Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Speech Communication 44 (2004) 55–61

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Selective adaptation and recalibration of auditory speech by lipread information: dissipation

Jean Vroomen [a,*], Sabine van Linden [a], Mirjam Keetels [a], Béatrice de Gelder [a,b], Paul Bertelson [a,b]

[a] *Department of Psychology, Tilburg University, Warandelaan 2, 5000 LE Tilburg, The Netherlands*
[b] *Laboratoire de Psychologie Expérimental, Université Libre de Bruxelles, Belgium*

Received 2 February 2004; accepted 20 March 2004

Abstract

Recently, we have shown that lipread speech can recalibrate auditory speech identification when there is a conflict between the auditory and visual information (Bertelson, P., Vroomen, J., De Gelder, B, 2003. Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychol. Sci. 14 (2003) 592–597). When an *ambiguous* sound intermediate between /aba/ and /ada/ was dubbed onto a face articulating /aba/ (or /ada/), the proportion of responses consistent with the visual stimulus *increased* in subsequent unimodal auditory sound identification trials, revealing recalibration. In contrast, when an *unambiguous* /aba/ or /ada/ sound was dubbed onto the face (with no conflict between vision and audition), the proportion of responses *decreased*, revealing selective adaptation. In the present study we show that recalibration and selective adaptation not only differ in the direction of their aftereffects, but also that they dissipate at different rates, confirming that the effects are caused by different mechanisms.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Audio-visual speech; Aftereffect; Recalibration; Selective adaptation; Perceptual learning

1. Introduction

The question of how sensory modalities cooperate in forming a coherent representation of the environment is the focus of much current work at both the behavioral and the neuroscientific level. A substantial part of that work is carried out with *conflict* situations, in which incongruent information about potentially the same distal event is presented to different modalities (Bertelson and De Gelder, 2003). Exposure to such conflicting inputs produces two main effects: *immediate biases* and *aftereffects*. Immediate biases are

[*] Corresponding author. Tel.: +31 13 4662394; fax: +31 13 4662067.
E-mail address: j.vroomen@uvt.nl (J. Vroomen).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.03.009

changes in the perception of stimuli in a target modality produced by the presentation of incongruent stimuli in a distracting modality. One well-known example is the *ventriloquist effect*, in which the perceived location of target sounds is displaced toward light flashes delivered simultaneously at some distance, in spite of instructions to ignore the lights (Bertelson, 1999; Vroomen and DeGelder, 2004). Aftereffects are shifts following exposure to an intersensory conflict, when data in one or in both modalities are later presented alone. For the ventriloquist situation, aftereffects have been reported in which unimodal auditory localization is displaced in the direction of the light as seen in the exposure phase (Radeau and Bertelson, 1974; Frissen et al., 2003). The occurrence of such aftereffects implies that exposure to conflicting inputs *recalibrates* processing in the respective modalities as the previously experienced conflict is reduced.

Immediate biases and aftereffects have both been demonstrated for *spatial conflict* situations, but the existing evidence was, until recently, less complete for conflicts regarding *identities* for which biases had been reported consistently, but no aftereffects. The main example of cross-modal identity bias is the so-called McGurk-effect (McGurk and MacDonald, 1976) obtained when a particular speech token is delivered in synchrony with the visual presentation of a face articulating an incongruent token. In that situation, the reported speech token can be shifted toward the lipread distracter. For example, listeners perceive /da/ after hearing auditory /ba/ combined with visual /ga/. Though the McGurk-effect has been demonstrated repeatedly, for a long time no aftereffect consequent on exposure to McGurk pairs of stimuli was reported showing recalibration.

Recently, though, we managed to demonstrate recalibration of auditory speech by lipread information (Bertelson et al., 2003). When an *ambiguous* sound intermediate between /aba/ and /ada/ (henceforth A?) was dubbed onto a face articulating either /aba/ or /ada/ (A?Vb or A?Vd), the proportion of responses consistent with the visual stimulus *increased* in subsequent unimodal auditory sound identification trials. For example, when participants were exposed to A?Vb, they reported *more* /aba/ responses in subsequent testing. This was taken as a demonstration that the visual information had shifted the interpretation of the ambiguous auditory phoneme in its direction. This shift, then, was observable in subsequent testing.

In the same experiment, we also showed that when an *unambiguous* sound was dubbed onto a congruent face (AbVb or AdVd), the proportion of responses consistent with the visual stimulus *decreased*. Thus, when participants were, for example, exposed to AbVb, they reported *fewer* /aba/ responses in subsequent testing. This phenomenon was interpreted as *selective speech adaptation* (Eimas and Corbit, 1973). In selective speech adaptation, it is the repeated presentation of a particular speech utterance by itself (and thus in the absence of any conflict between auditory and visual information) that causes a *reduction* in the frequency with which that token is reported in subsequent identification trials. It probably reveals *fatigue* of some of the relevant processes, most likely acoustic or phonetic, although criterion setting may also play some role (Samuel, 1986). Within the same experimental situation, we thus showed that the *audio-visual conflict* in the audio-visual discrepant adaptors A?Vb (or A?Vd) caused recalibration to occur, whereas the *auditory component* of the unambiguous adaptor AbVb (or AdVd) caused selective adaptation.

In the present study, we further explored the possible differences between recalibration and selective adaptation. Here, we focused on how long the effects lasted. There is no doubt that recalibration and selective adaptation effects are both transient, but at present very little is known about how fast they dissipate, and whether they dissipate at equal rates or not. Participants were, as in Bertelson et al. (2003), exposed to audio-visual speech stimuli that contained either non-ambiguous or ambiguous auditory tokens taken from an /aba/–/ada/ speech continuum combined with the video of a face articulating /aba/ or /ada/ (A?Vb, A?Vd, AdVd, or AbVb). The effect of exposure to these tokens was measured on a subsequent auditory speech identification task such that we could trace aftereffects as a function of time of testing.

2. Method

2.1. Stimuli

A nine-point /aba/–/ada/ speech continuum was created and dubbed onto the video of a face articulating /aba/ or /ada/. Stimulus preparation started with a digital audio (Philips DAT-recorder) and video (Sony PCR-PC2E MiniDV) recording of a male speaker producing multiple repetitions of /aba/ and /ada/ utterances. Clearly spoken /aba/ and /ada/ tokens were selected and served as reference for the creation of the continuum. The stimuli were synthesized with the Praat program (http://www.praat.org/) (Boersma and Weenink, 1999). The glottal excitation source used in the synthesis was estimated from a natural /aba/ by employing the inverse filtering algorithm implemented in Praat. The stimuli were 640ms in duration with a stop consonant closure of 240 ms. A place-of-articulation continuum was created by varying the frequency of the second (F2) formant in equal steps of 39 Mel. The onset (before the closure) and offset (after the closure) frequency of the F2 was 1250Hz. The target frequency was 1100Hz for /aba/ and 1678Hz for /ada/ (see Fig. 1).

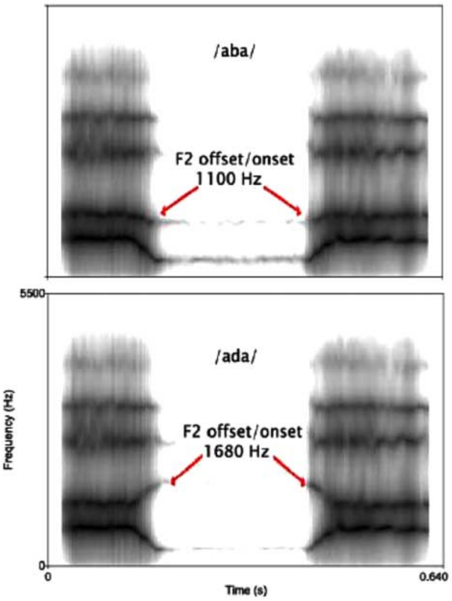

Fig. 1. Spectrogram of the /aba/ and /ada/ end-points.

The F1 transition changed from 750Hz to 350Hz before the closure for both stimuli. After the closure a mirror image of the transition was used. The duration of the transition was 40ms both before and after the closure of the consonant. The third (F3), fourth (F4), and fifth (F5) had fixed frequencies of 2500Hz, 3200Hz, and 4200Hz, respectively. The amplitude and the fundamental frequency contour followed those of the original /aba/ token.

The video recording showed the speaker facing the camera with the video frame extending from the neck to the forehead. Two video fragments were selected, different from the ones of the auditory tokens, one in which the speaker articulated /aba/, the other /ada/. The videos were digitized at 352 × 288 pixels at 30 frames per second. Each fragment lasted 2.5s and had a fade-in and fade-out of 330ms (10 video frames). The original audiotrack was replaced by one of the synthetic tokens such that the release of the consonant was synchronized with the original recording to the nearest video frame.

2.2. Procedure

Participants were tested individually in a soundproof booth. The videos were presented on a 17 in. monitor connected to a computer. The video filled about one third of the screen (10 × 9.5cm), and was surrounded by a black background. The sound was presented through a Fostex 6301B speaker placed just below the monitor. The loudness was 73dBa when measured at ear level. Participants were seated in front of the screen at a distance of 60cm.

The session involved three successive phases: a *calibration phase* to determine the stimulus that was nearest to the phoneme boundary (A?), followed by an auditory identification task that served as a *pre-test*, and finally three blocks of *audio-visual adaptation, each followed by a post-test*.

In the *calibration phase*, the participant was presented all stimuli of the continuum in random order and categorized them as /aba/ or /ada/. Tokens from the middle of the continuum were presented more often than tokens at the extreme

(6, 8, 14, 14, 14, 14, 14, 8 and 6 presentations for each of the nine stimuli, respectively). Participants were instructed to listen to each stimulus and to respond by pressing a 'b' or a 'd' on a keyboard upon hearing /aba/ or /ada/, respectively. The stimulus nearest to the 50% cross-over point was estimated via probit analysis, and this stimulus (A?) served as the most ambiguous stimulus in subsequent testing.

The *pre-test* consisted of 60 auditory-only test trials (2.5 s ITI), divided into 20 triplets. Each triplet contained the three auditory test stimuli nearest to the individually determined phoneme boundary (A? – 1, A?, A? + 1). Trials within a triplet were presented in different random orders. Participants responded by pressing a 'b' or a 'd' upon hearing /aba/ or /ada/, respectively.

For the *audio-visual exposure* phase, participants were randomly assigned to one of four groups (six participants each). A between-subjects design was used because we were concerned with possible transfer-effects. Participants were exposed to either AbVb, AdVd, A?Vb or A?Vd for three blocks of 50 trials each (1.5 s ITI). Five catch trials were interspersed during audio-visual exposure to ensure that participants were attending the face. Catch trials consisted of the presentation of a small white spot (12 pixels) between the lips and the nose of the speaker for three video frames (~100 ms). Participants had to press a key whenever a catch trial occurred (thus no phonetic categorization was required during the audio-visual exposure phase). Each of the three audio-visual exposure blocks was followed by an auditory-only identification task. These post-tests were the same as the pre-test, and thus consisted of 20 triplets of the three boundary stimuli (A? – 1; A?; A? + 1). Three quasi-random orders were used for the post-tests so that each of the three test-stimuli appeared once at each serial position.

3. Results

Calibration. The percentage of /aba/ responses in the auditory identification task was calculated for each of the nine auditory stimuli of the continuum (Fig. 2). The data showed the typical s-shaped identification curve. Each of the participants heard, as intended, the first tokens of the continuum as /aba/, and the last tokens as /ada/. The individually determined most ambiguous auditory stimulus (A?) ranged between stimulus 4 and 6.

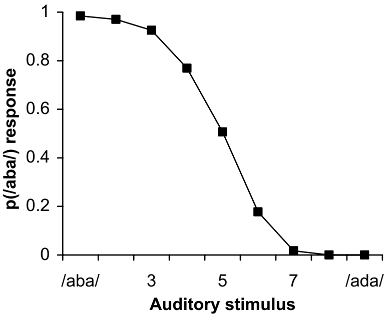

Fig. 2. Mean proportion of /aba/ judgements for each item of the continuum in the calibration phase.

Audio-visual exposure. Participants detected on the average 91% of the catch trials, indicating that they were indeed looking at the video during exposure. Aftereffects were calculated by subtracting the proportion of /aba/ responses in the pre-test from their proportion in the post-tests, so that a positive sign referred to an increase in responses consistent with the visual distracter as seen during the exposure phase. For example, when a participant responded in the pre-test on 50% of the trials /aba/, and following exposure to A?Vb, it was 60% in the post-test, then the aftereffect was +10%.

Fig. 3 shows the thus determined aftereffects as a function of the serial position of the test triplet. As in Bertelson et al. (2003), exposure to ambiguous sounds *increased* the number of post exposure judgements consistent with the visual distracter (i.e., *more* /aba/-responses after exposure to A?Vb, and *more* /ada/-responses after exposure to A?Vd; i.e., *recalibration*). The opposite effect was found after exposure to non-ambiguous sounds (*fewer* /aba/-responses after exposure to AbVb, and *fewer* /ada/-responses after exposure to AdVd; i.e., *selective adaptation*). As is apparent in Fig. 3, the recalibration effect was short-lived and lasted for only six test trials (the first and second triplet positions of the test items), whereas selective adaptation lasted for the whole test.

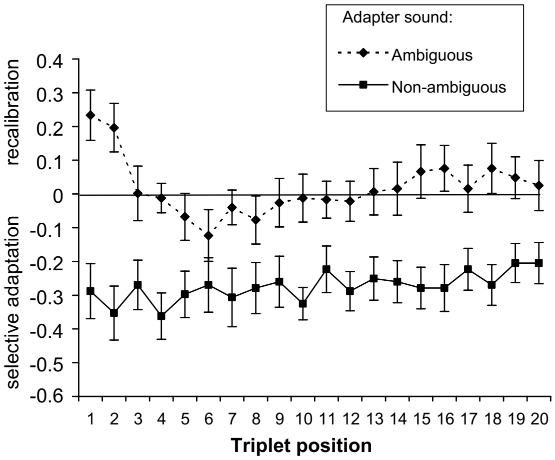

Fig. 3. Aftereffects as a function of the serial position in the post-test. After exposure to ambiguous sounds (A?Vb and A?Vd), the number of responses consistent with the video *increased* (= recalibration) at triplet positions 1 and 2 (test trials 1–6). After exposure to non-ambiguous sounds (AbVb and AdVd), the number of responses consistent with the video *decreased* (=selective adaptation) from triplet positions 1 thru 20 (test trials 1–60).

A 2 (non-ambiguous-sound exposure vs. ambiguous-sound exposure) × 20 (triplet position) ANOVA (with the sign of the effect reversed for non-ambiguous sound exposure) on the aftereffects showed that the size of the aftereffect following exposure to non-ambiguous sounds (selective adaptation) was, on average, bigger than the one after exposure to ambiguous sounds (=recalibration), $F(1,22) = 9.44$, $p < .006$. An main effect of triplet position was found, $F(19,418) = 3.63$, $p < .001$, as all aftereffects dissipated. Importantly, there was an interaction between the two factors, $F(19,418) = 2.92$, $p < .001$, indicating that aftereffects dissipated faster for recalibration than for selective adaptation. Separate *t*-test for each triplet showed that recalibration-effects were significantly bigger than zero ($p < .01$) only at triplet positions 1 and 2, whereas selective adaptation effects were significant at all triplet positions.

4. Discussion

Exposure to a particular visual speech token combined with the corresponding non-ambiguous auditory token resulted in a reduced tendency to produce that token, i.e., the typical selective speech adaptation effect. The same visual token combined instead with an ambiguous auditory token resulted in the opposite shift, a more frequent production of the response consistent with the visual adapter, indicative of cross-modal recalibration. Thus, as in Bertelson et al. (2003), a dissociation between the two adaptation effects was obtained under otherwise identical conditions, just by manipulating the ambiguity of the auditory speech presented during adaptation.

The new finding in the present study is that the two effects dissipate at different rates. Whereas recalibration lasted, in the present set-up, only about six test trials, selective adaptation could be observed even after 60 test trials. This difference confirms that the two adaptation phenomena result from different underlying mechanisms.

Interestingly, despite the fact that aftereffects of clear and ambiguous speech tokens were very different in terms of their direction and rate of dissipation, participants were hardly able to distinguish between these two kinds of adapters. In subsequent identification tasks of the adapter stimuli, virtually all A?Vb and AbVb tokens were labeled as /b/ (98% and 100%, respectively), and all A?Vd and AdVd tokens were labeled as /d/ (both 100%). Moreover, in an ABX task in which participants had to judge whether the audio of an audiovisual adapter stimulus was identical to A?Vb or AbVb, (or A?Vd vs. AdVd), performance was only 52% correct (with chance level being 50%). Thus even when explicitly asked to discriminate between clear and ambiguous speech tokens, listeners performed at chance level when these tokens were combined with a video. This implies that conscious response strategies of the participants cannot be held responsible for the effects we observed, as listeners found it virtually impossible to distinguish clear from ambiguous adapter tokens.

The conditions under which recalibration and selective adaptation were obtained may also shed light on a study in which aftereffects due to recalibration *might* have been observed, namely the one by Roberts and Summerfield (1981; later replicated by Saldaña and Rosenblum, 1994). The original purpose of this study was not to explore recalibration as a consequence of exposure to audio-visual conflict, but to separate acoustic from phonetic contributions to selective speech adaptation. The experiment involved the repeated presentation of an audio-visual discrepant adaptor (auditory /bε/ combined with visual /gε/, henceforth AbVg) followed by post-tests with speech tokens from an auditory /bε/–/dε/ continuum. The authors reported that following exposure to AbVg, more /dε/ responses were given. This increase in /dε/ responses, though, is difficult to attribute uniquely because it might be caused by both selective adaptation (a fatigue of the auditory '/bε/-detector') *and* by recalibration (the intersensory conflict in the AbVg adapter shifted auditory phoneme perception towards /dε/ [1]). The authors also used an auditory /bε/ as adapter (Ab), and found that aftereffects of this audio-alone stimulus were not different from the audio-visual incongruent AbVg adapters. This absence of a difference between Ab and AbVg adapters might, at first sight, rule out a contribution from recalibration. Yet, it might also be the case that ceiling effects were at play with the AbVg adapter, such that recalibration effects were overwhelmed by selective adaptation.

Recalibration of phoneme boundaries has, since our initial report (Bertelson et al., 2003), now also been reported to occur via lexically induced knowledge. Norris et al. (2003) replaced the final fricative (/f/ or /s/) from critical words by an ambiguous sound, intermediate between /f/ and /s/. Listeners heard this ambiguous sound /?/ either in /f/-final words (e.g., /witlo?/, from *witlof*, chicory) or in /s/-final words (e.g., /naaldbo?/, from *naaldbos*, pine forest). Listeners who heard /?/ in /f/-final words were in subsequent testing more likely to report /f/, whereas those who heard /?/ in s-final words were more likely to report /s/. These results are thus analogous to the present ones, implying that both lipread and lexical information can recalibrate phoneme boundaries. Both phenomena therefore seem to reflect perceptual learning effects.

Interestingly, Samuel (2001) used adapter stimuli very similar to Norris et al., but did not find recalibration effects. He presented an ambiguous /s/–/š/ sound in the context of an /s/-final word (e.g., bronchitis) or /š/-final word (demolish), followed by a test involving /s/–/š/ identification. In contrast with Norris et al. (2003), no recalibration effect was observed, but a (small) selective adaptation effect (e.g., *less* /s/ responses after hearing 'bronchiti?). For the time being, the origin of this difference remains unclear, so that we can only speculate. One possibility would be that selective adaptation and recalibration both occur at the same time, but that one outweighs the other. For example, consider a 'not-so-good' /s/ (i.e., a stimulus intermediate between a good /s/ and a completely ambiguous /?/) in the context of 'bronchiti_'. One can imagine that if this stimulus were used as adapter, there is selective adaptation because there is acoustic information in the stimulus that specifies /s/. At the same time, there may be recalibration because there is a context which specifies that this somewhat ambiguous /s/ is indeed an /s/. Recalibration and selective adaptation might then play a role within the same stimulus, and aftereffects will be dependent on the relative weight of the two. If so, it becomes important to chart the conditions under which recalibration and selective adaptation occur, as they may, for example, not only dissipate at different rates, but also be acquired differently.

5. Conclusions

Exposure to audio-visual speech can modify auditory speech identification through both visual recalibration and unimodal selective speech adaptation. The distinction between these two forms of adaptation is supported by our earlier finding in that they produced aftereffects in opposite directions. The present study (a) confirms this direction of adaptation argument, and (b) provides the new argument that the two aftereffects dissipate at different rates.

Acknowledgment

We like to thank Jyrki Tuomainen for help with creating the auditory speech continuum.

[1] Note that lipread /g/ is similar to lipread /d/.

References

Bertelson, P., 1999. Ventriloquism: a case of crossmodal perceptual grouping. In: Aschersleben, G., Bachmann, T., Müsseler, J. (Eds.), Cognitive Contributions to the Perception of Spatial and Temporal Events. Elseviers, Amsterdam, pp. 347–362.

Bertelson, P., De Gelder, B., 2003. Multisensory integration, perception and ecological validity. Trends Cogn. Sci. 7, 460–467.

Bertelson, P., Vroomen, J., De Gelder, B., 2003. Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychol. Sci. 14, 592–597.

Boersma, P., Weenink, D., 1999. Praat, a system for doing phonetics by computer. http://www.fon.hum.uva.nl/praat/.

Eimas, P.D., Corbit, J.D., 1973. Selective adaptation of linguistic feature detectors. Cognitive Psychol. 4, 99–109.

Frissen, I., Vroomen, J., De Gelder, B., Bertelson, P., 2003. The aftereffects of ventriloquism: are they sound frequency specific? Acta Psychol. 113, 315–327.

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

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognitive Psychol. 47, 204–238.

Radeau, M., Bertelson, P., 1974. The aftereffects of ventriloquism. Q. J. Exp. Psychol. 26, 63–71.

Roberts, M., Summerfield, Q., 1981. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Percept. Psychophys. 30, 309–314.

Saldaňa, A.G., Rosenblum, L.D., 1994. Selective adaptation in speech perception using a compelling audiovisual adaptor. J. Acoust. Soc. Am. 95, 3658–3661.

Samuel, A.G., 1986. Red herring detectors and speech perception: In defence of selective adaptation. Cognitive Psychol. 18, 452–499.

Samuel, A.G., 2001. Knowing a word affects the fundamental perception of the sounds within it. Psychol. Sci. 12 (11), 348–351.

Vroomen, J., DeGelder, B., 2004. Perceptual effects of crossmodal stimulation: The cases of ventriloquism and the freezing phenomenon. In: Calvert, G., Spence, C., Stein, B.E. (Eds.), Handbook of multisensory processes, MIT Press, Cambridge, pp. 141–150.