STUDIES OF THE MCGURK EFFECT: IMPLICATIONS FOR THEORIES OF SPEECH PERCEPTION

Kerry P. Green

University of Arizona

ABSTRACT

Studies of the McGurk effect demonstrate that observers integrate auditory information with visual information from a talker's face during speech perception. The findings from these studies pose challenges for theories of speech perception that must account for how and why the auditory and visual information are integrated. One theoretical issue concerns the objects of speech perception. Some researchers claim that the objects of speech perception are articulatory gestures while others argue that the objects are auditory in nature. The McGurk effect is often taken as evidence for gestural approaches because such theories provide a good account for why the auditory and visual information are integrated during perception. The findings from various studies of the McGurk effect including cross-modal context effects, developmental influences, and neuromagnetic measures of brain activation will be reviewed. The implication of these findings will be discussed with regard to whether the metric for combining the auditory and visual information is best thought of as auditory or gestural in nature.

1. INTRODUCTION

There is now abundant evidence that speech processing is a multimodal rather than a unimodal process even for normal hearing listeners presented with clear speech. This was first demonstrated by McGurk and MacDonald [1] who found that auditory syllables such as /ba/ dubbed onto a videotape of talkers articulating different syllables such as /ga/, were perceived as something different from either the auditory or the visual signals: typically "tha or "da". The findings of McGurk and MacDonald raised two important questions with regard to auditory-visual speech processing: (1) when are the two signals combined or integrated during speech processing, and (2) what metric is used to combine the two signals? Since the original findings of McGurk and MacDonald, several studies have attempted to address one or the other of these questions. These attempts have been complicated by the fact that the "McGurk effect" is complex. Dubbing an auditory syllable such as /ga/ onto a face articulating /ba/ does not produce "tha" or "da". Instead, subjects (Ss) report that the talker was saying something like "bga". The findings from the McGurk effect pose a challenge to theories of speech perception which must not only address the two questions listed above, but also consider why visual information is combined with the auditory information when the auditory signal by itself provides sufficient information for accurate speech perception under most conditions.

One issue addressed by theories of speech perception concerns the "objects" of perception and whether they are articulatory or auditory in nature [2,3]. With regard to this issue, the McGurk effect and other findings of auditory-visual (AV) speech perception have played an important role by demonstrating that the perception of speech is not solely an auditory process, even under normal listening conditions. The McGurk effect is often seen as evidence for gestural theories because such theories provide a good account for why the auditory and visual information are integrated during perception. They are integrated because both signals provide the observer with information about articulatory gestures. Articulatory gestures also become the common denominator or metric with which to integrate the information from the two modalities. The account is different for auditory theories. Visual information is thought to influence the perception of speech because associations between visual features and phonological representations have been acquired through the experience of watching talkers mouths move while listening to them speak. The exact metric used to combine the information is not always described. It may be auditory in nature or it may be something more abstract such as fuzzy truth values representing the independent strength of the available information in each modality for a particular segment [4].

In this paper, evidence from three different areas of research are described that relate to the issue of the metric used to integrate the auditory and visual information. The three areas are: (1) a recent study on cross-modal context effects; (2) developmental studies of the McGurk effect in children and infants; and (3) studies of neuromagnetic imaging of brain activations during the presentation of McGurk type speech tokens. The data are discussed with regard to whether AV speech perception is best accounted for by auditory or gesture-based theories of perception.

2. CROSS-MODAL CONTEXT EFFECTS

Context effects are situations in which the phonetic perception of the auditory signal is modified by the nature of the surrounding phonetic context. Usually, the change in perception is in accord with the coarticulatory effects of the surrounding context on a target phoneme during production. The congruence between production and perception has led some researchers to argue that context effects reflect the use of tacit knowledge of coarticulation during phonetic perception [6]. Others however, have argued that context effects reflect auditory principles that serve to enhance the perceptual distinctiveness between different speech sounds [3,7].

Recently, we have investigated whether context effects occur when the context is presented in the visual modality and the relevant segmental information is presented in the auditory modality. Our most recent study examined the impact of a bilabial stop consonant on the production and perception of /l/ and /r/ in stop clusters as in /bri/ and /bli/ [8]. The production data are shown in Table 1. The bilabial context produced a significant reduction in the onset frequency of the second formant (F2) for /l/, and a reduction in the onset frequency of the third formant (F3) for /r/ that was not quite significant. There

was also a reliable increase in the slope of F2 for /l/. The changes in production raised the question of whether the perceptual system was sensitive to such changes during speech perception. This question was addressed by synthesizing an auditory /iri-ili/ continuum and a single /ibi/ token. The continuum was created by increasing the third formant (F3) onset frequency. Three different types of stimuli were constructed from these tokens. The first type consisted of a diotic presentation of the /iri-ili/ tokens over headphones. For the second type, a bilabial release burst was added into the waveform, preceding the onset of each /r-l/ token. These tokens were also presented diotically and perceived as ranging from /ibri-ibli/. For the third type, each member of the /iri-ili/ continuum was paired with the auditory /ibi/. The /iri-ili/ token was presented to one ear and the /ibi/ to the other ear in a dichotic presentation. These tokens were also perceived as ranging from /ibri-ibli/. The tokens were blocked by type and presented to Ss who identified whether the syllable contained an /r/ or an /l/.

	F2 Onset Frequent	F3 Onset Frequency	F2 Slope	F3 Slope
/iri/	1401	2167	13.8	12.25
/ibri/	1405	2058	10.8	11.82
Difference	-4.0	109	3.8	-.43
/ili/	1601	2794	13.6	2.2
/ibli/	1333	2731	17.1	6.2
Difference	268**	63	-3.5**	-4.0

Table 1: Mean formant frequency values (Hz) associated with the initial onset of /r/ and /l/ in different contexts, as well as the formant transition rates (Hz/ms). A ** indicates a significant difference, $p < .05$.

The /r-l/ boundaries for the three types of tokens are presented in Table 2. Raising the onset frequency of F3 changed the identification of the tokens from /r/ to /l/ for all the tokens. More important, the dichotic /ibri-ibli/ tokens produced a reliable shift in the boundaries towards a lower F3 onset frequencies relative to the diotic /iri-ili/ tokens. This shift was consistent with the production data which showed that /r/ is produced with a lower F3 onset frequency when it is preceded by a bilabial stop consonant. Finally, the /iri-ili/ tokens with the stop release burst did not produce a reliable shift in the /r-l/ boundary, even though they were perceived as varying from /ibri-ibli/. A follow-up experiment indicated that listeners discerned no difference in the overall "goodness" of the /b/ in the two types of /ibri-ibli/ tokens. Apparently, just having the perception of a stop consonant in the token is not enough to cause a shift in the /r-l/ boundary.

This experiment demonstrates that in the perception of /r/ and /l/, the perceptual system compensates for a preceding bilabial context when it is specified in the auditory signal. The purpose of the second experiment was to determine if the perceptual system would also compensate for the bilabial context when it was specified only in the visual signal. Illusory stop cluster pairs were created by pairing a visual /ibi/ with each of the /iri-ili/ tokens used in the first experiment. In medial position, the illusory presence of the bilabial stop is very strong, especially when paired with auditory tokens that form a natural stop cluster in English. If the visual information is simply serving to produce a perception of the bilabial stop, as occurred in the /iri-ili/ tokens with the stop burst, then the visual information should produce no shift in the /r-l/ boundary. However, if the visual signal provides information about coarticulation that is taken into account during perception, then a shift in the /r-l/ boundary should occur. A new group of Ss were presented with the AV /ibri-ibli/ tokens in one condition, and just the /iri-ili/ tokens in a separate AO condition. As in Experiment 1, the Ss identified whether the tokens contained an /r/ or an /l/.

Condition	Token	Perceived	Boundary
Experiment 1			
diotic	/iri-ili/	iri-ili	2360 Hz
diotic	/iri-ili/+burst	ibri-ibli	2282 Hz
dichotic	/iri-ili/+/ibi/	ibri-ibli	2192 Hz
Experiment 2			
AO	/iri-ili/	iri-ili	2250 Hz
AV	/iri-ilil/ + ibi	ibri-ibli	2094 Hz

Table 2: Mean /r-l/ boundaries in F3 onset frequency for the different conditions in Experiment 1 and 2.

The /r-l/ boundaries for these two conditions are also presented in Table 2. Analysis of the mean boundaries indicated there was a significant shift ($p < .01$) between the AO and the AV conditions. Moreover, the magnitude and direction of the shift is comparable to that which occurred between the dichotic /ibri-ibli/ and diotic /iri-ili/ tokens in Experiment 1. A follow-up, visual-only experiment ruled out the possibility that the shift in the AV boundary was the result of the visual bilabial looking like an /l/ articulation and producing some kind of visual response bias.

One question that arises is: what type of articulatory information might be provided by the visual signal that could influence the perception of the /r-l/ tokens? It can't simply be the case that the perception of /b/ caused the shift in the /r-l/ boundary. As shown in Experiment 1, there were situations in which a /b/ was perceived but the shift in the boundary did not occur. An alternative possibility is that the visual signal provided information that was also consistent with the coarticulatory influences of the bilabial stop on the acoustic realization of /r/ and /l/. For example, the visual signal might have provided information about the rate of change in the opening of the oral cavity. The more rapid opening indicated by the visual bilabial could have been taken as coarticulatory evidence for the presence of an /l/ token. As shown in our production data, there was a significant increase in the slope of F2 for /l/ in the stop cluster environment and F2 is affected by changes in the size and shape of the oral cavity. It may be that knowledge of the coarticulatory influence of the bilabial on the F3 onset frequency for /r/ and evidence of a more rapidly opening oral cavity are both necessary to produce a reliable shift in the /r-l/ boundary.

The results of Experiment 2 are problematic for an auditory account of cross-modal speech perception for several reasons. First, there was no possibility of a direct auditory interaction between the information for the bilabial and the information for the /r-l/ that could account for the change in the /r-l/ boundary. Second, it is not obvious how the visual information could establish a context that would influence the perception of the onset frequency of F3 in the auditory signal. Third, the results from Experiment 1 rule out an interaction between the two modalities at the phonological level. If the context effects were

due to interactions at this level, a shift in the /r-l/ boundary would occur any time there was enough information to activate a /b/ percept, regardless of the modality. As shown in Experiment 1, this was not the case. There are however, two possible accounts that are consistent with an auditory theory. First, visual information about rate of change in the oral cavity could be converted to an auditory metric and then combined with the information from the auditory signal. Alternatively, the perceptual system may have learned an association between a change in the visual oral cavity over time and F2. However, this association is at an earlier level of analysis than is typically assumed by auditory theories to account for auditory-visual interactions. The challenge is to account for how such associations might be built up since we have no isolated awareness of F2 independent of the phonetic percept.

For gestural theories, the account is more straightforward. Both the auditory and visual signals provide information about the gestures involved in articulating stop clusters. The visual signal provides information about the presence of a bilabial. It also provides information about the rate of change in the opening of the oral cavity which interacts with that derived from the slope of F2 provided by the auditory information. This would make it appear more rapid than when it is specified by just the auditory signal. Both pieces of information jointly serve to influence the /r-l/ decision.

3. DEVELOPMENT

The auditory account of AV speech effects depends upon the occurrence of perceptual learning to build up associations between the visual information and phonological representations. If such learning occurs, then developmental differences in the degree to which visual information influences speech perception ought to arise. For example, young infants might get little or no McGurk effect because they are still in the process of forming their phonological prototypes and they have had little opportunity to correlate visual gestures with auditory speech sounds. However, as kids get older, the visual information ought to have a stronger impact.

There is evidence that age influences the magnitude of the McGurk effect in children. Young children typically have smaller McGurk effects than older children or adults [1, 9, 10]. These data seem to support the notion that experience may be improving the associations between the visual and phonological representations. However, there are several reasons for questioning this conclusion. First, recent studies have demonstrated that 4-5 month old infants do get McGurk type percepts [11, 12]. Thus, the capability to integrate the auditory and visual speech information occurs at a very young age. Second, there are several factors that might influence whether young children get strong or weak McGurk effects. Our research has shown that using a different talker with the same face can have a significant impact [9]. With one talker, there was a 55% difference in the McGurk effect between young and old children. However, this difference declined to only 19% for a second talker. This was due to a large increase in the McGurk effect in the younger children for the second voice (nearly 51%) and only a moderate increase for the older children (16%). Age and experience may have less to do with the amount of influence that the visual signal has on speech perception then the characteristics of the auditory signal. Experience might alter the way kids attend to various dimensions of the auditory signal. Young children may weight the auditory dimensions differently than older children [13], and this alternate weighting might result in reduced interaction with the visual information.

Finally, there is little evidence that experience improves the way the visual information is associated with phonological representations. One place where experience might be expected to play a major role in the mapping between visual information and phonological representations is in the ability to speechread. However, existing studies reveal a considerable amount of variability in speechreading capability for people with normal hearing or with severe hearing-impairments, and the variation usually shows little correlation with experience. Even specific training on speechreading usually produces little benefit in mapping the visual gestures onto phonological representations. The benefits that do occur are often specific to a particular talker or phonetic environment, or involve the improvement of linguistic or general communication strategies.

Overall, there appears to be little evidence that children are learning to associate visual articulations with phonological representations. Young children and even infants are capable of integrating the auditory and visual information. The age effects that do occur may be the result of other factors such as how children weight the underlying auditory dimensions of the speech signal. This would pose an obstacle for auditory theories that depend upon perceptual learning to account for why the visual information influences the perception of speech. The data are more consistent with gestural theories, if one assumes that children change the perceptual weighting of various gestural dimensions as they get older.

4. BRAIN ACTIVATION DURING AV PRESENTATIONS

Lately, researchers have been investigating speech perception from a neurophysiologic perspective [14]. The mismatch negativity (MMN) response has been used to examine auditory and speech discrimination in children and adults and is thought to reflect processing of acoustic differences in auditory stimuli [14]. It therefore represents an electrophysiological measure of discriminability at the level of the auditory cortex. The MMN is obtained by presenting Ss with numerous examples of two acoustically different tokens, one occurring less frequently than the other. Recording and averaging of the brain's electrical activity is made for each of the two stimuli and the difference between the two waveforms is a measure of the MMN response.

In a recent study, Sams and his colleagues [15] obtained MMN responses to AV tokens consisting of the same auditory token (/pa/) paired with different visual articulations (/pa/ or /ka/). Neuromagnetic recordings indicated that the "McGurk" token produced an MMN response even though the auditory portion was identical for both AV tokens. Moreover, the same visual articulations presented without the auditory stimulus did not elicit an MMN response, even though the articulations were clearly distinguishable.

There are two reasons why Sams' [15] data are of interest. First, the MMN response is very sensitive to differences in the auditory dimensions of different stimuli but relatively insensitive to visual or tactile qualities. Thus, two tones of different frequency or loudness will produce an MMN but two

lights of different colors will not. Moreover, the MMN occurs even for two different speech sounds that are both members of the same category, indicating that it is a response to auditory rather than category differences. What is interesting is that an MMN is produced for the same speech sound paired with different visual articulations, which by themselves do not produce an MMN. The second reason why the data are of interest is the localization of the MMN response to the AV tokens. Sams used a technique that was quite accurate for localizing the source of cortical activation of the MMN. For these stimuli, the MMN was found in the left hemisphere and depending upon the S, also in the right hemisphere (although usually smaller). Moreover, the response occurred in the temporal lobe just posterior to primary auditory cortex. This localization is consistent with the perceptual impact of the McGurk effect: that of "hearing" a different speech sound from the one actually presented in the auditory signal.

How might these data be accounted for in a gestural or auditory theory? These data might reflect the activation of a common mode of representation in auditory cortex that is gestural in nature. An MMN occurs for the AV tokens because the combined gestural information for the frequent and rare stimuli is different. There are two problems with this account. First, the MMN occurs in auditory cortex and is clearly sensitive to auditory differences between different sounds. It is not clear why it would also be sensitive to differences in the gestural qualities of speech sounds. Second, and more problematic, is the fact that no MMN is produced for the same visual articulations presented without the sound even when the two gestures are quite distinguishable (and at least one, /p/, readily identifiable). If the MMN reflects the difference in the processing of articulatory gestures at the level of auditory cortex, then it shouldn't matter which modality provides the information about the gestures.

An alternative interpretation is that the MMN occurs because the visual information is being mapped onto auditory rather than gestural dimensions. This interpretation is consistent with auditory theories but certain issues need to be addressed. For example, studies of the MMN indicate that it reflects precategorical differences among the speech stimuli. The fact that it occurs for AV tokens suggests that the visual information is mapped onto the auditory dimensions prior to phonetic categorization. This is problematic for auditory theories that assume visual gestures are associated with existing phonological representations by experience. Alternatively, it may be the case that MMNs are produced at several different levels in the auditory system as a result of perceptual differences as well as auditory differences. Simply getting an MMN for McGurk tokens wouldn't specify which level of processing was responsible. Additional studies using AV and AO stimuli will be necessary to tease apart these possibilities. With the advances being made in imaging techniques and the capability to store auditory and video signals on-line, such studies should be forthcoming in the near future.

5. CONCLUSIONS

Theories of speech perception must be able to account for the McGurk effect and the conditions under which it occurs. This paper has described three different kinds of data that need to be addressed by auditory and gestural theories of speech perception. Neither type of theory does a completely satisfactory job with all three kinds of data. However, by examining such data with regard to these theories and others, a more complete understanding will emerge of how and why auditory and visual information are integrated during spoken language processing.

6. REFERENCES

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

2. Liberman, A.M. & Mattingly, I.G. The motor theory of speech perception revised. *Cognition*, 21, 1-36, 1985.

3. Diehl, R.L. & Kluender, K.R. On the objects of speech perception. *Ecological Psychology*, 1, 121-144, 1989.

4. Massaro, D. *Speech perception by ear and eye: A paradigm for Psychological Inquiry*. London: Erlbaum, 1987.

6. Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92. 81-110, 1982.

7. Diehl, R.L., Kluender, K.R., & Walsh, M. Some auditory bases of speech perception and production. *Advances in Speech, Hearing and Language Processing*, 1, 243-267, 1990.

8. Norrix, L. W. & Green, K.P. Auditory-visual context effects on the perception of /r/ and /l/ in a stop cluster. J*ournal of the Acoustical Society of America*, 99, 2591, 1996.

9. Boliek, C., Green, K.P., Fohr, K. & Obrzut, J. Auditory-visual perception of speech in children with learning disabilities: The McGurk effect. *Annual meeting of the International Neuropsychological Society*, Chicago, 1996.

10. Hockley, S.N., & Polka, L. A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America*, 96, 3309, 1994.

11. Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. The McGurk effect in infants. *Perception & Psychophysics*, in press.

12. Desjardins, R. & Werker, J. 4-month-old infants are influenced by visible speech. *International Conference on Infant Studies*. Providence, 1996.

13. Nittrouer, S. & Studdert-Kennedy, M. The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research*, 30, 319-329, 1987.

14. Kraus, N., McGee, T., Carrell, T.D., & Sharma, A. Neurophysiologic bases of speech discrimination. *Ear and Hearing*, 16, 19-37, 1995.

15. Sams, M., Aulanko, R., Hamalainen, M., Hari., R., Lounasmaa, O.V., Lu, S-T., & Simola, J. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, 127, 141-145, 1991.