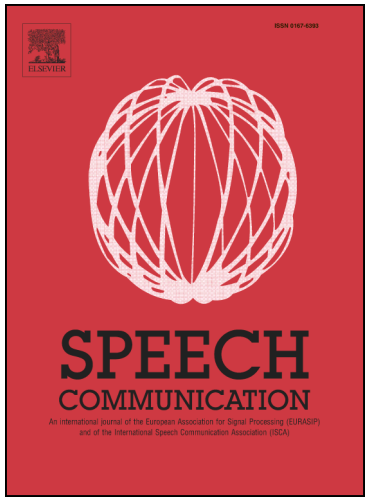

Hearing words helps seeing words: A cross-modal word repetition effect

Patrick van der Zande, Alexandra Jesse, Anne Cutler

PII: S0167-6393(14)00002-8
DOI: http://dx.doi.org/10.1016/j.specom.2014.01.001
Reference: SPECOM 2210

To appear in: *Speech Communication*

Received Date: 18 March 2013
Revised Date: 14 December 2013
Accepted Date: 6 January 2014

Please cite this article as: Zande, P.v.d., Jesse, A., Cutler, A., Hearing words helps seeing words: A cross-modal word repetition effect, *Speech Communication* (2014), doi: http://dx.doi.org/10.1016/j.specom.2014.01.001

Hearing words helps seeing words:

A cross-modal word repetition effect

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

[*]Corresponding author: p.zande@gmail.com, tel: +31 6 28 116023

[a]Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands

P.Zande@gmail.com

[b]Department of Psychology, University of Massachusetts Amherst, 135 Hicks Way, Massachusetts 01003, USA

AJesse@psych.umass.edu

[c]MARCS Institute, University of Western Sydney, Penrith South, New South Wales 2751, Australia

A.Cutler@uws.edu.au

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, in press; 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; 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 modality. In many previous studies, explicit memory for spoken words has shown effects of speaker repetition (Craik and Kirsner, 1974; Goldinger, 1996; Luce and Lyons, 1998; Palmeri et al., 1993). Words presented audiovisually are also better recognised as old if the voice of the speaker is preserved (Sheffert and Fowler, 1995; further, listeners in that study better remembered the voice in which sounds were produced than the face of the speaker producing them). Repetition priming has not necessarily displayed such effects (Luce and Lyons, 1998, with the same materials that had shown speaker repetition effects in explicit memory; however, Goldinger, 1996, found parallel implicit and explicit memory effects with a different task). Given that hearing a word produced by the same speaker a second time could

provide additional contextual cues for recognition, explicit memory might prove more susceptible than implicit memory to changes in surface form. Perceivers are certainly able to detect whether the same speaker produced auditory-only speech and visual-only speech, both for isolated words (Lachs and Pisoni, 2004) and for sentences (Kamachi et al., 2003).

In summary, this study investigates whether cross-modal effects of long-term word repetition priming appear in an auditory-to-visual priming paradigm with an identification task. Finding effects of word repetition priming across these modalities will strengthen previous evidence that the processing of auditory and visual speech involves the same lexical representations. We use long-term priming to ascertain whether cross-modal word repetition priming persists over large intervals, and we use an identification task to shed light on the putatively phonological locus of the priming effect. We further test whether speaker repetition effects occur across modalities, and whether speaker repetition affects the strength of repetition priming; here the results will provide evidence concerning storage of knowledge about speaker idiosyncrasies. Finally, we use an explicit memory task; if, as previous research suggests to be likely, only this task shows cross-modal effects of speaker repetition, then speaker-specific information, though associated with lexical representations that are shared across modalities, may not be necessarily activated in recognition irrespective of modality.

2. Experiment 1

2.1. Participants

Fifty-three native speakers of Dutch (mean age = 20.8; 10 male) were paid for their participation in Experiment 1. All participants reported normal hearing and normal or corrected-to-normal vision, and none had received prior explicit training in lipreading. Equipment failure caused the loss of data from six participants. The final data set for analysis came from 47 participants, of whom 23 heard Speaker 1 during the exposure phase and 24 heard Speaker 2. Eleven further participants from the same population took part in a pilot experiment (mean age = 21; all female).

2.2. Materials

The initial stimulus set consisted of 195 monosyllabic and disyllabic Dutch nouns, all morphologically simple. Words were selected such that the stimulus set included all ten viseme categories distinguished for Dutch (Van Son et al., 1994). Visemes are sets of speech sounds that are produced with similar external articulatory configurations, and cannot be conclusively distinguished from visual evidence alone; Dutch viseme categories are shown in Table 1.

--- INSERT TABLE 1 ABOUT HERE ---

One male and one female speaker of Dutch (Speaker 1 and 2, respectively) were recorded. Both speakers belonged to the same population as the participants and had not received any specific speech training. Recordings, with a stand-alone Sennheiser MKH50 microphone for the audio and a Sony DCR-HC1000e camera for the video, were made in front of a neutral background and each speaker was visible from the top of the shoulders to the top of the head. The speakers produced multiple tokens of all 195 words and were instructed to start and end each individual utterance with a neutral mouth position with the lips slightly open (as a result speakers' initial and final mouth shape was similar across items). A native speaker of Dutch (the first author) selected one audiovisual token of each word for the pilot study. Videos were digitised as uncompressed avi files (720 × 576 pixels) in PAL format. Auditory-only stimuli used the audio of the same tokens (sampling rate 44.1 kHz).

A pilot experiment was conducted, in which 11 participants (from the same population as for the main experiment) performed a visual-only identification task on all 195 words presented in random order. Participants were randomly assigned to lipread one speaker (six to Speaker 1, five to Speaker 2), and saw only this speaker throughout. Their task was to identify the words the speaker produced using visual

speech information only and to type in their response to each word on the computer keyboard. Before analysing participants' responses, typographical errors were corrected when it could clearly be determined what the intended response had been (e.g., misspellings, switched characters); where the intended response could not be unequivocally established, participants' input was left unchanged. Homophone responses were scored as correct. Phonetic transcriptions for all responses were added to the dataset using the Celex lexical database for Dutch (Baayen et al., 1993). Responses that did not occur in this database were considered incorrect but were not excluded from the analyses. Viseme transcriptions, using the Van Son et al. (1994) categories, were added to the dataset on the basis of the phonetic transcriptions.

As a measure of accuracy, we calculated the overlap between the visemes in the input and the visemes provided in participants' responses. This measure is less strict than a measure of phoneme overlap or correct word identification, since viseme categories include multiple phonemes and multiple responses may thus be scored as correct (e.g., answering /p/ to a visually presented /b/ would be correct as both are members of the {p} viseme). The viseme overlap score was calculated by counting the number of visemes in the response that also occurred in the input, divided by the larger of the total number of visemes in either the input or the response. The number of overlapping visemes was always divided by the

larger of the two totals to ensure that longer responses could not reach 100% correct simply due to exceeding the length of the input. Syllable boundaries were also counted so that participants' overlap score was higher when they provided an answer with the correct number of syllables. For example, if a participant saw the input *lamp* "lamp" and gave the response *lamp*, their viseme overlap score would be 100%. If the same participant had given the response *lam* "lamb", the viseme overlap score would be 75%. The response *lampen* "lamps" to *lamp* gave a viseme overlap score of 57%, since only four of the seven total characters in the response (i.e., *lam-pen*) overlap with the input visemes. In addition to the viseme overlap scores, we also recorded identification scores that showed whether participants had provided the correct word (scored in phonemes).

--- INSERT TABLE 2 ABOUT HERE ---

Independent samples t-test revealed no difference between the two speakers across the 195 pilot words for the correct word identification scores (Speaker 1: M = 7.08%; SD = 13.71%; Speaker 2: M = 8.10%; SD = 14.29%; t(388) = -0.72, p = 0.47) nor for the viseme overlap scores (Speaker 1: M = 57.73%; SD = 13.74%; Speaker 2: M = 59.07%; SD = 14.75%; t(388) = -0.92, p = 0.36). The 120 words that were lipread most

accurately for both speakers using the viseme overlap measure were selected for use in Experiments 1 and 2 (see Appendix A). Across the selected 120 target words, independent samples t-tests again showed no difference between the two speakers on the correct word identification scores (Speaker 1: M = 11.08%; SD = 16.05%; Speaker 2: M = 12.50%; SD = 16.41%; t(238) = -0.67, p = 0.50) and the viseme overlap scores (Speaker 1: M = 61.20%; SD = 13.27%; Speaker 2: M = 63.38%; SD = 13.86%; t(238) = -1.25, p = 0.21). These 120 words were divided into four word sets that were matched on their visual intelligibility for both speakers (see Table 2) and on average length in syllables. These lists were used to counterbalance the presentation of all words over the four experimental conditions. A 2 × 4 (speaker × word set) analysis of variance using viseme overlap scores as the dependent variable showed no significant main effects for speaker or word set and no significant interaction between the factors (all F values < 1). The word sets were rotated through the four experimental conditions in the test phases of Experiment 1 and 2 such that all 120 words occurred in all conditions.

2.3. Procedure

Participants were tested individually in a sound-attenuated booth. The experiment had two phases: an auditory-only exposure phase and a visual-only test phase. Each phase consisted of an identification task.

Participants were informed that there would be two separate phases, but were not told at the outset about the nature of the second task. In the exposure phase, the task was to identify 60 auditory-only words spoken by a single speaker. These 60 words were taken from two of the four experimental word sets with sets counterbalanced across participants. Twenty-three participants heard Speaker 1, the other 24 heard Speaker 2. Words were presented in random order over Sennheiser HD280 headphones at a fixed level. No noise was added to the auditory input. Participants were informed that a real Dutch word would be presented on each trial and that their task was to identify this word by typing in a response using the computer keyboard. Answers could be corrected until confirmed; pressing the return key confirmed an answer and initiated the next trial.

In the test phase, participants performed a visual-only identification task on all 120 words from the four word sets. Sixty of these 120 words had previously occurred in the auditory-only exposure phase and the other 60 were new. In both cases, half the tokens were produced by the exposure speaker and the other half were produced by the novel speaker. There were 30 tokens in each of the four experimental conditions (i.e., *new words/new speaker; new words/old speaker; old words/new speaker; old words/old speaker*). Presentation of words and speakers in each condition was counterbalanced across participants, and

the 120 items were presented in fully randomised order. As in the preceding phase, participants were told to expect only real Dutch words, typed their answers using the computer keyboard, and started new trials by pressing the return key to confirm an answer.

2.4. Analysis

Participants' responses were checked for typographical errors. Responses were scored for correct word recognition in phonemes. In addition, viseme overlap scores were calculated for responses given during the test phase using the same procedure as described in Section 2.2. The resulting data set was analysed with linear mixed-effect models in the R statistical package (R Development Core Team, 2007), using the lmer function of the lme4 library (Bates and Sarkar, 2007). The dependent variable was the binomial correct word identification (correct or incorrect). A logistic linking function was used for this categorical dependent variable. Best-fitting models were established through systematic model comparison using likelihood-ratio tests. Factors that did not contribute to a better model fit were pruned from the full model, starting from the factor with the highest *p*-value. All best-fitting models included participants as a random factor. Word repetition (old, new), speaker repetition (old, new) and exposure speaker (Speaker 1, Speaker 2) were evaluated as contrast-coded fixed factors.

3. Results and discussion

3.1. Exposure phase

Participants' auditory-only word identification scores in the exposure phase were high (M = 95.00%; SD = 5.64%). To test whether these results differed by exposure speaker, an lmer analysis was conducted with exposure speaker as a contrast-coded fixed factor and participants as random factor. The dependent variable was the binomial word recognition score (correct or incorrect). There was no significant effect of exposure speaker (β = -0.05, SE = 0.28, p = .83), i.e., Speaker 1 (M = 95.56%) and Speaker 2 (M = 94.51%) means did not reliably differ.

3.2. Test phase

Participants' visual-only word identification scores in the test phase were, as expected, relatively low (M = 15.71%; SD = 6.62%). Participants lipread repeated words more accurately than they lipread new words (β = -0.75, SE = 0.08, $p < .001$), indicating an overall effect of cross-modal word repetition priming. The effect of speaker repetition varied by exposure speaker (β = 0.84, SE = 0.15, $p < .001$) and the results were therefore further analysed separately by exposure speaker (see Table 3).

--- INSERT TABLE 3 ABOUT HERE ---

Participants who heard Speaker 1 during the auditory exposure phase were better at lipreading words that were repeated from the auditory-only exposure phase than they were at lipreading new words (β = -0.72, SE = 0.11, $p < .001$). This effect was not influenced by the identity of the speaker at test ($\chi^2(1) = 1.42$, $p = .23$). But the old speaker (i.e., here Speaker 1) was generally lipread better than the new speaker (Speaker 2; β = -0.40, SE = 0.11, $p < .001$). Participants who heard Speaker 2 during the auditory exposure phase were also better at lipreading repeated words than new words (β = -0.79, SE = 0.11, $p <$ .001). There was no interaction of word repetition and speaker repetition ($\chi^2(1) = 0.32$, $p = .57$), indicating that new words were lipread better than old words regardless of speaker identity. There was a main effect of speaker repetition ($\beta = 0.44$, $SE = 0.11$, $p < .001$), but the reverse of what had been predicted. Participants who had heard Speaker 2 during the auditory exposure were better at lipreading (the new) Speaker 1 than (the old) Speaker 2 at test. This reversal explains why the results of the combined model for the complete data set showed an interaction of exposure speaker and speaker repetition. Both participant groups lipread Speaker 1 better than Speaker 2, irrespective of their exposure speaker,

and despite the careful matching of word sets on the visual intelligibility of the speakers.

Additional analyses were performed on participants' ability to identify individual visemes in the visual-only test stimuli. Viseme identification was high (M = 64.11%; SD = 6.67%). Viseme overlap scores also showed an overall main effect of word repetition (β = -0.14, SE = 0.02, p < .001), i.e., participants' identification of individual visemes improved with word repetition. Again, analyses were split by exposure speaker because the effect of speaker repetition varied as a function of exposure speaker (β = 0.38, SE = 0.05, p < .001). These results revealed the same pattern as observed for the word identification results: Word repetition benefits viseme recognition, regardless of exposure speaker (Speaker 1: β = -0.12, SE = 0.04, p < .01; Speaker 2: β = -0.16, SE = 0.04, p < .001). Speaker 1 was again generally more intelligible than Speaker 2, thus reversing the effect of speaker repetition for participants who heard Speaker 2 during the auditory exposure phase (Speaker 1 as exposure speaker: β = -0.16, SE = 0.04, p < .001; Speaker 2 as exposure speaker: β = 0.18, SE = 0.04, p < .001). There was no interaction between word repetition and speaker repetition regardless of exposure speaker (Speaker 1: $\chi^2(1) = 1.87$, p = .17; Speaker 2: $\chi^2(1)$ = 0.01, p = 0.90). So, the viseme overlap results too show a benefit from prior auditory exposure on lipreading visual speech segments: Previously

heard speech affects perceivers' visual identification of individual speech segments.

Overall, the results of Experiment 1 revealed long-term repetition priming across modalities. Participants were better at identifying words and their parts from visual speech when they had previously heard the words. This cross-modal effect was found regardless of whether words were repeated by the same or a new speaker. The processing of auditory speech and visual speech thus utilises the same representations in the mental lexicon and the present results suggest that these representations do not contain speaker-specific information that is useful across modalities.

4. Experiment 2

Experiment 2 was identical to Experiment 1 except that, at test, participants were first asked to indicate whether the word they perceived visually was a new word or a word repeated from the auditory exposure phase (explicit memory task) before giving their identification response (identification task, reflecting implicit memory).

4.1. Participants

Fifty-two new participants from the same population as in Experiment 1 (mean age = 20.5; 9 male) took part in return for payment.

Four participants' data were lost due to equipment failure. The final analysed data set consisted of data from 48 participants, of whom 24 heard each speaker during exposure.

4.2. Materials and procedure

The materials in Experiment 2 were as in Experiment 1. The procedure differed from Experiment 1 only in that, during test, participants also performed a recognition memory task on each trial. Participants indicated after each visual-only presentation whether or not they had encountered the word during the auditory exposure phase, regardless of the identity of the speaker who produced the word; responses were given by pressing one of two buttons corresponding to labels "old" and "new" on the computer screen, with button assignment counterbalanced across participants. Participants had three seconds to respond and after a response had been given (or after the trial timed out) they identified the word by typing in their response as in Experiment 1. For the explicit memory task, the instructions stressed the importance of providing an answer as quickly and as accurately as possible.

4.3. Analysis

Typographical errors in participants' responses were again corrected, and results analysed using linear mixed-effect models, as for

Experiment 1. For the recognition memory task, the dependent variable was the binomial recognition memory judgement (correct or incorrect). A logistic linking function was used for this categorical dependent variable. The dependent variables for the identification tasks during exposure and at test were word identification scores. For the identification task at test, viseme overlap was also analysed. For both identification and recognition memory, word repetition (old or new), speaker repetition (old or new), and exposure speaker (Speaker 1 or Speaker 2) were evaluated as contrast-coded fixed factors. Participants were included as a random factor in all best-fitting models.

5. Results and discussion

5.1. Exposure phase

Participants' auditory-only word identification scores in the exposure phase were high (M = 96.22%; SD = 2.91%). A mixed-effect analysis evaluated exposure speaker as a contrast-coded fixed factor and participants as a random factor, with the binomial word recognition score (correct or incorrect) as the dependent variable. This revealed a significant effect of speaker (β = 1.20, SE = 0.23, $p < .001$). Although identification approached ceiling for items spoken by each speaker, there was a numerically small but reliable difference between the scores for Speaker 1 (M = 94.24%) and Speaker 2 (M = 98.19%).

5.2. Test phase

5.2.1. Recognition memory

Participants' overall correct word recognition was quite low (M = 48.18%; SD = 6.03%) and was similar following both exposure speakers (Speaker 1: M = 48.65%; SD = 5.05%; Speaker 2: M = 47.71%; SD = 6.95%). The complete model for the recognition memory task showed a significant three-way interaction (β = -0.58, SE = 0.21, p < .01), indicating that the results varied as a joint function of word repetition, speaker repetition, and exposure speaker. The results were therefore analysed separately by exposure speaker (see Figure 1).

For participants who heard Speaker 1 during auditory-only exposure, speaker repetition led to more correct new/old judgements for new but to fewer for old words (although the crossover interaction of word repetition and speaker repetition did not reach significance: β = -0.27, SE = 0.15, p = .07). No main effect was significant for this group (p > .05). Participants who heard Speaker 2 during exposure also showed a crossover interaction, in this case fully significant (β = 0.30, SE = 0.15, p < .05), but the pattern here is the reverse of that for the former group: more correct judgements for old words, fewer for new with repeated speaker. Again, no main effect was significant (p > .05). These results together suggest that new words were somewhat more accurately judged

as new when produced by Speaker 1 than when produced by Speaker 2. Overall, participants' scores were very close to chance, however.

Participants' recognition memory was also analysed for just those items that were correctly identified in the subsequent visual-only identification phase. No main effect and no interaction was significant in this analysis (all $p > .05$). Thus recognition memory was not affected by ability to identify the word in visual-only speech.

--- INSERT FIGURE 1 ABOUT HERE ---

A d' analysis, again using linear mixed-effect models, assessed participants' sensitivity in the recognition memory task. The effect of word repetition could not be evaluated since for the d' calculations hits were defined as correct "old" responses to old words and false alarms as incorrect "old" responses to new words. The best-fitting model showed no significant main effect of speaker repetition and no significant interaction between speaker repetition and exposure speaker (all $p > .05$). There was a trend (non-significant) towards a main effect of exposure speaker (β = -0.27, SE = 0.16, p = .08); participants who had heard Speaker 1 during exposure tended to have better recognition memory performance than those who had heard Speaker 2.

Although the recognition memory results indicated that new visually presented words were more accurately classified as new when spoken by Speaker 1, the *d′* results show that participants' ability to recognise whether they had previously heard a word was unaffected by who the speaker was, either at test or during exposure. This finding suggests that the inter-speaker difference in new/old classification accuracy may actually have been due to a bias in responses to the visually presented words.

5.2.2. Identification

The Experiment 2 visual-only word identification performance (M = 14.95%; SD = 7.19%) resembled that in Experiment 1. The overall results of the visual-only identification task showed a main effect of word repetition (β = -0.67, SE = 0.08, $p < .001$), replicating the Experiment 1 cross-modal repetition priming effect. Participants were better at lipreading words that they had previously heard in the auditory-only exposure than words that were new. There was again a significant interaction between speaker repetition and exposure speaker (β = 0.81, SE = 0.15, $p < .001$). The results were therefore again analysed separately by exposure speaker (see Table 3).

Visual-only identifications by participants who heard Speaker 1 in exposure showed a significant main effect of word repetition (β = -0.64,

SE = 0.11, $p < .001$); repeated words were more accurately lipread than new words. There was also a main effect of speaker repetition (β = -0.49, *SE* = 0.11, $p < .001$); the repeated speaker was easier to lipread than the new speaker. The word repetition effect was not influenced by speaker repetition ($\chi^2(1) = 2.07$, $p = .15$). Participants who heard Speaker 2 in exposure also lipread repeated words more accurately than new words (β = -0.70, *SE* = 0.11, $p < .001$) and also showed no significant interaction between word repetition and speaker repetition ($\chi^2(1) = 0.11$, $p = .74$). They showed a main effect of speaker repetition but, as in Experiment 1, this effect was reversed in that the new speaker (Speaker 1) was lipread more accurately than the old speaker (Speaker 2) (β = 0.33, *SE* = 0.10, $p < .01$). The speaker repetition effects are apparently driven by differences in visual intelligibility of the two speakers, not by memory factors.

Viseme overlap scores in Experiment 2 were also comparable to those in Experiment 1 (*M* = 62.08%; *SD* = 7.47%) and, as expected, higher than the correct word identification scores. Analyses on viseme overlap scores showed a similar pattern of results as the analyses on word scores. There was a main effect of word repetition (β = -0.13, *SE* = 0.03, $p < .001$) and an interaction of speaker repetition and exposure speaker (β = 0.29, *SE* = 0.05, $p < .001$). We therefore split the data by exposure speaker and found that participants who had heard Speaker 1 during exposure showed a significant interaction between the factors word

repetition and speaker repetition (β = 0.17, SE = 0.08, $p < .05$). This finding indicates that while both the main effect of word repetition (β = -0.13, SE = 0.04, $p < .001$) and the main effect of speaker repetition (β = -0.14, SE = 0.04, $p < .001$) were significant, the advantage of identifying visemes in the repeated words compared to new words was mainly driven by a difference in the old speaker condition. Participants who had heard Speaker 2 in exposure lipread new words better than old words (β = -0.12, SE = 0.04, $p < .01$) and were better at lipreading the new Speaker 1 than the old Speaker 2 (β = 0.29, SE = 0.05, $p < .001$). For these participants there was no significant interaction between the two main effects ($\chi^2(1) = 0.01$, $p = .94$).

In summary, the Experiment 2 identification results largely replicated those for Experiment 1. The main finding is again a cross-modal long-term effect of word repetition priming. This repetition priming holds despite the repeated words being presented in a different modality on first and second occurrence. Correct word identification results are the same across the two exposure groups. For the viseme overlap scores, however, participants who heard Speaker 1 in exposure subsequently lipread the visemes in repeated words by the same speaker better than the visemes in repeated words by the novel speaker. This suggests that in this case speaker repetition enhanced participants' ability to identify individual sounds. This influence of speaker repetition was not

found for participants who had heard Speaker 2 in exposure, however, nor did it appear in the correct word identification scores.

6. General discussion

Listeners are able to perceive words more quickly and more accurately when they have been encountered previously (Ellis, 1982; Jackson and Morton, 1984; Schacter and Church, 1992). This facilitation for processing repeated words is observed even when there is a change in modality between the first and second presentation of a word (Buchwald et al., 2009; Dodd et al., 1989; Kim et al., 2004). The present study has shown this cross-modal repetition effect to hold also when the words are encountered first in auditory-only speech and later in visual-only speech. This supports earlier claims (Buchwald et al., 2009; Kim et al., 2004) that auditory and visual speech processing draw on shared underlying lexical representations.

Our significant word repetition priming results come from a task requiring identification of visual speech, thus providing evidence of facilitation of phonological processing. Our task involved a longer delay between exposure and test than had been used in previous work, showing that cross-modal word repetition priming can persist over a significant time interval. Moreover, speaker familiarity did not modulate the size of

the effect, showing that the speaker information in lexical representations is not transferred across modalities.

Experiments 1 and 2 thus demonstrated that having heard a word previously improved how well the same word could later be identified from visual-only speech. Hearing a word improves the later identification both of the exact word and of the visemes that form that word. The effects of cross-modal word repetition on both word and viseme identification are statistically significant, though numerically relatively small: an improvement of 4-12% for the recognition of the complete word and about 4% for recognition of visemes. The stronger effect is thus on how visemes are interpreted as a word, suggesting that having heard words before also influences which lexical item is considered the most suitable interpretation of a given input.

Previous demonstrations of word priming across modalities (Buchwald et al., 2009; Kim et al., 2004) had shown a benefit for auditory word recognition after visual-only exposure; we have shown that the reverse (arguably more ecologically valid, as we will discuss below) situation also holds. The previous demonstrations of cross-modal repetition priming had also only used short prime-to-target intervals; our results show that such priming persists over a longer term, also consistent with real-world utility. Our results further confirm that auditory-to-visual priming, like auditory-to-auditory priming, involves phonological

representations in the lexicon, and the lexical representations invoked in processing are the same for visual and heard speech. Thus our results extend previous findings in at least three ways.

One puzzling aspect of our results is that although we found no speaker repetition effect on word repetition priming, we did observe an overall speaker effect in our data. This took the form of a global benefit for processing the visual speech of Speaker 1 over that of Speaker 2 at test, independent of exposure condition. Speakers certainly differ in intelligibility, with some speakers being consistently easier to understand than others (Bond and Moore, 1994; Gagné et al., 1994). It seems unlikely, however, that visual-only perception for Speaker 1 in Experiments 1 and 2 was inherently easier than visual-only perception for Speaker 2. The word stimuli, and the sets into which they were divided for the experiments, were closely matched on visual intelligibility across the speakers based on the result of the pilot study; if anything, it was Speaker 2 who was slightly easier to lipread there. To attempt to explain this pattern in our results, we separately analysed the results from the *new words/new speaker* condition of Experiments 1 and 2. These conditions are similar to the situation in the pilot experiment, in that in both cases participants had no prior exposure to either the speaker or the words they had to lipread. Independent samples t-tests on viseme overlap scores showed a small and not quite significant difference between the two

speakers in Experiment 1 ($t_1(45) = 1.76$, $p = 0.08$; $t_2(236) = 1.88$, $p = 0.06$), but a significant difference between them in Experiment 2 ($t_1(46) = 2.33$, $p = 0.05$; $t_2(236) = 2.72$, $p < 0.01$). In both cases, the viseme overlap scores for Speaker 1 were higher than those for Speaker 2. No such difference exists between speakers for the same words in the pilot experiment ($t_1(9) = -0.34$, $p = 0.74$; $t_2(236) = -1.19$, $p = 0.24$). Note that the overall mean performance in the *new words/new speaker* condition of Experiments 1 and 2 was similar to that in the pilot for this subset of words (all p values > 0.05).

It thus seems unlikely that Speaker 1 was just generally easier for participants to lipread[1]. Cross-speaker differences in visual-only identification scores in the two main experiments were also not due to differences in *auditory* identification of the speaker's speech during the auditory-only exposure. Although we found a significant difference in identification scores for Speaker 1 and 2 in Experiment 2, listeners' auditory performance was actually worse for Speaker 1 than for Speaker 2. We therefore can only suggest that Speaker 1's advantage in the experimental situation reflects some as yet unidentified dimension of visual articulation that can prove memorable and useful for recognising articulated versions of previously heard words. This topic certainly

[1] The difference between the two speakers is also unlikely to have been due to sex differences between the participant groups. Sex differences in visual speech recognition are controversial (see, e.g., Irwin et al., 2006; Strelnikov et al., 2009); we found better visual-only identifications for Speaker 1 than for Speaker 2, independent of the participants' sex.

deserves further empirical investigation[2], but does not affect the conclusions drawn from the present study.

Those conclusions, as described earlier, primarily concern the representations involved in processing speech in the auditory and visual modalities. With respect to the common representations that subserve both auditory and visual lexical processing, our results are not consistent with obligatory availability of all traces of prior experience irrespective of modality. As described in the introduction, overall evidence is in favour of both abstract and episodic components to lexical representations. Spoken-word recognition studies suggest that the degree to which each component type is called into play depends on the level of processing involved; easy tasks, requiring only shallow processing, are more likely to engage veridical traces in the lexicon (McLennan et al., 2003), while harder tasks, including priming over longer terms, are more likely to call on abstract knowledge such as the canonical form of words (McLennan et al., 2003; Sumner and Samuel, 2005). It is certainly the case that our participants found the tasks we gave them hard! Thus it is consistent with this depth of processing account that although in our study listeners' visual-only identification performance was better for repeated words than for new words, the magnitude of this effect of word repetition was not modulated by changes in the identity of the speaker.

[2] We could find no systematic relation between degree of priming and particular phonemes or visemes uttered by Speaker 1.

The lack of a speaker repetition effect here suggests that the processing involved abstract components of underlying representations, and the lexical content that was primed did not include speaker-specific detail.

It is in fact unclear to what extent speaker information can transfer across modality at all. Stored indexical information about speakers in lexical representations could be modality-specific even though the representations are used in more than one modality. If this is the case, then both speakers perceived during the visual-only identification task in the test phase could be considered new speakers because neither one had previously been perceived visually. Although Rosenblum and colleagues (2007) have shown transfer of speaker-specific information across modalities, there are methodological differences between that study and ours. One potentially critical difference is that Rosenblum et al. gave listeners substantially more exposure, in sentences rather than in isolated words. Listeners have been shown to tune in to different speaker-specific properties depending on the kind of speech materials they experience (Cvejic et al., 2012; Grant et al., 1998; Nygaard and Pisoni, 1998), so that speaker-specific information obtained from isolated words could be less susceptible to cross-modal transfer than information from sentences. Future research should assess auditory-to-visual transfer of speaker-specific information gained from sentences rather than from words. Most importantly, Rosenblum et al. showed transfer of indexical information

from visual to auditory speech, while we examined transfer from auditory to visual speech. It could thus also be the case that visual speech can provide useful information about auditory idiosyncrasies, but auditory speech is insufficient to define visual idiosyncrasies. This interpretation is consistent with our own earlier finding that listeners' retuning of auditory phonetic categories by the use of lexical knowledge (Jesse and McQueen, 2011; McQueen et al., 2006) does not transfer to visual categories unless listeners have also been exposed to a speaker's visual speech (Van der Zande et al., 2013). To draw this conclusion from our present finding of no cross-modal speaker repetition effect would of course be problematic for speech perception theories, such as motor theory or direct realism, that propose perception of auditory speech input in terms of the underlying gestures of the speaker's vocal tract (Fowler et al., 2003; Liberman and Mattingly, 1985). If listeners are able to extract such information about the movements or position of the speaker's articulatory features from the auditory signal, prior experience with a speaker's voice should also be useful in subsequent processing of visual-only speech. This was, however, not what we have observed.

Cross-modal speaker repetition also failed to affect *explicit* recognition memory. Listeners in Experiment 2 were equally likely to correctly classify words as "old" for different-speaker as for same-speaker repetitions. This is in contrast to prior findings of a same-speaker

advantage in auditory explicit memory (Goldinger, 1996; Sheffert and Fowler, 1995). When a task involves different modalities, therefore, it does not seem that repetition of the speaker will improve explicit memories of repeated words. Although remembering 60 individual words from auditory speech without an explicit prompt to do so may well have been a difficult task for our participants, it was not beyond the capability of listeners in other long-term recognition memory studies, some with even higher numbers of items (Bradlow et al., 1999; Craik and Kirsner, 1974; Schacter and Church, 1992; Sheffert, 1998). Rather, the finding that speaker repetitions failed to facilitate explicit memory of repeated words across modalities suggests that, as argued above for the visual-only identification task, the depth of processing required by any difficult task is one which calls on the more abstract components of the words' lexical representations.

Finally, we stress that the perception of just visual speech without auditory information plays only a limited role in our normal interaction with others. Auditory-only communication (e.g., telephone conversation) and audiovisual communication (e.g., face-to-face interaction) are far more likely to occur. Although we may see many people speaking together from afar without ever hearing them, choosing to communicate with someone through visual-only speech production is, for people with normal hearing, quite rare. Note, though, that it is most likely to happen

with speakers with whom we are familiar and whom we have heard speak before. In most cases, then, visual-only exposure before auditory-only exposure (as used by Rosenblum et al., 2007) is unlikely because familiarity with a speaker through auditory speech will usually precede familiarity with a speaker on the basis of visual-only speech. When someone mouths something to us across a busy conference room, it would be beneficial for our visual-only identification performance if we could be primed by auditory words perceived earlier. Our results show that such priming across modalities indeed occurs, even though it may be limited in its extent. In the same situation, our visual-only identification of speech from an *unfamiliar* speaker will also benefit from containing words that we have recently perceived auditorily, showing that when necessary we can even lipread people whom we have not heard before.

6.1. Conclusions

This study investigated the effects of word repetition and speaker repetition on implicit and explicit memory in an auditory-to-visual, long-term cross-modal priming paradigm. The results indicate that auditory processing and visual processing share lexical representations, because the processing of repeated words is facilitated across speech modalities. The shared components of these lexical representations are not adjusted on the basis of speaker-specific information. Repeated words and their

segments are consistently identified better than new words, regardless of the identity of the speaker. Neither implicit memory nor explicit memory of repeated words was enhanced by repetitions being produced by the same speaker on both instances. Speaker-specific information therefore may not be transferable across modalities at the lexical level.

References

Arnold, P., Hill, F., 2001. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. Brit. J. Psychol. 92, 339-355.

Baayen, R.H., Piepenbrock, R., Van Rijn, H., 1993. The Celex lexical database [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Bates, D., Sarkar, D., 2007. lme4: Linear mixed-effects models using S4 classes [Software]. Available from http://lme4.r-forge.r-project.org/. Last accessed March 7th, 2013.

Bond, Z., Moore, T., 1994. A note on the acoustic-phonetic characteristics of inadvertently clear speech. Speech Commun. 14, 325-337.

Bradlow, A.R., Nygaard, L.C., Pisoni, D.B., 1999. Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Percept. Psychophys. 61, 206-219.

Buchwald, A.B., Winters, S.J., Pisoni, D.B., 2009. Visual speech primes open-set recognition of spoken words. Lang. Cogn. Process. 24, 580-610.

Craik, F.I.M., Kirsner, K., 1974. The effect of speaker's voice on word recognition. Q. J. Exp. Psychol. 26, 274-284.

Creelman, C.D., 1957. The case of the unknown talker. J. Acoust. Soc. Am. 29, 655.

Cvejic, E., Kim, J., Davis, C., 2012. Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. Cognition 122, 442-453.

Dodd, B., Oerlemans, M., Robinson, R., 1989. Cross-modal effects in repetition priming: A comparison of lipread, graphic and hear stimuli. Visible Lang. 22, 58-77.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A., Schwartz, J.-L., 2004. Visual perception of contrastive focus in reiterant French speech. Speech Commun. 44, 155-172.

Ellis, A., 1982. Modality-specific repetition priming of auditory word recognition. Curr. Psychol. Res. 2, 123-128.

Ferguson, S.H., 2004. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. J. Acoust. Soc. Am. 116, 2365-2373.

Foulkes, P., Docherty, G., 2006. The social life of phonetics and phonology. J. Phonetics 34, 409-438.

Fowler, C.A., Brown, J.M., Sabadini, L., Welhing, J., 2003. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. J. Mem. Lang. 49, 396-413.

Gagné, J.-P., Masterson, V.M., Munhall, K.G., Bilida, N., 1994. Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. J. Acad. Rehabil. Audiol. 27, 135-158.

Goldinger, S.D., 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. J. Exp. Psychol. Learn. 22, 1166-1183.

Grant, K.W., Walden, B.E., Seitz, P.F., 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. J. Acoust. Soc. Am. 103, 2677-2690.

Irwin, J.R., Whalen, D.H., Fowler, C.A., 2006. A sex difference in visual influence on heard speech. Percept. Psychophys. 68, 582-592.

Jackson, A., Morton, J., 1984. Facilitation of auditory word recognition. Mem. Cognition 12, 568-574.

Jesse, A., Massaro, D.W., 2010. The temporal distribution of information in audiovisual spoken-word identification. Atten. Percept. Psychophys. 72, 209-225.

Jesse, A., McQueen, J.M., 2011. Positional effects in the lexical retuning of speech perception. Psychon. Bull. Rev. 18, 943-950.

Jesse, A., McQueen, J.M., In press. Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition. Q. J. Exp. Psychol.

Kamachi, M., Hill, H., Lander, K., Vatikiotis-Bateson, E., 2003. 'Putting the face to the voice': Matching identity across modality. Curr. Biol. 13, 1709-1714.

Kim, J., Davis, C., Krins, P., 2004. Amodal processing of visual speech as revealed by priming. Cognition 93, B39-B47.

Krahmer, E.J., Swerts, M., 2004. More about brows: A cross-linguistic analysis-by-synthesis study, in: Ruttkay, Z., Pelachaud, C. (Eds), From brows to trust: Evaluating Embodied Conversational Agents. Kluwer Academic Publishers, Dordrecht, pp. 191-216.

Kricos, P.B., Lesner, S.A., 1982. Differences in visual intelligibility across speakers. The Volta Review, 84, 219-225.

Lachs, L., Pisoni, D.B., 2004. Cross-modal source information and spoken word recognition. J. Exp. Psychol. Hum. Percept. Perform. 30, 378-396.

Ladefoged, P., 1980. What are linguistic sounds made of? Language 56, 485-502.

Laver, J., Trudgill, P., 1979. Phonetic and linguistic markers in speech, in: Scherer, K.R., Giles, H. (Eds.), Social markers in speech. Cambridge University Press, Cambridge, pp. 1-32.

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

Luce, P.A., Lyons, E.A., 1998. Specificity of memory representations for spoken words. Mem. Cognition 26, 708-715.

Macleod, A., Summerfield, Q., 1987. Quantifying the contribution of vision to speech perception in noise. Brit. J. Audiol. 21, 131-142.

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

McLennan, C.T., Luce, P.A., Charles-Luce, J., 2003. Representation of lexical form. J. Exp. Psychol. Learn. 29, 539-553.

McQueen, J.M., Cutler, A., Norris, D., 2006. Phonological abstraction in the mental lexicon. Cogn. Sci. 30, 1113-1126.

Mullennix, J.W., Pisoni, D.B., Martin, C.S., 1989. Some effects of talker variability on spoken word recognition. J. Acoust. Soc. Am. 85, 365-378.

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E., 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychol. Sci. 15, 133-137.

Norris, D., Butterfield, S., McQueen, J.M., Cutler, A., 2006. Lexically guided retuning of letter perception. Q. J. Exp. Psychol. 59, 1505-1515.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60, 355-376.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech perception as a talker-contingent process. Psychol. Sci. 5, 42-46.

Palmeri, T.J., Goldinger, S.D., Pisoni, D.B., 1993. Episodic encoding of voice attributes and recognition memory for spoken words. J. Exp. Psychol. Learn. 19, 309-328.

R Development Core Team, 2007. R: A language and environment for statistical Computing [Software]. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.r-project.org/. Last accessed June 8th, 2010.

Reisberg, D., McLean, J., Goldfield, A., 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli, in: Dodd, B., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading. Lawrence Erlbaum, London, U.K., pp. 97-113.

Rosenblum, L.D., 2008. Speech perception as a multimodal phenomenon. Curr. Dir. Psychol. Sci. 17, 405-409.

Rosenblum, L.D., Miller, R.M., Sanchez, K., 2007. Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. Psychol. Sci. 18, 392-396.

Schacter, D.L., Church, B.A., 1992. Auditory priming: Implicit and explicit memory for words and voices. J. Exp. Psychol. Learn. 18, 915-930.

Sheffert, S.M., 1998. Voice-specificity effects on auditory word priming. Mem. Cognition 26, 591-598.

Sheffert, S.M., Fowler, C.A., 1995. The effects of voice and visible speaker change on memory for spoken words. J. Mem. Lang. 34, 665-685.

Strelnikov, K., Rouger, J., Lagleyre, S., Fraysse, B., Deguine, O., Barone, P., 2009. Improvement in speech-reading ability by auditory training: Evidence from gender differences in normally hearing, deaf and cochlear implanted subjects. Neuropsychologia 47, 972-979.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212-215.

Summerfield, Q., 1987. Some preliminaries to a comprehensive account of audio-visual speech perception, in: Dodd, B., Campbell, R. (Eds.), Hearing by Eye: The Psychology of Lip-Reading. Lawrence Erlbaum, London, U.K., pp. 3-51.

Sumner, M., Samuel, A.G., 2005. Perception and representation of regular variation: The case of final /t/. J. Mem. Lang. 52, 322-338.

Van der Zande, P., Jesse, A., Cutler, A., 2013. Lexically guided retuning of visual phonetic categories. J. Acoust. Soc. Am. 134, 562-571.

Van Son, N.J.D.M.M., Huiskamp, T.M.I., Bosman, A.J., Smoorenburg, G.F., 1994. Viseme classifications of Dutch consonants and vowels. J. Acoust. Soc. Am. 96, 1341-1355.

Walden, B.E., Prosek, R.A., Worthington, D.W., 1974. Predicting audiovisual consonant recognition performance of hearing-impaired adults. J. Speech Hear. Res. 17, 270-278.

Yakel, D.A., Rosenblum, L.D., Fortier, M.A., 2000. Effects of talker variability on speechreading. Percept. Psychophys. 62, 1405-1412.

Yehia, H., Rubin, P., Vatikiotis-Bateson, E., 1998. Quantitative association of vocal-tract and facial behavior. Speech Commun. 26, 23-43.

Appendix

Viseme Overlap for All 120 Experimental Words.

			Pilot		Experiment 1		Experiment 2	
Dutch word	Dutch transcription	English gloss	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2
baby	ˈbeː-bi	baby	62.22	75.56	63.64	63.23	53.15	66.96
bar	bɑːr	bar	79.44	57.38	86.44	60.80	84.78	45.84
beek	beːk	brook	50.00	53.33	55.72	81.94	51.32	76.39
bende	ˈbɛn-də	gang	62.04	78.33	45.34	70.19	50.20	67.99
bezem	ˈbeː-zəm	broom	51.59	48.10	55.57	71.70	49.26	78.49
bijbel	ˈbɛi-bəl	bible	63.89	59.17	84.00	73.78	72.00	73.94
boef	buf	thug	76.67	40.00	73.96	33.86	67.01	44.10

boel	bul	bunch	58.89	66.67	75.05	68.75	64.51	65.56
bom	b□m	bomb	78.33	79.50	95.83	51.42	98.61	61.20
bon	b□n	ticket	67.78	95.00	65.53	81.12	69.93	73.68
boog	bo□x	arch	57.22	70.00	76.27	74.38	85.07	64.58
boom	bo□m	tree	82.14	88.00	97.71	64.36	92.33	74.70
burger	□b□r-x□r	citizen	42.86	42.86	34.42	50.80	37.27	45.13
cheque	□□k	check	58.33	59.00	68.12	70.05	63.26	63.31
chic	□ik	chic	66.67	46.33	84.91	69.57	77.08	65.69
faam	fa□m	fame	60.56	55.00	66.30	52.60	74.03	48.89
fabel	□fa□-b□l	fable	78.57	71.43	90.97	64.91	87.70	69.54
fan	f□n	fan	45.00	63.33	48.12	68.84	55.83	69.73
feit	f□□t	fact	85.83	61.33	81.16	49.63	77.78	56.80

fik	f□k	fire	72.22	85.33	83.33	67.02	76.88	72.71
folder	□f□l-d□r	leaflet	59.82	77.14	59.86	49.72	56.55	43.15
fout	f□ut	error	49.35	62.17	65.14	66.39	78.87	65.32
fuif	fœyf	party	62.22	50.67	73.67	45.50	63.67	46.75
fut	f□t	energy	48.61	55.00	64.47	44.66	51.39	48.06
gaas	xa□s	gauze	52.50	38.67	67.54	39.01	62.92	42.22
gang	x□ŋ	hallway	55.33	78.67	63.12	69.17	63.54	63.61
gevel	□xe□-v□l	façade	67.86	71.43	68.45	76.29	66.77	69.74
gif	x□f	poison	51.94	70.33	68.75	53.13	66.67	56.60
gil	x□l	yell	41.52	72.50	40.33	55.61	40.36	55.63
gordel	□x□r-d□l	seatbelt	33.33	48.57	48.21	53.42	41.67	42.26
kamer	□ka□-m□r	room	58.33	63.33	64.49	70.04	59.92	53.99

kater	ˈka-tər	hangover	66.67	67.50	67.99	62.67	65.97	53.67
keizer	ˈkɛi-zər	emperor	68.25	54.76	48.41	54.61	47.22	52.18
kip	kɪp	chicken	70.56	75.71	67.39	72.35	68.24	82.15
koffer	ˈkɔ-fər	suitcase	62.70	63.33	61.83	57.14	60.88	72.42
kom	kɔm	bowl	55.00	83.33	54.87	62.48	58.28	61.12
kop	kɔp	mug	83.61	66.67	86.81	50.74	84.17	59.08
kuif	kœyf	quiff	26.67	46.67	47.92	45.98	46.81	60.87
lach	lɑx	smile	83.33	86.67	72.22	65.58	70.35	67.69
leger	ˈle-xər	army	80.56	66.67	74.74	48.61	65.28	52.78
leraar	ˈle-raːr	teacher	69.44	66.67	75.78	61.56	60.88	63.10
les	lɛs	lesson	65.56	41.33	62.33	59.00	52.14	52.05
liefde	ˈliv-də	love	59.26	63.33	70.14	78.26	56.65	69.25

maag	ma□x	stomach	72.22	88.33	79.93	65.58	70.97	62.50
merel	□me□-r□l	blackbird	56.75	80.00	63.10	72.59	61.61	66.91
moed	mut	courage	64.58	61.67	63.57	61.30	61.96	62.29
mok	m□k	mug	62.78	74.67	62.77	70.36	65.72	72.92
motor	□mo□-t□r	motorcycle	56.75	73.33	73.51	72.49	67.60	61.31
mug	m□x	mosquito	65.28	43.33	71.53	75.00	67.85	63.74
muis	mœys	mouse	28.61	58.33	41.81	64.33	37.29	56.71
muur	myr	wall	78.06	88.33	61.41	72.78	69.62	57.38
naad	na□t	seam	59.44	33.05	70.80	44.79	67.15	40.14
nagel	□na□-x□l	nail	61.90	57.50	61.59	50.00	68.45	55.85
neef	ne□f	cousin (M)	68.06	39.83	79.86	64.54	77.85	53.45
negen	□ne□-x□	nine	65.24	72.00	66.19	69.78	63.91	61.07

nek	n□k	neck	63.89	60.33	57.61	62.85	56.67	67.15
nier	nir	kidney	59.44	73.33	57.15	84.72	56.79	74.16
nis	n□s	niche	61.11	72.33	77.29	58.49	56.46	59.22
noot	no□t	nut	47.33	86.67	52.46	73.96	50.70	68.54
nummer	□n□-m□r	number	63.89	63.33	73.44	54.37	53.27	60.16
parel	□pa□-r□l	pearl	41.67	70.00	58.12	71.74	46.51	52.26
pas	p□s	pass	65.28	86.67	70.20	79.84	70.27	69.27
paus	p□us	pope	80.00	63.33	71.01	69.38	66.81	64.18
piek	pik	peak	80.56	81.67	75.16	66.23	71.16	64.77
pijp	p□□p	pipe	53.10	74.17	75.43	48.61	72.33	50.95
pit	p□t	stone	64.09	57.57	77.18	65.86	79.65	67.57
poging	□po□-x□ŋ	attempt	59.23	65.83	52.38	58.34	53.37	54.64

pool	po□l	pole	42.78	78.67	71.88	72.59	75.35	70.42
put	p□t	well	55.56	68.00	58.33	62.20	56.88	60.91
raad	ra□t	advice	55.56	57.33	72.78	47.39	69.20	45.31
reep	re□p	strip	41.67	71.90	51.25	72.28	66.36	81.04
regen	□re□-x□	rain	56.67	52.67	59.86	43.47	52.71	47.28
rel	r□l	riot	50.56	70.00	40.83	58.17	36.81	50.40
riem	rim	belt	72.22	76.90	61.81	63.76	77.08	73.19
ring	r□ŋ	ring	47.22	72.00	50.35	64.49	56.25	54.31
rook	ro□k	smoke	31.03	73.67	49.54	55.42	50.14	53.62
rubber	□r□-b□r	rubber	67.17	56.67	74.90	38.61	63.89	54.46
rug	r□x	back	42.78	55.00	70.83	60.20	68.06	65.56
ruzie	□ry-zi	row	42.86	57.33	67.47	47.06	60.14	45.27

sap	s□p	juice	69.44	53.33	68.42	46.91	63.54	55.75
satan	□sa□-t□n	satan	67.46	50.95	60.23	52.03	57.00	54.80
saus	s□us	sauce	34.72	40.00	41.10	69.24	36.23	56.81
sein	s□□n	sign	61.11	49.17	66.01	47.69	67.71	58.48
set	s□t	set	75.48	50.79	74.17	57.27	63.02	47.30
shampoo	□□□m-po□	shampoo	51.39	45.24	78.34	51.43	81.25	47.45
sik	s□k	goatee	66.67	31.67	79.62	65.30	72.83	68.47
sinas	□si-n□s	orange soda	58.33	54.29	56.14	65.36	51.12	62.24
soep	sup	soup	70.00	55.00	74.64	75.28	64.58	69.38
suiker	□sœy-k□r	sugar	51.19	66.67	46.14	74.68	38.42	60.76
taak	ta□k	task	73.61	52.00	77.90	48.97	79.58	48.00

taal	ta□l	language	38.89	43.33	36.11	28.24	47.15	30.95
tafel	□ta□-f□l	table	97.22	67.14	95.96	71.63	92.66	71.97
tak	t□k	branch	56.67	68.33	76.04	65.49	82.64	61.31
tempel	□t□m-p□l	temple	78.87	73.77	77.10	68.06	73.81	66.15
titel	□ti-t□l	title	50.00	42.22	54.96	61.73	57.74	60.86
toeval	□tu-v□l	coincidence	63.89	66.19	71.22	42.34	60.91	59.82
toon	to□n	tone	51.39	45.33	57.99	42.33	49.93	41.62
vaas	va□s	vase	70.56	42.86	72.60	59.28	81.80	45.66
vak	v□k	square	70.56	61.67	79.24	68.35	74.64	54.87
val	v□l	fall	38.06	32.33	45.14	44.71	57.99	37.64
vat	v□t	barrel	71.11	85.00	64.44	46.35	66.83	54.58
veer	ve□r	feather	66.67	83.33	70.83	73.20	69.72	68.25

vijf	v□□f	five	77.78	90.00	85.42	79.30	70.49	77.08
voedsel	□vut-s□l	food	73.02	68.57	69.94	63.74	59.33	66.07
voeg	vux	joint	51.51	57.90	82.23	74.58	68.91	78.59
vogel	□vo□-x□l	bird	78.57	60.12	92.64	82.61	93.48	77.98
vuur	vyr	fire	51.79	56.29	73.08	66.72	72.00	50.84
zaad	za□t	seed	70.00	55.00	62.63	46.27	61.76	46.05
zak	z□k	sack	77.78	67.00	72.27	66.84	80.43	62.13
zebra	□ze□-bra□	zebra	45.24	70.48	61.41	61.34	57.77	66.05
zeef	ze□f	sieve	54.37	49.90	64.20	47.45	76.13	55.73
zeep	ze□p	soap	76.39	68.00	78.47	58.85	63.06	70.67
zeil	z□□l	tarpaulin	38.33	51.67	60.97	54.48	52.37	50.86
zes	z□s	six	59.52	75.33	63.29	66.11	63.47	48.13

zet	z□t	move	65.00	70.50	73.60	59.91	69.06	56.16
zoemer	□zu-m□r	buzzer	61.90	66.67	61.37	69.18	72.72	60.68
zoen	zun	kiss	65.00	63.33	71.56	62.92	64.58	57.44
zomer	□zo□-m□r	summer	74.21	70.00	53.66	64.88	59.61	63.91
zuivel	□zœy-v□l	dairy	67.86	67.62	54.35	65.87	60.91	67.06
zuurkool	□zyr-ko□l	sauerkraut	64.29	54.29	59.33	58.94	49.49	60.42

Note. All viseme overlap scores are listed in percentages and the means for Experiment 1 and 2 are calculated over all four experimental conditions.

Tables

Table 1

Viseme Categories (Visually Confusable Sets) of Dutch Consonants and Vowels (Van Son et al., 1994).

Consonants		Vowels	
Viseme Category	Phonemes	Viseme Category	Phonemes
{p}	/p, b, m/	{i}	/i, □, e, □/
{f}	/f, v, □/	{a}	/□ □, a, □/
{s}	/s, z, □/	{u}	/u, □, □/
{t}	/t, d, n, j, l/	{o}	/□ □, o/
{k}	/k, r, x, ŋ, h/	{au}	/œy, □u/

Table 2

Mean Percentages of Viseme Overlap Scores in the Visual-only Pilot for the Word Sets Created for Experiment 1 and 2 (with Standard Deviations in Parentheses).

		Set 1	Set 2	Set 3	Set 4
Speaker (M)	1	59.46 (14.90)	60.65 (12.63)	60.57 (14.92)	64.10 (10.28)
Speaker (F)	2	62.91 (15.17)	63.86 (10.50)	62.28 (16.01)	64.47 (13.75)

Table 3

Mean Percentages of Correct Word Identification in the Experimental Conditions of the Visual-only Identification Task in the Test Phase of Experiment 1 and 2 (with Standard Deviations in Parentheses).

		New words		Old words	
		New speaker	Old speaker	New speaker	Old speaker
Experiment 1	Speaker 1	12.64 (7.98)	12.92 (7.04)	22.64 (13.19)	25.23 (11.73)
	Speaker 2	10.46 (6.06)	8.06 (5.47)	17.06 (8.10)	16.81 (10.83)
Experiment 2	Speaker 1	13.75 (7.51)	11.25 (6.80)	23.19 (12.06)	21.25 (11.03)
	Speaker 2	8.75 (6.28)	10.00 (7.74)	12.92 (8.06)	18.47 (11.12)

Figure Captions

FIG 1. Experiment 2: Mean percentage correct old/new word judgements at the test phase following auditory exposure to Speaker 1 (gray bars) and 2 (white bars) across the four experimental conditions. Error bars show the standard error of the mean.

Highlights

- Long-term word repetition priming occurs across modalities from auditory-only exposure to visual-only test.
- Neither speaker repetition nor word repetition affect explicit memory for auditory-only items during visual-only test.
- The magnitude of the word repetition priming effect is not modulated by speaker identity.
- Speaker-specific information is not transferred across modalities at the lexical level.