Cognition 151 (2016) 96–107

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access

Rachel Ostrand [a,*], Sheila E. Blumstein [b], Victor S. Ferreira [c], James L. Morgan [b]

[a] Department of Cognitive Science, University of California, San Diego, United States
[b] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, United States
[c] Department of Psychology, University of California, San Diego, United States

ARTICLE INFO

Article history:
Received 3 January 2015
Revised 17 February 2016
Accepted 27 February 2016
Available online 21 March 2016

Keywords:
Auditory–visual integration
Lexical access
McGurk effect
Multisensory perception

ABSTRACT

Human speech perception often includes both an auditory and visual component. A conflict in these signals can result in the McGurk illusion, in which the listener perceives a fusion of the two streams, implying that information from both has been integrated. We report two experiments investigating whether auditory–visual integration of speech occurs before or after lexical access, and whether the visual signal influences lexical access at all. Subjects were presented with McGurk or Congruent primes and performed a lexical decision task on related or unrelated targets. Although subjects perceived the McGurk illusion, McGurk and Congruent primes with matching real-word auditory signals equivalently primed targets that were semantically related to the auditory signal, but not targets related to the McGurk percept. We conclude that the time course of auditory–visual integration is dependent on the lexicality of the auditory and visual input signals, and that listeners can lexically access one word and yet consciously perceive another.

1. Introduction

Speech comprehension is a complex, multi-staged process. Although speech perception is primarily driven by the auditory signal (Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008; Erber, 1975), visual information, such as that provided by mouth movements, can have an influence as well (Fort, Spinelli, Savariaux, & Kandel, 2010; Green, 1998; Summerfield, 1987), especially in noisy or degraded environments (Erber, 1975; Grant & Seitz, 2000; Sumby & Pollack, 1954). This implies that auditory and visual signals are integrated into a single representation at some point during processing. The present work addresses whether auditory–visual (AV) integration occurs before or after the component stimuli access the lexical-semantic network, and thus what role (if any) visual speech information plays in lexical access.

McGurk and MacDonald (1976) first reported the McGurk Effect, in which mismatching auditory and visual signals perceptually combine. The result is that listeners consciously perceive a stimulus which is a fusion of the auditory and visual inputs, and thus is different from what would be perceived by hearing the auditory signal alone. To create these integrated auditory–visual percepts, a video of a speaker mouthing a stimulus is dubbed with an auditory track differing by one consonant's place of articulation. People often report perceiving McGurk stimuli as a fusion of phonetic features from the auditory and visual signals. For example, auditory [ba] paired with visual [ga] or [da] is generally consciously perceived as *da*. This effect is remarkable because of its illusory status – listeners report perceiving tokens that are distinct from the auditory signal, even though the auditory input is perceptually clear.[1]

A visual signal can especially affect the perception of a degraded auditory signal. In addition to having a stronger influence in noisy environments, there is some evidence from perceptual identification tasks that subjects perceive the McGurk illusion more frequently when the auditory signal is less "good" than the integrated signal. For example, two previous studies (Barutchu et al., 2008; Brancazio, 2004) provide evidence for a lexical bias in auditory–visual integration. Subjects were shown incongruent auditory–visual stimuli, and reported perceiving the fused AV

[1] In the literature, the "McGurk Effect" is sometimes used to refer more narrowly to instances in which the resulting percept corresponds neither to the auditory nor the visual stimulus but rather to something between the two. It is also sometimes used more broadly to refer to instances in which the visual stimulus causes perceivers to report something other than what they would report if exposed to the auditory signal alone. In this article, we use "McGurk Effect" in its broader sense.

* Corresponding author at: Department of Cognitive Science, 9500 Gilman Drive, #0515, La Jolla, CA 92093-0515, United States.
E-mail address: rostrand@cogsci.ucsd.edu (R. Ostrand).

http://dx.doi.org/10.1016/j.cognition.2016.02.019
0010-0277/

percept more often when the auditory signal was a nonword (e.g., *besk*) than when the auditory signal was a real word (e.g., *beg*). Similarly, subjects reported perceiving the fused percept more often when the visual signal (and fused percept) was a word (e.g., *desk*) than when it was a nonword (e.g., *deg*; but see Sams, Manninen, Surakka, Helin, & Kättö, 1998, for conflicting results). Thus, visual information seems to affect perception when it enhances access to the lexicon, by increasing the likelihood that a nonword auditory signal is comprehended as a real word.

These studies suggest that lexical characteristics of the auditory and visual signals affect whether the visual signal influences the outcome of conscious perception – what listeners report they heard. However, it remains unknown whether the same lexical characteristics of the auditory and visual signals influence lexical access – that is, the extent to which a given stimulus input activates lexical representations and information in the associated semantic network. The present work addresses this second question, specifically investigating the timing of AV-integration, and the situations in which each sensory signal does or does not influence lexical access.

Our research speaks to an ongoing debate as to whether AV-integration is an early or late process – pre- or post-lexical – and thus whether the combined percept or the separate sensory signals drive lexical access. There is some evidence for both possibilities; however, much of the previous work used nonword syllables (such as [ba]) rather than real words, making it impossible to draw conclusions about the time course of AV-integration relative to lexical access specifically. The current work challenges the assumption of a strict pre- or post-lexical dichotomy and considers alternative points at which the auditory and visual signals could be integrated.

The dominant view is that AV-integration is a strictly early, pre-lexical process; that is, that the separate auditory and visual inputs are fused into a single stimulus before lexical access occurs (see Massaro & Jesse, 2007, for a discussion). In this case, the integrated McGurk percept – not the auditory signal – is the lookup key in the lexicon, accessing its own lexical-semantic entry and associates. This would imply that AV-integration operates in a purely bottom-up direction, and occurs similarly regardless of the lexicality or other characteristics of either sensory input signal. Supporting early integration, Sams et al. (1998) found that subjects were equally likely to fuse conflicting auditory and visual streams into nonwords as into words. This was true regardless of whether stimuli were isolated words or were predictable from the preceding sentence context, suggesting that AV-integration occurred before (and irrespective of) word identification (though note, as discussed below, Brancazio, 2004 reports different results). Additionally, some neuropsychological evidence suggests that AV-integration is an early process. Colin et al. (2002) exposed subjects to a high proportion of congruent AV stimuli (e.g., $bi_{Aud}bi_{Vis}$), interspersed with a few incongruent AV stimuli (e.g., $bi_{Aud}di_{Vis}$). The incongruent stimuli elicited a mismatch negativity (MMN), an automatic and pre-attentive electroencephalography (EEG) component. However, infrequent visual-only stimuli (e.g., $Ø_{Aud}di_{Vis}$ presented interspersed with frequent $Ø_{Aud}bi_{Vis}$) elicited no MMN. As infrequent visual stimuli seem not to elicit an MMN, the differing visual signals of the incongruent and congruent AV stimuli could not have triggered the observed MMN component, and thus subjects must have integrated the auditory and visual streams of the McGurk stimuli. And because the incongruent AV items elicited an MMN even though the auditory signal was identical to that of the congruent items (both bi_{Aud}), AV-integration must have occurred early in processing (before the MMN occurred), and the MMN must have reflected the integrated AV percept (see also Besle, Fort, Delpuech, & Giard, 2004; Colin, Radeau, Soquet, & Deltenre, 2004; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007; Soto-Faraco, Navarra, & Alsius, 2004). This view is also supported by models of AV-integration such as the Fuzzy Logical Model of Perception (FLMP; Massaro, 1987) or Pre-labeling Model (Braida, 1991). For example, the FLMP predicts that although the differing sensory signals are initially evaluated separately, they are integrated prior to perception and interpretation of the stimulus.

In contrast, other research suggests that AV-integration is a late, post-lexical process, and thus lexical access should occur based on the information in (one or both) separate, un-integrated sensory input signals. As the auditory stream is usually more informative about speech than lip-reading is, under this account, the activated lexical item should derive from the auditory stimulus. Only later, after lexical access, would AV-integration occur, producing the perceptual experience of the McGurk Effect. Thus the combined percept, and the word or nonword it forms, would not (initially) contact the lexicon. Unlike the early integration account, under the post-lexical (late AV-integration) account, AV-integration does *not* occur irrespective of the input signals' properties. In this case, some incongruent stimuli may never get integrated, or may take longer to do so. Supporting the late time course, Brancazio (2004) and Barutchu et al. (2008) found a lexical bias in the McGurk effect. They found that AV stimuli with nonword auditory signals were more likely to be perceived as the McGurk fusion than those with word auditory signals, implying that lexical access had already occurred on the unimodal auditory signal to determine its lexicality. (Note, however, that these results are inconsistent with the results of Sams et al., 1998; Brancazio, 2004 suggests that this is due to shortcomings of the McGurk stimuli used by Sams et al.)

Similarly, a late integration account raises the possibility that top-down factors and semantic knowledge might influence whether listeners perceive the McGurk Effect at all. Indeed, listeners report more McGurk illusions and rate their perception closer to the fused word when the AV fusion is semantically congruent with a preceding sentence (Windmann, 2004), suggesting that listeners have access to the meaning of the sensory input signals (and their semantic associates) before integrating them (or not). Models of AV-integration such as the Post-labeling Model (Braida, 1991) support a late time course of integration.

The results from the experiments reported here suggest that the time course of AV-integration is more nuanced than a strict binary choice between pre-lexical or post-lexical integration. We propose a third possibility – that the time course, and the likelihood of success, of AV-integration is dependent on the lexicality of the two input signals. Whether AV-integration occurs before or after lexical access could depend on properties of the specific auditory and visual inputs, rather than having a fixed time course. There are some hints in prior research supporting this hypothesis. For example, Baart and Samuel (2015) presented subjects with spoken words and nonwords that differed at the onset of the third syllable (like "banana" and "banaba"). Additionally, the third syllable was either presented auditory-only, visual-only (i.e., mouthed), or auditory–visual. They found that both lexical status and presentation modality modulated subjects' ERP activity. However, the two factors did not affect each other's degree of influence, and occurred at the same time points. Although Baart and Samuel (2015) did not test incongruent AV stimuli, their results suggest that lexical access and the integration of auditory and visual signals might, in certain circumstances, occur in parallel.

In the present experiments, subjects performed lexical decisions on auditory target items that were semantically related or unrelated to preceding auditory–visual primes. The primes were either created from mismatching AV signals (McGurk) or matching AV signals (congruent controls). This priming task allows for the detection of words that the auditory–visual prime stimuli activate in the lexicon. In Experiment 1, for each McGurk prime, either the auditory signal or the integrated auditory–visual (McGurk) percept was a word; the other was a nonword. Congruent primes paired