Lipread Aftereffects in Auditory Speech Perception: Measuring Aftereffects After a Twenty-Four Hours Delay

Jean Vroomen, Sabine van Linden, and Martijn Baart

Tilburg University, Dept of Psychology
Tilburg, The Netherlands

Abstract

Lipreading can evoke an immediate bias on auditory phoneme perception [e.g. 6] and it can produce an aftereffect reflecting a shift in the phoneme boundary caused by exposure to an auditory ambiguous stimulus that is combined with non-ambiguous lipread speech (recalibration, [1]). Here, we tested the stability of lipread-induced recalibration over time. Aftereffects were measured directly after exposure and after 24 hours. Aftereffects dissipated quickly during testing and were not observable anymore after a 24 hours delay.

Index Terms: Audio-visual speech, Recalibration, Aftereffects, Lipreading.

1. Introduction

Lipread information and lexical information can both influence perception of spoken language. These information sources not only change the immediate percept of an ambiguous speech sound (a direct bias) [e.g. 6, 4], but they can also produce a shift in the categorization of an ambiguous phoneme following a period of exposure to the ambiguous phoneme that is combined with lipread or lexical context information [e.g. 1, 8, 9]. The shift in the phoneme boundary is observable as an aftereffect and it probably reflects an adaptive shift (i.e., recalibration).

Lipread speech is a so-called 'bottom-up' source of information, and as such it is an inherent property of the speech stimulus. Lexical context, on the other hand, depends on knowledge of the language stored in the listener's brain, and is considered a 'top-down' source of information. There are several studies suggesting that bottom-up lipread information exerts a stronger effect on auditory speech perception than lexical information does. For example, Munhall & Tohkura [7] used a gating paradigm to study the time course of audio-visual speech processing and reported that lipread information can be available to a listener even before the auditory speech signal is heard. In contrast, lexical effects occur relatively late in the identification process, and are usually only obtained after the word is recognized. Brancazio [2] found in phoneme categorization that lipread effects were stronger than lexical effects and were present in slow, medium and fast responses, whereas lexical effects were only significant in slow and medium responses. Van Linden and Vroomen [9 in press] also reported that lipread information produced bigger direct bias effects on the perception of ambiguous phonemes than lexical information.

As concerns recalibration, it appears that lipread and lexical induced aftereffects are very similar to each other, as both effects are about equally big and dissipated quickly with prolonged testing. For example, in a previous study, we observed that lipread and lexical recalibration effects disappeared after only 12 test tokens [10]. Others, though, have reported that lexical recalibration lasts much longer than the ones reported in our studies [e.g. 5, 3]. For example, Kraljic and Samuel [5] reported that a lexically-induced aftereffect was still present after a twenty-five minute delay, and it had not become smaller if measured directly after exposure. Eisner and McQueen [3] also reported lexically-induced aftereffects that were still observable after a 12 hours delay, and again, the magnitude of the effects seemed to remain stable over time.

One potentially relevant difference between studies on lipread and lexically induced aftereffects is that the latter not only used an ambiguous /?/ sound during exposure (e.g., 'witlo/?/'; *witlo**f*** = *chicory*), but also the unambiguous phoneme of the opposite category (radij**s**; *radish*). We demonstrated that the use of these contrast stimuli enhance aftereffects [9], possibly because they increase the contrast. On this view, participants who will hear, say, a clear /s/ on an /s/-/f/ continuum. will have a tendency to judge anything that is not a clear /s/ (the /?/ stimulus) as an /f/. Such a contrast effect will enhance the magnitude of the aftereffect, and the contribution of recalibration proper is then overestimated

Here we explored whether lipread-induced aftereffects do indeed become bigger and more stable over time when the exposure phase includes such contrast stimuli. Lipread recalibration was induced by exposing participants to an ambiguous speech sound /?/, halfway between /t/ and /p/, that was combined with non-ambiguous visual speech (A?Vt or A?Vp). In addition, we exposed participants to non-ambiguous contrast stimuli of the opposing category (ApVp or AtVt). Following exposure to these stimuli, participants, were tested on auditory-only test trials immediately after the exposure phase and after 24 hours. Dissipation rates of the aftereffects were measured as a function of the serial position in the test.

2. Method

Participants. Twenty native speakers of Dutch (18-25 yrs old) with normal hearing and normal seeing participated.
Materials. By using the Praat speech editor (http://www.praat.org) (Boersma and Weenink, 1999), a 10-point continuum between /t/ and /p/ was created by changing the second (F2) and third (F3) formants. Video and audio trials were created by recording a male native speaker of Dutch on digital audio and video tape (Philips DAT-recorder and Sony PCR-PC2E mini DV) with video trials showing the speaker's nose, mouth, chin and cheeks. The /?/ test tokens were dubbed onto the recordings of the pseudo word soo/?/ which was used during calibration, training and the post-test phase, and on the pseudo words foo/?/, woo/?/, kaffoo/?/ and dikasoo/?/ which were used during exposure. See [9] for a more elaborate description of the stimuli.
Procedure. Participants were tested individually in a soundproof booth. Half of the participants were recalibrated

towards /t/ and were thus exposed to A?Vt and ApVp; the other half of the participants was recalibrated towards /p/ and were exposed to A?Vp and AtVt.

Participants were seated at a distance of 60 cm in front of a 17-inch CRT-monitor on which the video fragments were presented. The audio samples were presented via two regular computer speakers (JBL Media 100WH/230) placed on left and right of the monitor. The video fragments were 10 by 9.5 centimetres in size, and were shown in a black background. Loudness of the words and non-words was measured at 70 dBa at ear level. A regular keyboard was used for data-acquisition. During the test, participants were instructed to press the p-key upon hearing /sop/ and t-key upon hearing /sot/.

The whole experiment consisted of four phases: a calibration phase, a training phase, an exposure-test phase and a second test phase after 24 hours.

Calibration. In the calibration phase it was determined, for each individual participant, which was the most ambiguous test token /?/ of the continuum. All test tokens were presented 10 times in pseudo random order and participants were asked to indicate whether they heard 'soot' or 'soop' by pressing the appropriate key. By using a logistic procedure, the obtained s-shaped response curve was used to determine the 50% crossover point for each participant. The test-item nearest to this point, served as the most ambiguous stimulus /?/ for subsequent testing.

Training. To acquaint participants with the test procedure, they had to categorize the ambiguous /?/ token, and the two tokens nearest to this stimulus; the more 'p'-like token /?-1/ and the more 't'-like token /?+1/. Each of the three tokens was presented twenty times in pseudo-random order.

Exposure – test. Participants were presented five blocks of 16 exposure stimuli followed by 60 test trials. In the t-condition, each exposure block contained eight A?Vt stimuli and eight contrast stimuli ApVp. In the /p/ condition, each exposure block contained eight A?Vp and eight AtVt stimuli. Table 1 provides an overview of the stimuli in both conditions. To ensure that participants were watching the monitor during the exposure phase, catch trials were included consisting of a short appearance (100 ms) of a small white dot on the upper lip of the speaker. Upon detecting a catch trial, participants pressed a special key.

Each exposure block was immediately followed by 60 auditory-only test trials. In this test phase, the three test tokens (/?-1/, /?/, and /?+1/) were presented twenty times in the pseudoword /so?/ in counterbalanced order. Participants were asked to indicate whether they heard 'soop' or 'soot' the same way as they did in the calibration and training phase.

Second test after a 24 hours delay.

The second test was delivered 24 hours after exposure. Participants were again presented five blocks of 60 auditory test stimuli. Stimuli and procedure were the same as in the immediate test phase, except that participants were not exposed anew to exposure trials. Instead they tried to solve a Rubick's cube (a visual puzzle) during a one-minute interval between the successive series of test trials.

Table 1. *Stimuli as used in the exposure phase*

T condition		P condition	
Auditory	Lipread	Auditory	Lipread
foo/?/	foot	foo/?/	foop
woo/?/	woot	woo/?/	woop
kaffoo/?/	kaffoot	kaffoo/?/	kaffoop
dikasoo/?/	dikasoot	dikasoo/?/	dikasoop

Contrast stimuli

T condition		P condition	
Auditory	Lipread	Auditory	Lipread
foop	Foop	foot	foot
woop	Woop	woot	woot
kaffoop	kaffoop	kaffoot	kaffoot
dikasoop	dikasoop	dikasoot	dikasoot

3. Results

The most ambiguous stimulus ranged between tokens 3 and 7 of the ten synthesized test tokens. In the training phase, 50 % of the stimuli were judged as /t/ in the t-condition, and 51 % in the /p/-condition indicating that the proportion of t- and p-responses before exposure started was in both groups alike. During exposure, 99 % of the catch trials were detected, indicating that participants kept their eyes fixated on the monitor.

To investigate the dissipation rate of lipread recalibration, the 60 test trials were binned in 10 subsequent serial positions with each position representing the mean average number of t-responses of 6 consecutive test trials. The proportion of t-responses is shown in figure 1 (immediate test) and figure 2 (delayed test). High values indicate that more 't' and thus less 'p' responses were given.

Aftereffects were calculated as in previous studies by subtracting the proportion of t-responses following p-exposure from t-exposure. Figure 3 shows the difference in t- versus p-exposure. For the immediate test, a 2 (t- or p-exposure) x 10 (test token position) ANOVA on the proportion of t-responses showed a significant main effect of exposure condition, F1,18, = 12.19, $p < .003$, because there were, as expected, more t-responses following t-exposure rather than p-exposure. The interaction with test token position was also significant, $F(9,162) = 7,39$, $p < .001$, as aftereffects dissipated with prolonged testing. Separate t-tests showed that the aftereffects was bigger than zero up to test token position 2 (or 12 consecutive test trials).

The same ANOVA on the data of the test after 24 hours showed that none of those effects was significant anymore. Thus, there was no overall difference between the t- and p-exposed groups, $F(1,18) = 1.66$, $p < .213$, and the interaction with test token position was not significant ($F< 1$).

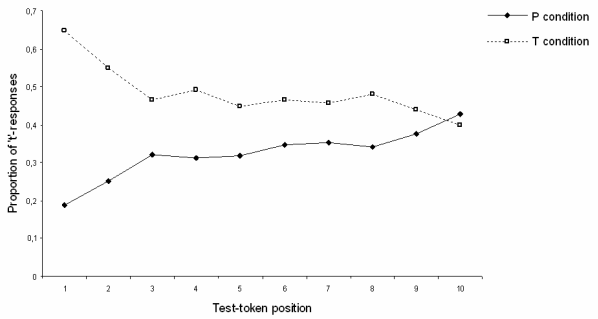

Figure 1: *Proportion of t-responses as a function of the serial position in the immediate test.*

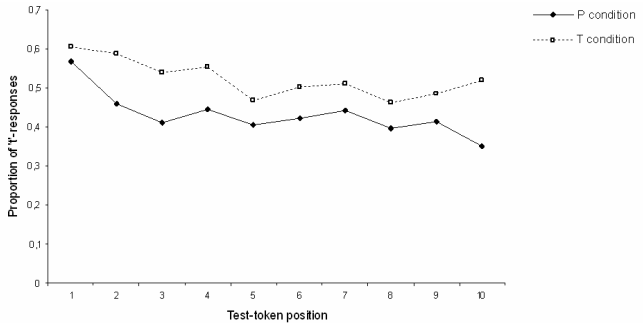

Figure 2: *Proportion of t-responses as a function of the serial position in the delayed test.*

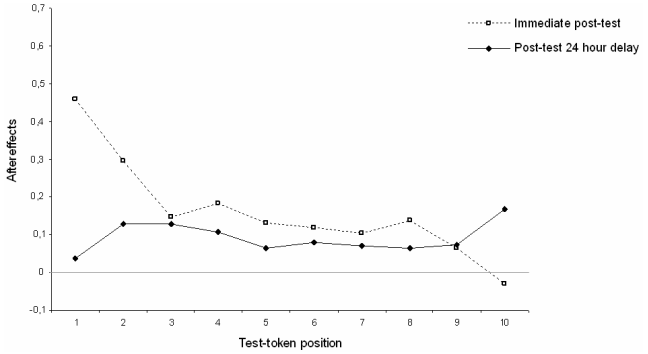

Figure 3: *Aftereffects as a function of the serial position in the test.*

4. Discussion

A substantial aftereffect induced by a combination of ambiguous speech sounds with lipread speech and contrast stimuli was observed following immediate testing. Presumably, this aftereffect reflects listeners' adjustment of the phoneme boundary that is driven by two distinct processes: On the one hand, there is reduction of the conflict between the lipread and ambiguous speech sound (i.e., recalibration proper). Second, there is a contrast effect driven by the auditory non-ambiguous stimuli from the opposing phoneme category. Both phenomena contributed to the large aftereffect on the first test token positions. Importantly, though, the aftereffect dissipated quickly (in 12 test trials) with prolonged testing [see also 10], and did not reappear following a 24 hours delay. Recalibration of phonetic categories and contrast effects are thus both transient phenomena.

Dissipation of the aftereffect is probably the result of a re-adjustment of the phoneme boundary back to normal. On this view, the immediate test has triggered dissipation and it explains why no effects were found after twenty-four hours.

The long-lasting lexical aftereffects reported by others [3,5,8] are most likely not caused by their use of contrast stimuli presented during the exposure phase. As observed before [9], contrast stimuli amplify the size of aftereffects, but they do not become more stable in time. Possibly, if others would measure aftereffects as a function of the serial position in the test, they would also observe that aftereffects dissipate quickly.

5. References

[1] Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science, 14(6)*, 592-597.

[2] Brancazio, L. (2004). Lexical Influences in Audiovisual Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance, 30(3)*, 445-463.

[3] Eisner, F., & McQueen, J. M. (2006) Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America, 119*(4), 1950-1953.

[4] Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6(1),* 110-125.

[5] Kraljic, T., & Samuel, A. G. (2005) Perceptual learning in speech. Is there a return to normal? *Cognitive Psychology, 51(2),* 141-178.

[6] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264* (5588), 746-748.

[7] Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America, 104*, 530- 539.

[8] Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204-238.

[9] Van Linden, S., & Vroomen, J. (in press). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance.*

[10] Vroomen, J., van Linden, S., Keetels, M., de Gelder, B. & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication, 44,* 55-62.