Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 59 (2014) 31–43

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Hearing words helps seeing words: A cross-modal word repetition effect

Patrick van der Zande [a,*], Alexandra Jesse [b], Anne Cutler [c,a]

[a] Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands
[b] Department of Psychology, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003, USA
[c] MARCS Institute, University of Western Sydney, Penrith South, New South Wales 2751, Australia

Received 18 March 2013; received in revised form 14 December 2013; accepted 6 January 2014
Available online 12 January 2014

Abstract

Watching a speaker say words benefits subsequent auditory recognition of the same words. In this study, we tested whether hearing words also facilitates subsequent phonological processing from visual speech, and if so, whether speaker repetition influences the magnitude of this word repetition priming. We used long-term cross-modal repetition priming as a means to investigate the underlying lexical representations involved in listening to and seeing speech. In Experiment 1, listeners identified auditory-only words during exposure and visual-only words at test. Words at test were repeated or new and produced by the exposure speaker or a novel speaker. Results showed a significant effect of cross-modal word repetition priming but this was unaffected by speaker changes. Experiment 2 added an explicit recognition task at test. Listeners' lipreading performance was again improved by prior exposure to auditory words. Explicit recognition memory was poor, and neither word repetition nor speaker repetition improved it. This suggests that cross-modal repetition priming is neither mediated by explicit memory nor improved by speaker information. Our results suggest that phonological representations in the lexicon are shared across auditory and visual processing, and that speaker information is not transferred across modalities at the lexical level.

Keywords: Speech perception; Audiovisual speech; Word repetition priming; Cross-modal priming

1. Introduction

Listeners encounter speech produced by many different speakers, whose articulators differ physiologically (Ladefoged, 1980; Laver and Trudgill, 1979) and whose dialectal or sociological backgrounds may also differ (Foulkes and Docherty, 2006), leading to specific idiosyncrasies in the way speech sounds are formed. Despite this speaker variability, spoken word recognition is generally quick and accurate. Listeners exploit recurrence of specific idiosyncrasies; words previously perceived are more efficiently recognised (Ellis, 1982; Jackson and Morton, 1984), and this is particularly true when the words are repeated by the same speaker (Goldinger, 1996; Mullennix et al., 1989; Schacter and Church, 1992).

Listeners also benefit from the availability of visual as well as auditory information about speech (Macleod and Summerfield, 1987; Reisberg et al., 1987; Sumby and Pollack, 1954). The benefit of visual speech information is particularly noticeable in situations where the auditory signal is difficult to interpret (Sumby and Pollack, 1954), but information from both sources is actually processed wherever possible (Arnold and Hill, 2001; McGurk and MacDonald, 1976; Reisberg et al., 1987). Visual speech facilitates the recognition of phonemes and words by providing segmental information that is complementary and redundant to the auditory signal (Grant et al., 1998; Jesse and Massaro, 2010; Summerfield, 1987; Walden et al., 1974). Visual speech also provides important prosodic information to help with speech recognition (Cvejic et al., 2012; Dohen et al., 2004; Jesse and McQueen,

* Corresponding author. Tel.: +31 6 28 116023.
E-mail addresses: P.Zande@gmail.com (P. van der Zande), AJesse@psych.umass.edu (A. Jesse), A.Cutler@uws.edu.au (A. Cutler).

0167-6393/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2014.01.001

2013; Krahmer and Swerts, 2004; Munhall et al., 2004). The visual speech signal thus constitutes an important source of information for listeners.

To understand spoken utterances, listeners must recognise the words they contain. This involves accessing the stored representations of these words in the listener's mental lexicon. Much recent research has addressed the content of such representations, and in particular the degree to which they may contain knowledge that is abstract, versus veridical traces of past recognition episodes. Evidence for the storage of episodic traces is provided by facilitation of recognition for words spoken in previously experienced voices (e.g., Mullennix et al., 1989); evidence for abstraction is provided by generalisation of learning about speaker-specific pronunciations to new words that are quite different from those experienced from a given speaker so far (e.g., McQueen et al., 2006). The consensus view has therefore come to be one that embraces lexical representation of both abstract and episodic information, with each type of information coming into play where task requirements encourage it (McLennan et al., 2003).

The simultaneous use of visual and auditory information to process speech bears on this issue, in that episodic traces of processing by different senses will differ in many ways. Particularly relevant is evidence for repetition priming across modalities. Repetition priming refers to facilitated recognition of words on second presentation (Jackson and Morton, 1984; Schacter and Church 1992); cross-modally, a spoken word is recognised more rapidly by listeners who have just seen a speaker articulate it (Buchwald et al., 2009; Kim et al., 2004). The auditory and visual input presumably activated the same representations in the perceiver's mental lexicon.

In these previous cross-modal priming studies, priming has been short-term (i.e., target immediately following prime). Such studies do not address the persistence of the facilitation. In the present study, we assess whether priming across modalities is long lasting by using a long-term (auditory-to-visual) priming paradigm. We also ask whether the priming involves phonological information. Long-term auditory-to-visual and visual-to-auditory word repetition priming occurs in semantic categorisation (Dodd et al., 1989), but in that task the priming could be either semantic or phonological in nature. The short-term visual-to-auditory priming results described above suggest a phonological locus of the cross-modal repetition priming (much like auditory-only repetition priming; Norris et al., 2006): the visual-only primes limit the range of phonemes used in both correct and incorrect responses to auditory targets (Buchwald et al., 2009), indicating that the cross-modal priming does not depend on correct identification of the prime. The long-term repetition-priming paradigm used in the present study provides a new view of the persistence of these effects, and by using a word identification task at test, we can also ask whether the locus of the priming effect is indeed phonological.

Moreover, we have not restricted our investigation of priming to speech from a single speaker. Speakers all have their own way of producing speech sounds, and speaker-specific idiosyncrasies occur in visual speech just as in auditory speech; also, speakers can differ widely in intelligibility (Bond and Moore, 1994; Ferguson, 2004; Gagné et al., 1994; Kricos and Lesner, 1982; Yehia et al., 1998). Perceivers clearly retain some speaker-specific information from exposure to a speaker, because recognition of subsequent speech from the same speaker is facilitated (Nygaard and Pisoni, 1998; Nygaard et al., 1994). Speaker variability taxes cognitive resources and reduces processing speed and accuracy due to the fact that such speaker information must be encoded; both auditory and visual speech are more accurately recognised with a single, constant speaker than when speakers vary from trial to trial (Creelman, 1957; Mullennix et al., 1989; Yakel et al., 2000). Crucially, speaker-specific knowledge acquired from visually presented speech benefits the subsequent recognition of auditory speech from the same speaker, suggesting that information about speaker idiosyncrasies is also encoded in a way that can generalise across the modalities (Rosenblum, 2008; Rosenblum et al., 2007). To put this generalisation to further test, we here investigate the reverse situation: does auditory exposure to a speaker's voice improve perceivers' subsequent identification of visually presented words from the same speaker? Even though not every visible movement in the speaker's face necessarily influences the resulting auditory signal, visual speech may hold sufficient information about that auditory signal to prime subsequent auditory recognition. But does auditory speech in turn provide good information about what the accompanying visual realisation would be?

Certainly there have been proposals that information about the shape of the vocal tract is extracted from auditory speech and used for auditory speech perception (Fowler et al., 2003; Liberman and Mattingly, 1985). Such proposals would indeed predict that hearing a speaker should provide sufficient information to affect the later processing of the speaker's visual speech. Also, the modality-general storage of speaker information argued for by Rosenblum (2008) and Rosenblum et al. (2007) on the basis of visual-to-auditory priming would likewise predict stronger auditory-to-visual priming for same-speaker than for different-speaker repetitions. Finding such a cross-modal speaker repetition effect would thus provide evidence that listeners can extract, from auditory speech, speaker-specific information that can then be readily applied to the perception of visual speech by the same speaker. Episodic traces could play a role, since if auditory speech is perceived in terms of the underlying gestures, lexical episodes obtained from listening would consist of this information and could then facilitate processing of new visual speech episodes involving the same gestures. In contrast, the absence of speaker repetition effects in auditory-to-visual priming would argue against such re-use of stored speaker-specific detail, or articulatory episodes being necessarily activated in word recognition irrespective of input modality.

We also include an explicit memory task to assess speaker repetition effects in explicit memory across