Journal of Memory and Language 93 (2017) 276–303

Contents lists available at ScienceDirect

Journal of Memory and Language

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

The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration

Alastair C. Smith [a,*], Padraic Monaghan [b], Falk Huettig [a,c]

[a] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*
[b] *Department of Psychology, Lancaster University, Lancaster, UK*
[c] *Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands*

ARTICLE INFO

Article history:
Received 10 May 2015
revision received 23 August 2016
Available online 18 November 2016

Keywords:
Visual world paradigm
Visual attention
Spoken word recognition
Connectionist modelling
Multimodal processing

ABSTRACT

Ambiguity in natural language is ubiquitous, yet spoken communication is effective due to integration of information carried in the speech signal with information available in the surrounding multimodal landscape. Language mediated visual attention requires visual and linguistic information integration and has thus been used to examine properties of the architecture supporting multimodal processing during spoken language comprehension. In this paper we test predictions generated by alternative models of this multimodal system. A model (TRACE) in which multimodal information is combined at the point of the lexical representations of words generated predictions of a stronger effect of phonological rhyme relative to semantic and visual information on gaze behaviour, whereas a model in which sub-lexical information can interact across modalities (MIM) predicted a greater influence of visual and semantic information, compared to phonological rhyme. Two visual world experiments designed to test these predictions offer support for sub-lexical multimodal interaction during online language processing.

Introduction

One of the defining features of language is displacement, i.e., the fact that concepts need not refer to objects or events that are currently present (Hockett & Altmann, 1968). In line with this observation is a long tradition of research in the language sciences which has largely ignored potential influences of 'non-linguistic' information sources (e.g., Fodor, 1983). However, although language does not need to refer to objects which are physically present it is often used in such a way. Moreover, psycholinguistic research over recent years suggests that language processing (including spoken word processing) is highly interactive in terms of combining multiple information sources to form an interpretation of the signal (see Onnis & Spivey, 2012). It is therefore likely to be a profound misrepresentation to restrict models of spoken word recognition exclusively to auditory information, overlooking multimodal aspects of the speech processing system (e.g. Luce, Goldinger, Auer, & Vitevitch, 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010).

Indeed, the prevalence of ambiguity in natural language (Piantadosi, Tily, & Gibson, 2012) is evidence for the efficiency with which the human speech processing system integrates linguistic and extra-linguistic information. If we accept that language usage takes place in context (i.e., embedded within extra-linguistic factors, such as visual

[*] Corresponding author at: Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands.
E-mail address: alastair.smith@mpi.nl (A.C. Smith).

http://dx.doi.org/10.1016/j.jml.2016.08.005
0749-596X/© 2016 Elsevier Inc. All rights reserved.

environment, non-verbal communicative cues, world knowledge, and so on) then the amount of information an efficient language should convey must be less than the amount of information required out of context (Kurumada & Jaeger, 2015; Monaghan, Christiansen, & Fitneva, 2011). However, we know ambiguity in natural language is ubiquitous yet such ambiguity is rarely harmful to effective communication (Ferreira, 2008; Jaeger, 2006, 2010; Piantadosi et al., 2012; Roland, Elman, & Ferreira, 2006; Wasow & Arnold, 2003; Wasow, Perfors, & Beaver, 2005). This implies that the speech processing system is able to efficiently integrate extra-linguistic contextual information with the ambiguous speech stream it receives. The lack of explicit awareness we have of the level of ambiguity within the raw speech signal when processing speech in natural settings illustrates the speed and ease with which linguistic and non-linguistic information is integrated by the human speech processing system.

Models of speech recognition and speech comprehension have frequently overlooked this multimodal aspect of the speech processing system (e.g., Luce et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008; Scharenborg & Boves, 2010), with comparatively little known about the architecture that supports integration and the temporal structure of this process. In this study we test two explicit implementations of alternative hypotheses describing how visual, phonological and semantic information may be integrated when processing spoken words in a visual world. The first model is based on TRACE (McClelland & Elman, 1986) and multimodal information integration occurs over lexical representations. The alternative model permits integration of multimodal information over sub-lexical representations. These simulations generate similar predictions for the role of phonologically similar words in competition when the similarity is at the word onset. However, critically, they provide contrasting predictions for the influence of phonological rhyme information on fixation behaviour relative to visual and semantic information during online spoken word processing. We therefore tested these effects in two visual world eye-tracking experiments (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The results provide constraints on when and how such information is integrated in speech processing.

Models of multimodal integration during speech processing

A distinct division in perspectives continues to exist within both cognitive psychology and cognitive neuroscience regarding the characterisation of how and when non-linguistic and linguistic information interact during speech processing (e.g. Dilkina, McClelland, & Plaut, 2010; Leonard & Chang, 2014; Pulvermüller, Shtyrov, & Hauk, 2009).

The classical view within psycholinguistics argues that on hearing a spoken word information in the speech signal activates progressively larger units of representation within a modular phonological processing hierarchy, for example progressing from activation of primary phonetic features, to phonemes, to ultimately activating lexical units (e.g. McClelland & Elman, 1986). It is at this point, at the lexical level, that information in other modalities can connect to influence processing (e.g. Fodor, 1983; Friederici, 2002; Marslen-Wilson, 1987; Spivey, 2007), although such architectures can vary greatly in the extent to which information is able to interact between levels (see, e.g., McClelland, Mirman, & Holt, 2006; McQueen, Norris, & Cutler, 2006).

Alternatively, information in other modalities may be available to interact sub-lexically (e.g. Dilkina, McClelland, & Plaut, 2008; Dilkina et al., 2010; Gaskell & Marslen-Wilson, 1997; Pulvermüller et al., 2009). In such an architecture it becomes feasible for associations to develop between sub-lexical representations across modalities, for example between individual phonemes and individual semantic features.

In this paper we implement each of these alternative architectures in cognitively plausible (McClelland, Mirman, Bolger, & Khaitan, 2014) computational models. In both cases spoken word recognition and spoken word comprehension are framed in terms of multimodal constraint satisfaction (cf. MacDonald, Pearlmutter, & Seidenberg, 1994; McClelland, Rumelhart, & Hinton, 1986; McClelland et al., 2014), with words conceived as entities that connect representations across multiple modalities (e.g., phonological, orthographic, semantic, visual, etc.). In both models, speech processing occurs in a multimodal context, with activation of information passing between modalities to reflect real time sensory input. Both models are able to incorporate such multimodal cues to adapt their response in accordance to the current information available.

The two models differ however in the level at which multimodal information is able to interact. To represent a lexical level multimodal interaction model we extend the TRACE model of speech processing (McClelland & Elman, 1986) to allow activation cascading from visual and semantic representations to influence processing at the lexical level. TRACE provides a phonological processing hierarchy that allows activation to interact bi-directionally between three levels of representations: phonetic features, phonemes and words. We extend this system by injecting activation from visual and semantic levels into the TRACE hierarchy at the lexical level.

For contrast, we also implement a fully interactive system in which information at all levels of representation is free to combine across modalities. To represent such a system, we use the Multimodal Integration Model (MIM) of language processing which integrates concurrent phonological, semantic and visual information in parallel during spoken word processing (Smith, Monaghan, & Huettig, 2013, 2014a, 2014b; see also Monaghan & Nazir, 2009). The model is derived from the Hub-and-Spoke framework (Dilkina et al., 2008, 2010; Plaut, 2002; Rogers et al., 2004), a single system architecture that consists of a central resource (hub) that integrates and translates information between multiple modality specific sources (spokes). Critically, processing in the MIM is emergent, with minimal assumptions regarding initial connectivity or constraints on the flow of information within the network. Behaviour is thus a consequence of the system learning to map across modalities in which differing representational structures are embedded.