computer support.

Correspondence concerning this article should be addressed to either Alexandra Jesse or Dominic Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064. Electronic mail may be sent via internet to ajesse@cats.ucsc.edu or massaro@fuzzy.ucsc.edu .

.

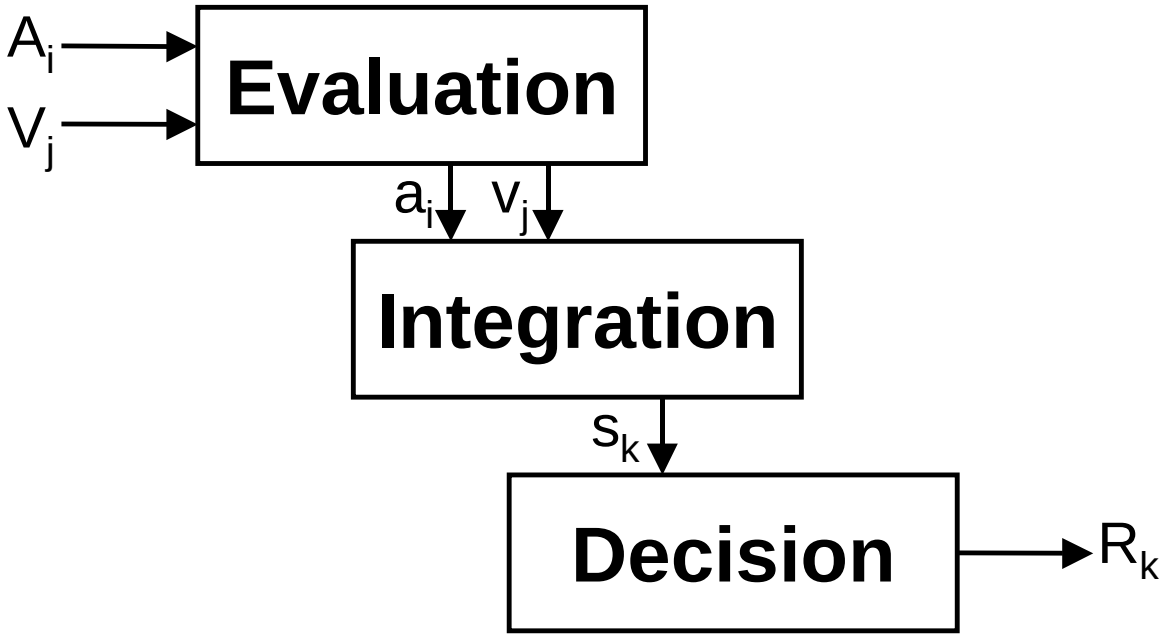

Figure 1. Schematic illustration of three stages of processing in the FLMP. The three processes involved in perceptual recognition include evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms sources of information (A_i & V_j) into psychological values (a_i and v_j), which are then integrated to give an overall degree of support (s_k) for each speech alternative. The decision operation maps the outputs of integration into some response alternative (R_k). The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. In this example, auditory and visual information are integrated to achieve perceptual recognition and, analogously, bottom-up and top-down sources of information are integrated in word and sentence processing.

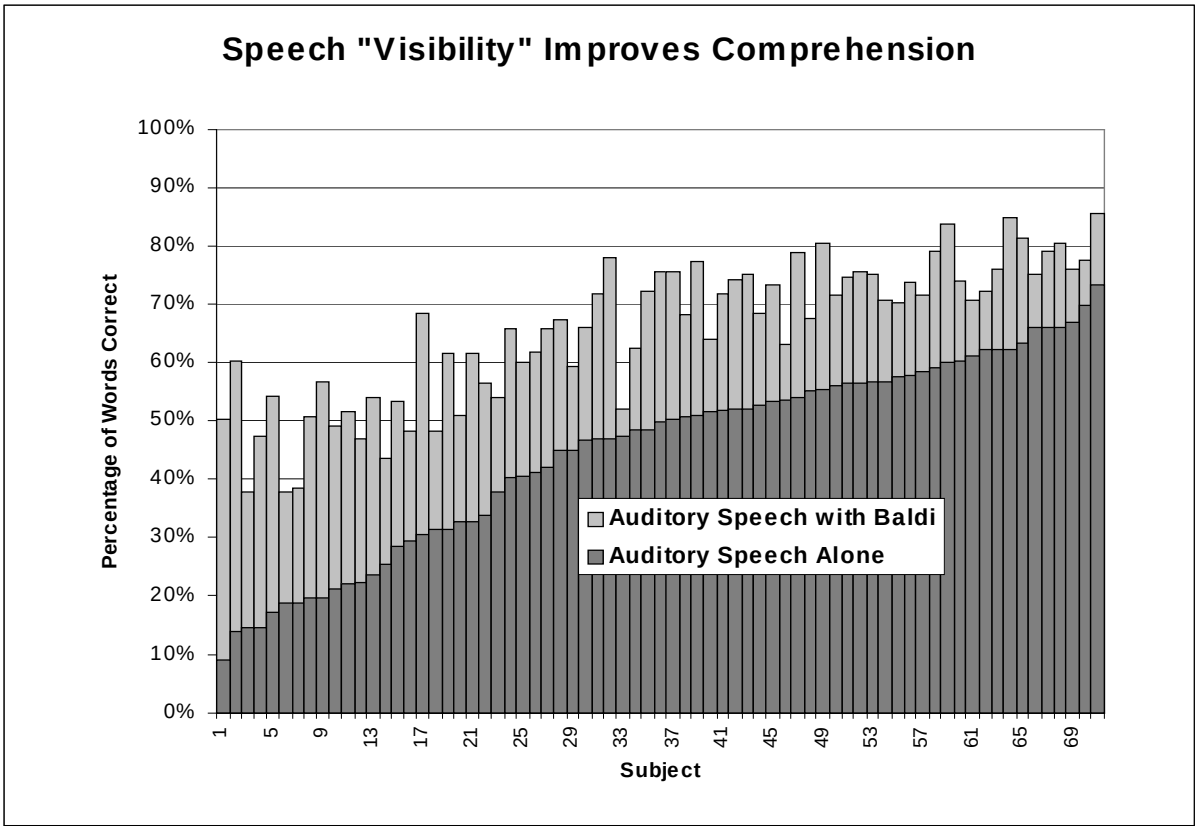

Figure 2. Proportion of words correctly reported for auditory speech alone and auditory speech plus Baldi conditions in experiment 1.

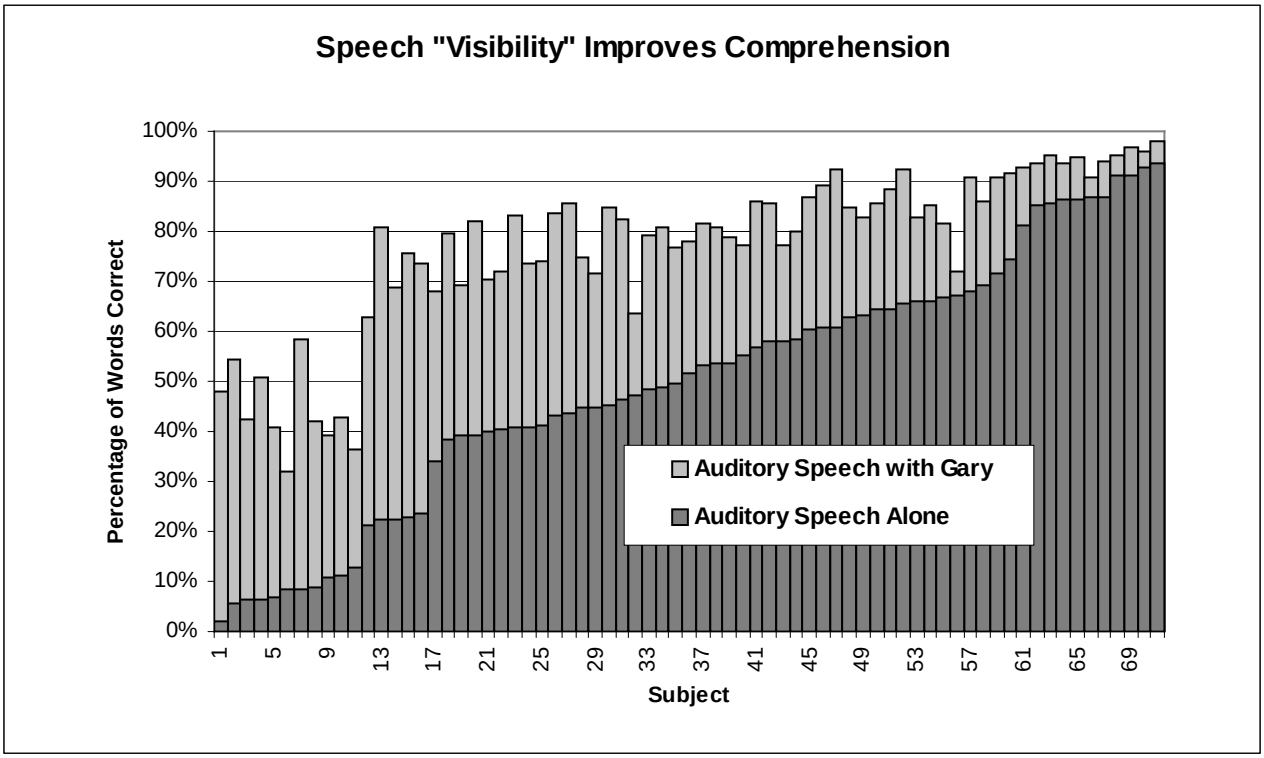

Figure 3. Proportion of words correctly reported for auditory speech alone and auditory speech plus Gary conditions.

Table 1. Mean percentage scores of correctly translated clauses for unimodal (UM) versus bimodal (BM) presentation per participant at time 1 in experiment 2

Participant	UM	BM
1	81	74
2	63	70
3	76	70
4	55	71
5	43	46
6	72	69
7	72	72
8	40	30
9	67	72
10	64	69
11	57	65
12	60	74
mean	62.5	65.2

Table 2. Mean percentage correct translations for the unimodal (UM) versus bimodal (BM) presentations for each of the eight participants at the first and second blocks in experiment 3

	Block 1		Block 2	
Participant	UM	BM	UM	BM
1	65	51	66	82
2	51	40	50	66
3	52	61	63	50
4	43	62	56	41
5	63	53	65	79
6	64	45	56	67
7	17	20	12	13
8	28	29	40	34
mean	47.51	45.21	51.09	54.07