Massaro, D. W. (1998). Illusions and Issues in Bimodal Speech Perception. Proceedings of Auditory Visual Speech Perception '98. (pp. 21-26). Terrigal-Sydney Australia, December, 1998.

ILLUSIONS AND ISSUES IN BIMODAL SPEECH PERCEPTION

Dominic W. Massaro

Perceptual Science Laboratory (http://mambo.ucsc.edu/psl/pslfan.html)

University of California, Santa Cruz, CA 95064

massaro@fuzzy.ucsc.edu

ABSTRACT

As witnessed by this conference and many other sources of evidence, the study of bimodal speech perception has attained the status of a cottage industry. The addition of just one more modality has made transparent several new phenomena, new theoretical endeavors, and a closer link between research and application. The goal of this paper is to review a series of relevant issues in our search for an understanding of speech perception by ear and eye. The issues include a discussion of viable explanations of the McGurk effect, the time course of auditory/visual processing, neural processing, the role of dynamic information, the information in visual speech, the fusion of written language and auditory speech, and the issue of generalizing from studies of syllables to words and larger segments.

1. SETTING THE STAGE

It has been well over two decades since the publication of hearing lips and seeing voices by the late Harry McGurk and his colleague John McDonald [1]. The so-called McGurk effect has obtained widespread attention in many circles of psychological inquiry and cognitive science. The classic McGurk effect involves the situation in which an auditory /ba/ is paired with a visible /ga/ and the perceiver reports hearing /da/. The reverse pairing, an auditory /ga/ and visual /ba/, tends to produce a perceptual judgment of /bga/.

We should not be surprised by the finding that auditory experience is influenced by the visual input. Certainly the McGurk effect was not the first crosstalk between modalities to be observed. We seem to have a little voice, not necessarily our own, in our heads as we read written language. Why do people watching a large screen in a movie theater hear an actor's voice coming from his face, even though the audio speakers are on the side of the screen? (This experience is equally powerful in theaters without stereoscopic sound, where indeed the auditory message has no information tying the sound to the actor.) This so-called visual capture is also exploited by ventriloquists, who contrary to popular belief, do not throw their voice at the puppet. The visual input changes our auditory experience of the location of the sound we are hearing. This situation represents a clear case of cross talk between modalities [2].

We should be relieved that the McGurk effect resembles other avenues of experience, such as localizing sound in space. Its similarity to other domains offers the hope of a more general account of sensory fusion and modality specific experience rather than one unique to speech perception by ear and eye. This result might simply mean that we cannot trust modality-specific experience as a direct index of processing within that modality. Speech information from two modalities provides a situation in which the brain combines both sources of information to create an interpretation that is easily mistaken for an auditory one. We believe we *hear* the speech because perhaps spoken language is usually heard.

We are attracted to bimodal speech perception as a paradigm for psychological inquiry for several reasons. It offers a compelling example of how processing information from one modality (vision) influences our experience in another modality (audition). Second, it provides a unique situation in which multiple modalities appear to be fused or integrated in a natural manner. Third, experimental manipulation of these two sources of information is easily carried out. Finally, the research project has the potential for valuable applications for individuals with hearing loss and for other domains of language learning.

1.1. A Downside to Current Inquiry

Many investigators have been misled by the traditional study of the McGurk effect. First of all it is not reasonable for an investigator to study an effect. For example, it would be foolish for someone to say I study the Ebbinghaus illusion. One investigates illusions to gain some insights into perceptual processing, not simply for the study of illusions. Similarly, it is important to keep in mind that the study of the McGurk effect should be aimed at understanding how we perceive speech. Focusing on the illusion tends to compromise the type of experimental study that is implemented. Most studies of the McGurk effect use just a few experimental conditions in which the auditory and visual sources of information are made to mismatch. Investigators also sometimes fail to test the

unimodal conditions separately so that there is no independent index of the perception of the single modalities.

The data analysis is also usually compromised because investigators analyze the data with respect to whether or not there was a McGurk effect, which often is simply taken to mean whether the visual information dominated the judgments. This approach is highly misleading because it is well-known that one modality does not dominate the other [3,4]. Both modalities contribute to the perceptual judgment with the outcome that the least ambiguous source of information has the most influence. McGurk's original interpretation, that place of articulation was cued by visual information and that manner and voicing were cued by auditory information, is wrong. Many studies have shown repeatedly that auditory information is important for perceiving place of articulation [3]. This is true even when it is paired with relatively unambiguous visible speech.

There is not complete consensus on the explanation of these results. One of the reasons is that highly discriminating data are not available. Investigators tend to 1) take too few observations under each of the stimulus conditions, 2) limit the number of response alternatives, and 3) do not test the unimodal conditions. A better understanding of the McGurk effect is attempted by enhancing the database and testing formal models of the perceptual process.

1.2 Exploring the McGurk Effect

To explore the McGurk effect more fully, we carried out a series of experiments in which the auditory syllables /ba/, /da/, and /ga/ were crossed with these same visible syllables in an expanded factorial design. Subjects were either limited to these three response alternatives or given a larger set of response alternatives. Why does auditory /ba/ paired with a visible /ga/ produce a perceptual report of hearing /da/? Our strategy in explaining this outcome has been to expect it to follow from the psychophysical properties of the audible and visible sources of information. This means that auditory /ba/ must be somewhat more similar to an auditory /da/ than to an auditory /ga/. Another possibility is that there are other sources of information (or constraints) contributing to performance. Higher-order context might be functional in that the segment /d/ appears to be more frequent in initial position than the segment /g/. This a priori bias for /d/ over /g/ (and /t/ over /k/) could be an important influence on the "fusion" response that is observed.

To address these issues, the natural auditory syllables /ba/, /da/, and /ga/ were crossed with the synthetic visual syllables /ba/, /da/, and /ga/ [5]. Participants also identified the unimodal syllables. Ten participants were tested for two sessions of 216 trials each, for a total of roughly 29 observations under each of the 15 conditions.

Figure 1 gives the probability of /ba/, /da/, /ga/,

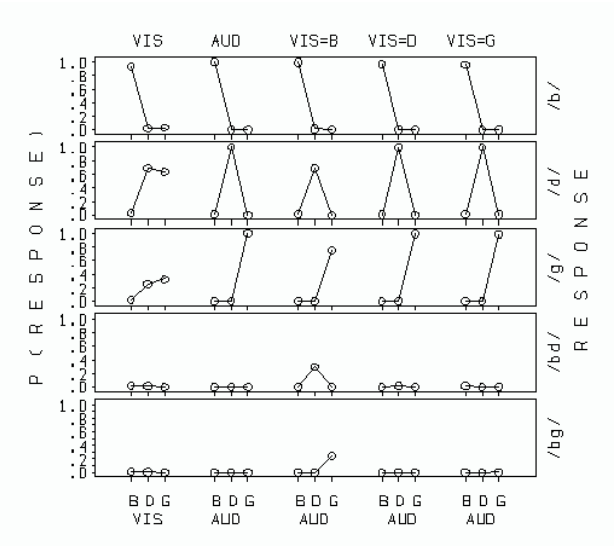

Figure 1: The percentage of /b/, /d/, /g/, /bd/, and /bg/ responses as a function of the three test stimuli in the unimodal visual, unimodal auditory, and bimodal conditions.

/bda/, and /bga/ responses for each of the 15 experimental conditions. Several results are of interest. As expected, there were confusions between visible /da/ and /ga/, because these syllables are very similar and belong to the same viseme category. What is important for our purposes, however, is that the participants respond to both visual /da/ and visual /ga/ about twice as often with the alternative /da/ than with the alternative /ga/. This observation is novel because previous investigators had not tested these unimodal visual conditions. This result offers a new explanation of why an auditory /ba/ paired with a visual /ga/ produces the response /da/. Apparently, people are biased to report /d/ over /g/ because /d/ occurs much more often than /g/ in spoken language [6].

Much to our dismay, however, we failed to replicate the prototypical McGurk fusion effect. Neither a visual /da/ or /ga/ biased the response to auditory /ba/. For whatever reason, the auditory information dominated the perceptual judgment. One possibility is that observers were not permitted to make other responses, such as /va/ or /tha/, which are frequently given to these conflicting syllables. To solidify our interpretation of the prototypical fusion effect, however, we will have to observe the traditional

McGurk effect in the same situation in which a bias for /d/ over /g/ is observed.

Although no fusion responses were observed, there were combination responses. A visual /ba/ paired with an auditory /da/ and /ga/ produced a significant number of /bda/ and /bga/ responses(see Figure 1).

2. TIME COURSE OF FUSION

An important issue in the understanding of bimodal speech perception is how the fusion of audible and visible speech occurs. Fusion models have to be explicit in their assumptions before they can be tested against available data. In addition, our data sets must be made richer before we can convincingly falsify some subset of these models. We formalize models of sensory fusion and test them against discriminating data.

We classify models as nonindependent if the information along one modality modifies the evaluation of the information along a different modality. One apparent piece of evidence against nonindependent models is the result observed when the audible and the visible speech signals are presented asynchronously. Models which assume nonindependence would predict that asynchronous presentations would be highly detrimental to integration. If an auditory consonant occurred 100 ms before the visible speech began, for example, then the static visual information would degrade the evaluation of the auditory speech. This should lead to a different result than if the two modalities were presented synchronously. However, if evaluation and integration occur in the same manner at an SOA of 100 ms as in the synchronous condition, it would provide support for the independent evaluation of auditory and visual speech. In this case, the evaluation of one source is not influenced by the nature of other sources. Integration of the two sources is based on these independent evaluations

To assess the robustness of the integration process across relatively small temporal asynchronies, the relative onset time of the audible and visible sources was systematically varied [7]. In the first experiment, bimodal syllables composed of the auditory and visible syllables /ba/ and /da/ were presented at five different onset asynchronies. The second experiment replicated the same procedure but with the vowels /i/ and /u/. The results indicated that perceivers integrated the two sources at asynchronies of 200 ms or less. We also varied the asynchrony between the audible and visible speech using both natural and synthetic syllables (/ba/, /va/, /tha/, and /da/) in an expanded factorial design [8]. The fuzzy logical model of perception (FLMP) was used to assess whether any type of nonindependence between the auditory and visual information existed. The model gave a good description of the results at asynchronies of about 200 ms or less. We may conclude that the independent evaluation and multiplicative integration of audible and visible speech is very robust across small changes in temporal occurrence. These findings appear to challenge the class of nonindependence models.

3. NEURAL PROCESSING

A third issue in the perception of speech by ear and eye has to do with neurological processing. Recent neuroimaging results imply that speechreading visible speech activates the auditory cortex [9]. How does this result inform us about the representation and evaluation of visible and audible speech in bimodal speech perception? It is not necessarily the case that the visual information invades the auditory pathways and confuses the auditory-specific nerves in terms of what is uniquely auditory and what is not. The influence of the visual information might simply result from a process called nonconvergent temporal integration [10]. One question has to do with the activation of the auditory cortex in the reading of written text. If reading excites auditory cortex, it would appear that there is nothing magical about visible speech. We would simply conclude that language engages the auditory cortex.

4. VISIBLE SPEECH INFORMATION

A fourth issue addresses the nature of the information in visible speech. It has been claimed that the most salient information available for perceiving visible speech is in the form of time-varying dimensions. We tested this hypothesis by creating visible speech with little or no time-varying information [11].

The time-varying information is eliminated from the visible speech signal by a temporal sampling algorithm in which time-varying information is replaced by static information. The nine viseme syllables used in many of our experiments were the test items. The visible speech was produced by our talking head, Baldi. The critical variable was the number of unique frames presented per second. Thirty frames/s corresponds to a normal rate. Five frames/s would mean that there were only five unique frames during the syllable presentation (which lasted about 1 s). To the extent there are only a few unique frames, then necessarily there is very little dynamic information in the visible speech signal.

Figure 2 gives the probability of correct identification as a function of the number of unique frames during the stimulus presentation. The curve parameter is the experimental session. The accuracy of perceptual judgments to these stimuli informs us about the information value of time-varying speech.

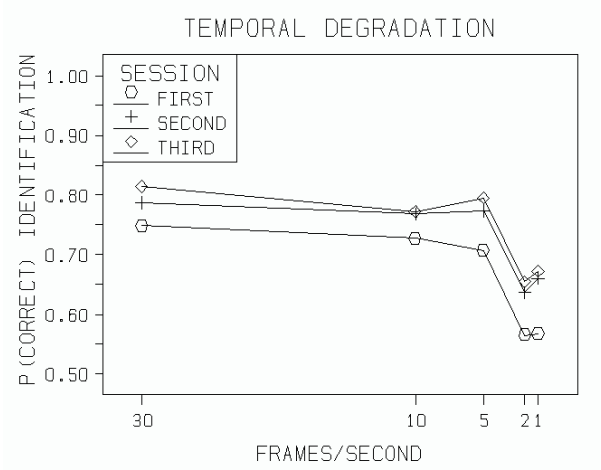

Figure 2. The probability of correct identification of the nine visemes as a function of the number of frames presented per second. The curve parameter is the three test sessions.

As can seen in the figure, perceptual recognition remains asymptotic until only two frames per second. This means that only two frames were used to specify the syllable. Analysis of the two frames for each of the syllables indicated that they no longer provided sufficient information to identify the syllables /la/ and /ʃa/. These were the only two syllables that showed a decrement at the two most difficult conditions. Finally, Figure 2 reveals that although performance improved somewhat with experience in the task, the influence of temporal degradation remained the same. We conclude that dynamic information does not appear to be essential in recognizing visible speech (at least at the syllable level; see Section 6).

5. WRITTEN TEXT AND SPEECH

A fifth issue concerns whether sensory fusion of auditory and visual inputs is limited to speech stimuli. We carried out a series of experiments that compared the perception of auditory speech paired with visible speech versus auditory speech paired with written language. The results from this study can help inform us about which theories of bimodal speech perception are viable.

Returning to the story of illusions (see Section 1), most seem to occur within a modality. A visual context creates an illusory visual impression: it does not change a visual experience into an auditory one. There might also be amodal influences on perceptual experience, however. Knowing or seeing the words to a rock song while hearing the song creates the impression of hearing a highly intelligible rendition of the words. Without this knowledge of the words, the listener cannot make heads or tails of the message. The first demonstration of this kind that I know of was by John Morton, who played a song by the Beatles. Members of the audience could not perceive clearly the words of the song until they were written on the viewing screen. Another variation on this type of illusion is the so-called phonemic restoration effect in which we claim to hear the /s/ in the word legislatures even though it is replaced by a cough, a buzz or even a pure tone [12].

Frost, Repp, and Katz [13] found that when a spoken word is masked by noise having the same amplitude envelope, subjects report that they hear the word much more clearly when they see the word in print at the same time. This result supports the idea that written text can influence our auditory experience. To show effects of written information on auditory judgment at the perceptual level, Massaro, Cohen, and Thompson [14] compared the contribution of lip-read information to written information. Subjects were instructed to watch a monitor and listen to speech sounds. The sounds were randomly selected from nine synthetic speech sounds along a /ba/ to /da/ continuum. On each trial, the subjects were presented with either 1) a visual representation of a man articulating the sound /ba/ or /da/, or 2) a written segment BA or DA. Although there was a large effect of visible speech, there was only a small (but significant) effect of the written segments on the judgments. Both the speech and written-text conditions were better described by the FLMP than by an alternative additive model.

To better test for the possible influence of text on speech perception, we aimed to obtain a larger effect of written text [15]. Given that letters of the alphabet have a strict spelling-to-sound mapping and are pronounced automatically and effortlessly, the letters B and D were used. The letter sequences BA and DA are not necessarily pronounced /ba/ and /da/. The letters B and D are only pronounced /bi/ and /di/--as they are named in the alphabet.

Nine participants from the University of California, Santa Cruz were tested. This experiment employed a within-subjects expanded factorial design. There were seven auditory levels between the syllables /bi/ and /di/ [15]. There were four visual levels--two letter conditions (the letters B and D) and two speech conditions (the visual syllables /bi/ and /di/), for a total of 39 trial types. The observers were specifically instructed to both watch the screen and listen for a sound and to report what they heard. On those trials in which only a visual stimulus was presented, they were to report the visual stimulus. On each trial, subjects identified stimuli as B or D by typing the appropriately marked keys. The stimuli were presented in 6 blocks of the 39 trial types for a total of 234 trials per session. The test conditions were selected at random without replacement. A practice block of 10 trials occurred prior to the experimental trials. Subjects had approximately three seconds to respond on each trial. Each subject participated on two days with two

sessions per day. Thus there were 24 observations per subject per condition. The dependent measure was the proportion of /di/ responses for each of the 39 experimental conditions.

Figure 3 displays the average results for the letter and speech conditions. The proportion of /di/ responses as a function of the seven auditory levels is shown with the visual B or D stimulus or no visual information (NONE) as the curve parameter. The average proportion of /di/ responses increased significantly as the auditory syllable went from the most /bi/-like to the most /di/-like level. There was also a significant effect on the proportion of /di/ responses as a function of the visual stimulus, with fewer /di/ responses for visual B than for a visual D. The interaction of these two variables was

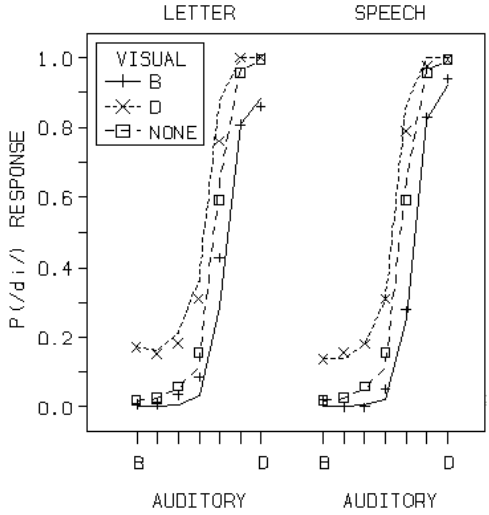

Figure 3: Observed (points) and predicted (lines) by the FLMP probability of a /di/ response as a function of the auditory and visual stimuli for the letter and word conditions.

also significant given that the effect of the visual variable was smaller at the less ambiguous regions of the auditory continuum.

The result of interest here is the difference between the visible speech and the letter conditions. As can be seen in the figure, the visual effect was substantial and of similar size for the letter and for the speech condition. The FLMP was fit to the average proportion of /di/ responses for each of the 9 participants. The FLMP gave a very good description of the observations. Thus, we conclude that written text, as well as visible speech, can influence our auditory experience and that the FLMP accounts for both types of influence. Other experiments with a larger number of response alternatives and without visual-alone trials reinforce this conclusion [15].

6. SYLLABLES TO SENTENCES

A sixth issue addresses the concern that research with syllables might not generalize to words and sentences. Experimental results with syllables should be compared with those with words and sentences to determine if the same model can be applied to these different test items. To move beyond syllables, we have begun to assess the processing of auditory and visual speech at the word level. Settling on an experimental task for evaluation is always a difficult matter. Even with adequate justification, however, it is important to see how robust the conclusions are across different tasks. The task we chose was the gating task, in which successively longer portions of a test word are presented [16].

Following our theoretical framework, we tested observers under auditory, visual, and bimodal conditions [17]. The test words were monosyllabic CVCs. Eight gating durations were tested. As expected, performance improved with increases in the duration of the test word. Auditory information was more informative than visual, but bimodal performance was significantly better than either unimodal condition. The FLMP was fit to both the accuracy of identification of the test words, as well as to the identification of the individual segments of the word. In both types of analyses, the FLMP outperformed a competing additive model, and gave a good description of the observed results. Thus, we have some evidence that the theoretical framework of the FLMP will generalize beyond the syllable level to words and even beyond.

7. OUR THEORETICAL FRAMEWORK

Our work has combined sophisticated experimental designs and quantitative model testing to understand speech perception and pattern recognition more generally. A wide variety of results have been described within the FLMP. The three processes involved in perceptual recognition are evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration into some response alternative. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The assumptions central to the model are: 1) each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall continuous degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives. In the

course of our research, we have found the FLMP to be a universal principle of perceptual and cognitive performance, which accurately models human pattern recognition. People are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation. Several of the sources together, however, usually provide a clear answer.

In spoken language perception, multiple sources of information are available to support the identification and interpretation of language. The experimental paradigm that we have developed allows us to determine which of the many potentially functional cues are actually used by human observers [4, Chapter 1]. This research strategy addresses how different sources of information are evaluated and integrated, and can also identify the sources of information that are actually used. These results show how visible speech is processed and integrated with other sources of information. The systematic variation of the properties of the speech signal and quantitative tests of models of speech perception allow the investigator to interpret the psychological validity of different ecological variables. This paradigm has proven to be effective in the study of audible, visible, and bimodal speech perception [3,4].

8. ACKNOWLEDGEMENTS

This paper is dedicated to the memory of Christian Benoit. The research is supported by grants from PHS, NSF, Intel Corporation, and UCSC. Michael Cohen, Jonas Beskow, Christopher Campbell, and David Merrill provided valuable help on the paper.

9 REFERENCES

1. McGurk, H., & MacDonald, J. “Hearing lips and seeing voices,” *Nature, 264*, 746-748,1976.

2 Bertelson, P., & Radeau, M. “Cross-modal bias and perceptual fusion with auditory-visual spatial discordance,” *Perception & Psychophysics, 29*, 578-584, 1981.

3. Massaro, D.W., *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, Cambridge, MA, 1998.

4. Massaro, D.W., *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

5. Bosseler, A. N., Cohen, M. M., & Massaro, D. W. Influences on the McGurk effect. Unpublished paper, *Perceptual Science Laboratory*, 1998.

6. Denes, P. B. “On the statistics of spoken English.” *Journal of the Acoustical Society of American, 35*, 892-904, 1963.

7. Massaro, D. W., & Cohen, M. M. “Perceiving aysnchronous bimodal speech in consonant-vowel and vowel syllables.” *Speech Communication, 13*, 127-134, 1993.

8. Smeele, P. M. T., Massaro, D. W., Cohen, M. M., & Sittig, A. C. “Laterality in Visual Speech Perception,” *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1232-1242 1998.

9. Calvert, G. A., et al. “Activation of auditory cortex during silent lipreading,” *Science, 276,*, 593-596, 1997.

10. Zeki, S. *AVision of the Brain*. Oxford: Blackwell Scientific Publications, 1993.

11. Campbell, C. S., & Massaro, D. W. Temporal degradation of visible speech. Unpublished paper, *Perceptual Science Laboratory*, 1998.

12. Warren, R. M. “Perceptual restoration of missing speech sounds,” *Science, 167*, 392-363, 1970.

13. Frost, R., Repp, B.H., & Katz, L. “Can speech perception be influenced by simultaneous presentation of print*?” Journal of Memory and Language, 27*, 741-755, 1988.

14. Massaro, D.W., Cohen, M.M., & Thompson, L.A. “Visible language in speech perception: Lipreading and reading,” *Visible Language, 1*, 8-31, 1988.

15. Barton, C. J., Cohen, M. M., & Massaro, D. W. The influence of written text on auditory experience. Unpublished paper, *Perceptual Science Laboratory*, 1998.

16. Grosjean, F. “Spoken word recognition processes and the gating paradigm,” *Perception & Psychophysics, 28*, 267-283, 1980.

17. De la Vaux, S. K., & Massaro, D. W. “The evaluation and integration of information in audiovisual gating of speech.” Unpublished paper, *Perceptual Science Laboratory*, 1998.