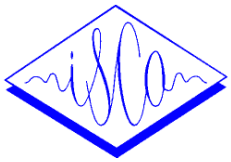

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Selective Adaptation and Recalibration of Auditory Speech by Lipread Information: Dissipation

Jean Vroomen[1], *Mirjam Keetels*[1], *Sabine van Linden*[1], *Béatrice de Gelder*[1,2], & *Paul Bertelson*[2,1]

[1]Department of Psychology, Tilburg University, The Netherlands
[2]Laboratoire de Psychologie Expérimental, Université Libre de Bruxelles, Belgium
j.vroomen@uvt.nl

Abstract

Recently, we have shown that visual speech can recalibrate auditory speech identification [1]. When an ambiguous sound intermediate between /aba/ and /ada/ was dubbed onto a face articulating /aba/ (or /ada/), then the proportion of /aba/ responses increased in subsequent unimodal auditory sound identification trials. In contrast, when an unambiguous /aba/ sound was dubbed onto the face articulating /aba/, then the proportion of /aba/ responses decreased, revealing selective adaptation. Here we show that recalibration and selective adaptation not only differ in the direction of their after-effects, but also that they dissipate at a different rates, confirming that the effects are caused by different brain mechanisms.

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 levels. 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 [2]. Exposure to such conflicting inputs produces two main effects: immediate biases and after-effects. Immediate biases are effects of incongruent inputs in a distracting modality on perception of inputs in a target modality. For example, in the so-called ventriloquist effect, the perceived location of target sounds is displaced toward light flashes delivered simultaneously at some distance, in spite of instructions to ignore the latter [3]. After-effects are shifts following exposure to an intersensory conflict, when data in one or in both modalities are later presented alone. For the ventriloquism situation, after-effects have been reported in which unimodal auditory localization is displaced in the direction occupied during the preceding exposure phase by the distracters in the other modality [4]. They show that exposure to conflicting inputs recalibrates processing in the respective modalities in a way that reduces the previously experienced conflict.

Although immediate biases and after-effects have been consistently demonstrated for spatial conflict situations, the existing evidence is less complete for conflicts regarding identities. Here, biases have been reported consistently, but not after-effects. The main example is the conflict resulting from the auditory delivery of a particular speech token in synchrony with the visual presentation of a face articulating an incongruent token. This situation can produce strong biases of auditory identification towards the lip-read distracter (e.g., auditory /ba/ combined with visual /da/ is 'heard' as /da/), generally known as McGurk effects [5]. On the other hand, several studies have failed to observe recalibration subsequent to exposure to McGurk-like stimulus pairs [6, 7, 8].

In a recent study [1], though, we have shown that the interpretation of these results is complicated by a phenomenon known as selective speech adaptation [9], by which the repeated presentation of a particular speech utterance by itself, and in the absence of any distracter, causes a reduction in the frequency with which that token is reported in subsequent identification trials. It is thus an adaptation phenomenon that, like recalibration, manifests itself by after-effects, but, unlike recalibration, does not depend on the presence of conflicting inputs. It probably reveals fatigue of some of the relevant processes, although criterion-setting operations, resulting in "range" or "contrast" effects, can also be involved under particular conditions [10].

We managed to isolate the effect of recalibration by using as auditory component of the audio-visual adapting stimulus an ambiguous synthetic token, intermediate between /aba/ and /ada/ (henceforth, A?). Items of this kind have been shown to cause no selective adaptation [11], and on the other hand to be susceptible to strong McGurk-biases [12]. Exposure to bimodal stimulus pairs combining an ambiguous auditory token with either of the two corresponding extreme visual tokens (henceforth A?Vb and A?Vd for an auditory ambiguous token combined with visual /aba/ and visual /ada/, respectively) was found to increase the number of post-exposure judgments consistent with the visual distracter, thus showing recalibration. In contrast, exposure to unambiguous auditory tokens combined with congruent visual tokens (AbVb and AdVd) decreased the number of post exposure judgments consistent with the visual information, revealing selective adaptation.

In the present study, we further explored the possible differences between the two adaptation phenomena. Here, we focused on the duration of the effects. 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 [1], exposed to audio-visual speech stimuli that contained either non-ambiguous or ambiguous auditory tokens taken from an /aba/ - /ada/ speech continuum (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 after-effects as a function of time of testing.

2. Method

2.1. Stimuli

A 9-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/) [13]. 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 640 ms 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 1250 Hz. The target frequency was 1100 Hz for /aba/ and 1678 Hz for /ada/ (see Figure 1).

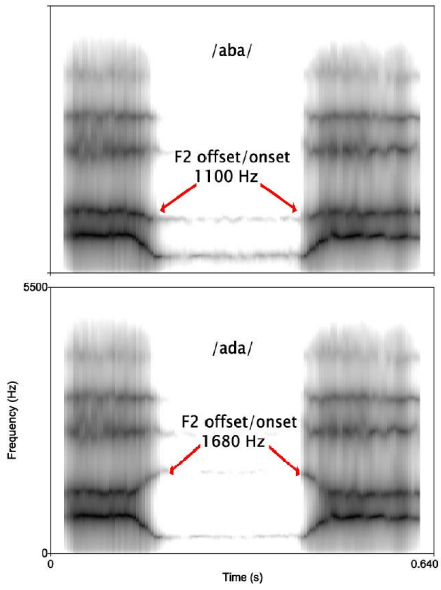

*Figure 1:*Spectrogram of the /aba/ and /ada/ end-points.

The F1 transition changed from 750 Hz to 350 Hz before the closure for both stimuli. After the closure a mirror image of the transition was used. The duration of the transition was 40 ms both before and after the closure of the consonant. The third (F3), fourth (F4), and fifth (F5) had fixed frequencies of 2500 Hz, 3200 Hz, and 4200 Hz, 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 352x288 pixels at 30 frames per s. Each fragment lasted 2.5 sec and had a fade-in and fade-out of 330 ms (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 sound-proof booth. The videos were presented on a 17-inch monitor connected to a computer. The video filled about one third of the screen (10 x 9.5 cm), and was surrounded by a black background. The sound was presented through a Fostex 6301B speaker placed just below the monitor. The loudness was 73 dBa when measured at ear level. Participants were seated in front of the screen at a distance of 60 cm.

Participants took part in three tests: An auditory identification task that served as pretest to determine which stimulus of the continuum was nearest to the phoneme boundary (henceforth A?), followed by an auditory identification task that served as a baseline, followed by an audio-visual exposure phase with a post-test.

In the pretest, we determined, for each participant individually, the stimulus that was nearest to the /aba/-/ada/ phoneme boundary. The test consisted of 98 auditory-only trials where each of the nine stimuli was presented in random order at 1.5 s ITI. 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 base-line 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 (6 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. They consisted of the presentation of a small white spot (12 pixels) between the lips and the nose of the speaker for three videoframes (~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 base-line 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

Pretest. The percentage of /aba/ responses on the auditory identification task was calculated for each of the nine auditory stimuli of the continuum (Figure 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.

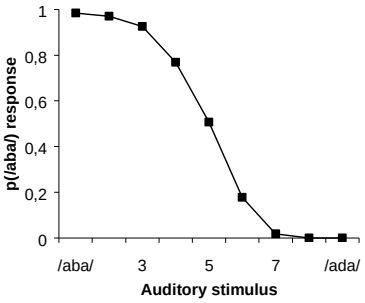

Figure 2: Auditory pre-tests. Mean proportion of /aba/ judgments for each item of the continuum.

Audio-visual exposure. Participants detected on the average 91% of the catch trials, indicating that they were indeed attending to the video during exposure. After-effects were calculated by subtracting the proportion of /aba/ responses in the baseline 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 base-line test on 50% of the trials /aba/, and following exposure to A?Vb, the /aba/ response in the post-test was 60%, then the after-effect was 10%.

Figure 3 shows the thus determined after-effects as a function of the serial position of the test triplet. As in [1], exposure to ambiguous sounds increased the number of post exposure judgments consistent with the visual distracter (i.e., more /aba/-responses after exposure to A?Vb, and more /ada/-responses after exposure to A?Vd), whereas the opposite effect was found after exposure to non-ambiguous sounds (less /aba/-responses after exposure to AbVb, and less /ada/-responses after exposure to AdVd). As is apparent in Figure 3, the recalibration effect was very transient and lasted for only about 6 test trials (the first and second triplet positions), whereas selective adaptation lasted for the whole test.

A 2 (non-ambiguous-sound exposure vs. ambiguous-sound exposure) x 20 (triplet position) ANOVA (with the sign of the effect reversed for non-ambiguous sound exposure) on the after-effects showed that the after-effect 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 effect of triplet position was found, $F(19,418) = 3.63$, $p < .001$, as after-effects dissipated. Importantly, there was an interaction between the two factors, $F(19, 418) = 2.92$, $p < .001$, as after-effects dissipated faster for recalibration than for selective adaptation. Separate t-test 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.

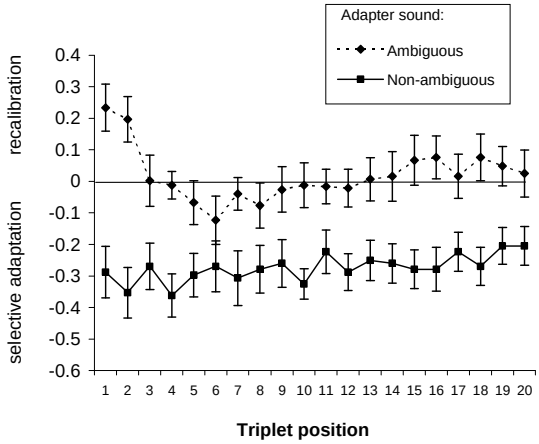

Figure 3: After-effects as a function of the serial position in the post-test. After exposure to ambiguous sounds (A?Vb and A?Vd), the number of responses consistent with the video increased (= recalibration) at triplet positions 1 and 2 (test trials 1-6). After exposure to non-ambiguous sounds (AbVb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

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 the ambiguous auditory token resulted in the opposite shift, a more frequent production of that token, indicative of crossmodal recalibration. Thus, as in [1], 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 also dissipate at different rates. Whereas recalibration lasted only about 6 test trials, selective adaptation could be observed even after 60 test trials. This confirms that the two adaptation phenomena result from different underlying mechanisms. Recalibration of speech is, like the well-known case of spatial recalibration, contingent on exposure to conflicting information from different sources [3]. It is a perceptual learning effect, and it may be of benefit to adjusting phoneme boundaries to new speakers. On the contrary, selective speech adaptation occurs in the absence of conflict, and could, to some extent at least, reflect the fatigue of some of the processing mechanisms by repeated extreme stimulation. These two forms of adaptation co-occur also in other domains of perception. Conflict-based recalibrations has been demonstrated, beyond the cases mentioned in the introduction of crossmodal spatial conflicts, also for analogous intra-modality conflicts, for instance between different cues to visual depth [14]. Conflict-free adaptation forms are manifested in visual cases of "sensory adaptation" like color, curvature, or motion [e.g., 15]. The two types of phenomena have generally been the objects of separate research lines, and their relations have rarely been investigated. Our study [1] was probably the first in which the two forms of adaptation have been dissociated within the same situation.

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 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.

6. Acknowledgements

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

7. References

[1] Bertelson, P., Vroomen, J. and De Gelder, B., "Visual recalibration of auditory speech identification: A McGurk after-effect", *Psych. Sci.*, in press.

[2] Bertelson, P. and De Gelder, B., "The psychology of multimodal perception", In Spence, C. and Driver, J., (Eds.), *Crossmodal Space and Crossmodal Attention*, Oxford University Press, London, in press.

[3] Bertelson, P., "Ventriloquism: A case of crossmodal perceptual grouping", In Aschersleben, G., Bachmann, T. and Müsseler, J., (Eds.), *Cognitive contributions to the Perception of Spatial and Temporal Events*, Elseviers, Amsterdam, 1999.

[4] Radeau, M. and Bertelson, P., "The after-effects of ventriloquism", *Q. J. Exp. Psychol., Vol. 26, 1974, p. 63-71.*

[5] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", *Nature, Vol. 264, 1976, p. 746-748.*

[6] Roberts, M. and Summerfield, Q., "Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory", *Percept. Psychophys., Vol. 33, 1981, p. 309-314.*

[7] Saldaňa, A. G. and Rosenblum, L. D., "Selective adaptation in speech perception using a compelling audiovisual adaptor", *J. Acoust. Soc. Amer., Vol. 95, 1994, p. 3658-3661.*

[8] Shigeno, S. (2002)., "Anchoring effects in audiovisual speech perception", *J. Acoust. Soc. Amer., Vol. 111, 2002, p. 2853-2861.*

[9] Eimas, P. D. and Corbit, J. D., "Selective adaptation of linguistic feature detectors", *Cognitive Psychol., Vol. 4, 1973, p. 99-109.*

[10] Samuel, A. G., "Red herring detectors and speech perception: In defence of selective adaptation", *Cognitive Psychol., Vol. 18, 1986, p. 452-499.*

[11] Sawusch, J. R. and Pisoni, D. B., "Response organisation in selective adaptation to speech sounds", *Percept. Psychophys., Vol. 20, 1976, p. 413-418.*

[12] Bertelson, P., Vroomen, J., Wiegeraad, G. and De Gelder, B. "Exploring the relation between McGurk interference and ventriloquism", *ICSLP Proc.*, 559-562, 1994.

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

[14] Epstein, W., "Recalibration by pairing: A process of perceptual learning", *Perception, Vol. 4, 1975, p. 59-72.*

[15] Gibson, J. J., "Adaptation, after-effects and contrast in perception of curved lines", *Percept. Psychophys,. Vol. 18, 1933, p. 31.*