Some Behavioral and Neurobiological Constraints on Theories of Audiovisual Speech Integration: A Review and Suggestions for New Directions

Nicholas Altieri\textsuperscript{1,*}, David B. Pisoni\textsuperscript{2} and James T. Townsend\textsuperscript{2}

\textsuperscript{1} Department of Psychology, University of Oklahoma, 3100 Monitor Ave., Two Partners Place, Suite 280 Norman, OK 73072, USA
\textsuperscript{2} Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405, USA

Received 10 August 2010; accepted 3 July 2011

Abstract

Summerfield (1987) proposed several accounts of audiovisual speech perception, a field of research that has burgeoned in recent years. The proposed accounts included the integration of discrete phonetic features, vectors describing the values of independent acoustical and optical parameters, the filter function of the vocal tract, and articulatory dynamics of the vocal tract. The latter two accounts assume that the representations of audiovisual speech perception are based on abstract gestures, while the former two assume that the representations consist of symbolic or featural information obtained from visual and auditory modalities. Recent converging evidence from several different disciplines reveals that the general framework of Summerfield’s feature-based theories should be expanded. An updated framework building upon the feature-based theories is presented. We propose a processing model arguing that auditory and visual brain circuits provide facilitatory information when the inputs are correctly timed, and that auditory and visual speech representations do not necessarily undergo translation into a common code during information processing. Future research on multisensory processing in speech perception should investigate the connections between auditory and visual brain regions, and utilize dynamic modeling tools to further understand the timing and information processing mechanisms involved in audiovisual speech integration.

© Koninklijke Brill NV, Leiden, 2011

Keywords

Audio-visual speech perception, McGurk effect, multisensory enhancement

\* To whom correspondence should be addressed. E-mail: nick.altieri@ou.edu

© Koninklijke Brill NV, Leiden, 2011 DOI:10.1163/187847611X595864
1. Introduction

The ability to obtain visual information from the movements of a talker’s articulators through lip-reading is important for both normal-hearing and hearing-impaired listeners when perceiving speech (Erber, 1969; Sumby and Pollack, 1954). Sumby and Pollack (1954) demonstrated over 50 years ago that visual information in speech perception enhances accuracy scores across a wide range of signal-to-noise (S/N) ratios and that the proportion of audiovisual gain generally remains constant across different S/N ratios. Another illustration of the effect of visual information in speech perception is the well-known McGurk effect, a perceptual fusion that occurs when the auditory and visual signals are mismatched (McGurk and Macdonald, 1976). Specifically, the McGurk effect occurs when incongruent audiovisual information, such as an auditory /ba/ combined with a visually articulated (ga), yields a novel fusion of the two streams; in this case the fusion typically leads to the percept of /da/ (see Fowler and Dekle, 1991, for further illustrations). While many phenomena in audiovisual speech perception have been extensively researched, the neuro-cognitive mechanisms that operate on auditory and visual speech inputs during the integration process have yet to be clarified. Broadly speaking, two important issues need to be addressed.

One important question relates to the nature of the neuro-cognitive processes involved in multisensory integration in speech perception. This includes a formal/mathematical description characterizing how the time-varying dynamics of the audiovisual integration processes operate (e.g., Altieri, 2010; Altieri and Townsend, under review).

Another related question, and a significant focus of this article concerns the representations (i.e., phonetic, gestural, etc.) upon which the neuro-cognitive system operates during the integration process. Over the past several decades, several theoretical explanations have been proposed to account for such phenomena in audiovisual speech perception. In a seminal contribution to theoretical discussion on representational issues in multisensory processing, Summerfield (1987) discussed several accounts of multisensory integration in speech perception including: (1) integration of discrete phonetic features in which information about place (e.g., bilabial or velar) is obtained from the visual modality while information about manner (e.g., voiced or voiceless) is obtained from the auditory modality, (2) vectors describing the values of independent acoustical and optical parameters, (3) the filter function of the vocal tract and (4) articulatory dynamics of the vocal tract. More detailed descriptions of these accounts will be presented in the following sections.

1.1. General Background

After briefly reviewing the theoretical accounts outlined by Summerfield (1987), we propose a new theoretical framework. First, we should mention that an inherent core assumption in each of Summerfield’s proposed accounts is that auditory and visual unisensory information is translated into a common code prior to the conflux of the streams (see also Rosenblum, 2005). We argue that this assumption is unnecessary